Development and validation of the startle and surprise Inventories and Visual Analogue scales

Chen, Jiayu; Landman, Annemarie; Derumigny, Alexis; Stroosma, Olaf; van Paassen, M. M.; Mulder, Max

DOI
10.1080/00140139.2025.2529317

Publication date
2025

Document Version
Final published version

Published in
Ergonomics

Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: www.tandfonline.com/journals/terg20

Development and validation of the startle and surprise Inventories and Visual Analogue scales

Jiayu Chen, Annemarie Landman, Alexis Derumigny, Olaf Stroosma, M. M. (René) van Paassen & Max Mulder

To cite this article: Jiayu Chen, Annemarie Landman, Alexis Derumigny, Olaf Stroosma, M. M. (René) van Paassen & Max Mulder (12 Jul 2025): Development and validation of the startle and surprise Inventories and Visual Analogue scales, Ergonomics, DOI: 10.1080/00140139.2025.2529317

To link to this article: https://doi.org/10.1080/00140139.2025.2529317

Published online: 12 Jul 2025.

ERGONOMICS
https://doi.org/10.1080/00140139.2025.2529317

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Development and validation of the startle and surprise Inventories and Visual Analogue scales

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

[a]Department of Control & Operations, Delft University of Technology, Delft, The Netherlands; [b]Department of Training and Performance Innovations, Organization for Applied Scientific Research (TNO), Soesterberg, The Netherlands; [c]Department of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

ABSTRACT

This paper 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, seven experts in the field assessed the content validity of 14 items for surprise, 7 items for startle derived from fundamental and applied literature. Elimination of items was based on a 50% agreement of relevance. In Phase 2, 81 participants completed the retained 19 items nine times, each time immediately after watching a video clip. A multilevel exploratory factor analysis was applied to assess the construct validity of items. In Phase 3, 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 a 11-item two-factor solution, corresponding to the constructs of startle and surprise. These results supported Startle-I and Surprise-I as measures of self-report startle and surprise, with Startle-VAS and Surprise-VAS as efficient alternatives.

PRACTITIONER SUMMARY: This study aimed to develop and preliminarily validate the self-report measuring instruments for startle and surprise. Drawing on content validity reviews, multilevel factor analysis and concurrent validation, we identified psychometrically-sound measures. These instruments enable practical assessment in applied, safety critical settings, supporting research and training that require accurate measurement.

ARTICLE HISTORY
Received 9 April 2025
Accepted 25 June 2025

KEYWORDS
Aviation; Psychometric Measures; Pilot Performance; Cognition; Experience; Emotions

1. Introduction

The leading cause of fatal aviation accidents worldwide between 2004 and 2024 was Loss of Control In-flight (LOC-I) in commercial aviation (Airbus 2025). Although improvements to flight systems and automation have greatly increased flight safety in the last decades, unexpected situations still require human intervention. These include system malfunctions, unforeseen weather events and bird strikes (Banks et al. 2022). Such events can be highly demanding and stressful for pilots (Stanton, Li, and Harris 2019). A frequently cited contributing factor in accident investigation reports is inappropriate pilot flight control inputs or actions (International Air Transport Association 2019). Pilots are required to assess situation, make sense of unexpected information, and take actions to maintain control of the aircraft. Startle and surprise have been contributing factors in LOC-I incidents and accidents, as upsets that occur in normal operations are unplanned and inadvertent, which could impact recognition or recovery (Casner, Geven, and Williams 2013; Federal Aviation Administration 2015; Landman et al. 2017b). To develop effective counter-measures, such as novel training methods or system interfaces, it is important to understand the responses of startle and surprise, as well as their effects on performance.

Startle refers to a coinciding physiological response elicited by a sudden, threatening, intense stimulus (Koch 1999). The startle reflex, typically within 20–50 milliseconds after a stimulus (Dreissen et al. 2012), involves the involuntary physiological reflexes and inhibition of muscular activities (e.g. eye-lid-closure, contraction of facial, neck and skeletal muscles; Koch 1999; Rivera et al. 2014), which can prepare the body for protection against adverse circumstances (Blumenthal 2015). If the threat persists, it is followed by the generalised stress response activated within the autonomic nervous system (Martin et al. 2015),

CONTACT Jiayu Chen J.Chen-12@tudelft.nl Department of Control & Operations, Delft University of Technology, Kluyverweg 1, Delft, 2629 HS, The Netherlands

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 (Dreissen et al. 2012; Holand et al. 1999; Jansen et al. 1995; Papadimitriou and Priftis 2009). 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.

The evidence suggests that the immediate psychomotor impact of startle may induce brief disorientation and short-term psychomotor impairments (Nakagawara et al. 2004). The startle response can inhibit muscular activity, but does not necessarily lead to degraded cognition (Civil Aviation Authority 2016). It was found that cognitive impairments could last for up to 30 seconds following a severe startle (Thackray and Touchstone 1983). Decreased information processing capability, distracted attention and pre-empted working memory can cause deterioration of task performance, which can result in lower performance accuracy and a considerable delay in decision-making (Eysenck et al. 2007; Martin, Murray, and Bates 2012; Stokes and Kite 2017; Thackray and Touchstone 1983).

Surprise, on the other hand, is an emotional and cognitive response to unexpected events that are (momentarily) difficult to explain (Horstmann 2006; Meyer, Reisenzein, and Schützwohl 1997). Its primary evolutionary function is thought to monitor the appropriateness of one's understanding of the world (Hansen and Topolinski 2011; Neisser 2014). When there is surprise, there is also a certain degree of unpreparedness for what is to come. Being able to predict relevant events before they occur is essential for survival. Surprise is a function of the discrepancy between expectations and perceived information from the environment, which can be positive, negative or neutral in valence (Noordewier, Topolinski, and Van Dijk 2016). When an individual is confronted with a surprising stimulus, ongoing thoughts and activities could be interrupted, and attention would be automatically directed to the surprising event (Horstmann 2006; Meyer, Reisenzein, and Schützwohl 1997). One's understanding of the situation needs to shift in order for the situation to make sense again. These "reframing" efforts require effortful, goal-directed attentional processing (Landman et al. 2017a). Such processing is vulnerable to the effects of stress (Eysenck and Calvo 1992). Stress may arise if the surprising event is threatening or startling, while the sense of uncertainty and lack of preparedness caused by the surprise may increase stress as well.

There exist methods to measure startle and surprise physiologically. The intensity of a startle response is widely measured through the eye-blink reflex. This can be done with surface electromyography(EMG), by which action potentials generated within the orbicularis oculi muscle can be detected (Blumenthal et al. 2005). Additionally, potentiometric (Hoffman, Marsh, and Stitt 1980), photoelectric (Flaten and Blumenthal 1998), vertical electro-oculographic (vEOG) (Gehricke, Ornitz, and Siddarth 2002) and magnetic search coil methods (Evinger and Manning 1993) can also measure eyelid movement. Beyond the blink reflex, pupil dilation is another physiological correlate of startle (Kinney and O'Hare 2020). Cardiovascular indicators, such as heart rate and blood pressure, have also been documented within 10 seconds of an acoustic startle, offering complementary autonomic markers (Holand et al. 1999).

Surprise can be physiologically measured through the EEG P300 event-related potential (ERP) (Noordewier et al. 2016), pupil dilation and activation in subcortical regions associated with dopamine (Antony et al. 2021). The P300 ERP originates in the anterior cingulate cortex (ACC), and peaks in ACC activity occur in aversive defence responses in general (Hajcak and Foti 2008). Pupil dilation was considered to represent the global state of the brain during cognitive processing as an effect of stress rather than solely elicited by surprise (Henckens et al. 2009). Pupil dilation related to surprise (maximum around 500 ms after the stimulus) was reported to be slower than pupil dilation due to startle (Kloosterman et al. 2015).

Physiological measures provide objective, real-time assessments and can overcome certain inherent biases, such as socially desirable answering patterns (Tran et al. 2007). However, the above-mentioned physiological measures are not specific to startle or surprise, but reflect broader autonomic, neural activation and affective responses (Bradley and Lang 2000). While these techniques offer high temporal resolution, they cannot distinguish between a startle response and similar defensive reactions triggered by fear or stress (Grillon and Baas 2003). They are also uneconomic for application to large numbers of participants (McCroskey 1984), and some are impractical and invasive to apply in operational settings. Physiological measures are 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 (Landman et al. 2017b), 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 (Rossato et al.

2018), 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; Izard et al. 1993) and the Positive and Negative Affect Schedule-Expanded Form (PANAS-X; Watson and Clark 1994) 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 12 fundamental emotions, in which surprise belongs to positive emotional factors. A principal component analysis with orthogonal varimax rotation (Izard et al. 1993) and a confirmatory factor analysis (Kotsch, Gerbing, and Schwartz 1982) 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 6-month interval (Ricard-St-Aubin et al. 2010). However, the subscale only showed moderate internal consistency (Cronbach's $\alpha = 0.65$), likely due to the small number of items (Watson and Vaidya 2003).

The same three items are also included in 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 Cronbach's $\alpha = 0.72$ to $\alpha = 0.80$. Test-retest coefficient over a 2-month interval was lower ($r = 0.23$) referring to the time frame "past week", and higher ($r = 0.52 - 0.56$) to "in general" (Watson and Clark 1994). The mean scores on surprise were the lowest compared to other affects in the PANAS-X over different samples (Watson and Clark 1994). Surprise was also the only subscale for which self-ratings did not correlate significantly with peer-ratings ($r = 0.14$). The relatively low stability and low mean scores could be explained by surprise being a transient emotional state, which relates to a specific stimulus, which was 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". 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 (Meyer et al. 1991; Reisenzein 2013), however, these measures have not been validated in a systematic way yet.

Concluding, for measuring startle there has been no systematic attempt to develop and validate a self-report measure. For measuring surprise, items in existing scales were not selected in a systematic manner and the scales were not developed nor validated to specific stimuli, even though the concept of surprise, in contrast to other affects, only makes sense referring to a stimulus or event. The goal of the current study is therefore to systematically develop measures for self-report startle and surprise. Accordingly, this study 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).

2. Method

2.1. Participants

Students and employees (N = 82) from Delft University of Technology were recruited. 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.

2.2. Procedure

2.2.1. Phase 1: Items set generation and content validity

An initial set of 14 items for surprise, and 7 items for startle (see Table 1) were formulated based on fundamental and applied literature on startle (Blumenthal 1988; Bradley, Moulder, and Lang 2005; Koch 1999; Lang, Bradley, and Cuthbert 1990; Martin et al. 2015) and surprise (Izard et al. 1993; Klein et al. 2007; Landman et al. 2017a; Meyer, Reisenzein, and Schützwohl 1997; Rivera et al. 2014), among which Items 1 and 8 for surprise were derived from the DES-IV (Izard et al. 1993) and PANAS-X (Watson and Clark 1994). The item "amazed" was not included due to it being likely associated with positive valence (i.e. "amazing" is often a positive expression). We aimed for

Table 1. The 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.[c]	85.7
10. It bewildered me.	57.1
11. It made my jaw drop.[b]	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.[b]	28.6
Initial set of items for startle	
15. It startled me.	85.7
16. It immediately made me feel scared or angry.[c]	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.[c]	85.7

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

a measure that could equally well be applied to surprise of positive, negative, or neutral valences.

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 (DeVellis 2012).

2.2.2. 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 (details in Section 2.3). The retained 19 items were arranged and presented in a randomised 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.

They were instructed to circle the number on a 5-point Likert scale that best represents their agreement with each statement. Participants indicated whether they (1) "Strongly disagreed", (2) Disagreed", (3) felt "Neutral", (4) "Agreed" or (5) "Strongly agreed" with each item using pen on the paper version of the measures. The specific event or stimulus of nine video clips that "it" in each item refers to is shown in the third column of Table 2.

2.2.3. Phase 3: the 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]?", where the specific stimulus or event in the preceding video clip (see the third column of Table 2) was inserted at [the stimulus]. The Startle-VAS and Surprise-VAS consisted of horizontal lines of 10 cm long, with tick marks at 1 cm intervals labelled "0" (left endpoint) to "10" (right endpoint). The left endpoint was additionally labelled with, "not startled at all" and "not surprised at all", respectively, and the right endpoint was 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 cross to the left endpoint in centimetres.

2.3. Video stimuli

To induce a variety of startle and surprise responses, nine video clips were selected from the internet. Table 2 summarises the description of the video clips and links where the videos can be viewed. 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.

Table 2. A description of the video clips.

ID	Description	Specific stimulus	Intended response	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	16s	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	10s	https://youtu.be/7p_iFLK9ldg
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	17s	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	17s	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	9s	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	14s	https://youtu.be/6em7aIoF5fI
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	12s	https://youtu.be/ZeJEVbpn8d8
Baseball	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	8s	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	12s	https://youtu.be/IkXMQznN5ck

The order in which the video clips were presented was counterbalanced between participants using the Latin square method (Hinkelmann and Kempthorne 2007) 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).

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.

2.5. Statistical analyses

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 factor structure of the items set was then analysed by performing a multilevel exploratory factor analysis (ML-EFA; Reise et al. 2005) with a repeated-measures design that 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 (Tabachnick, Fidell, and Ullman 2013).

Factor extraction at the within- and between-level was conducted based on: (a) eigenvalues greater than 1.0 (Kaiser's criterion; Kaiser 1958), (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 be greater than 50% (Streiner 1994). Items were excluded from the final inventory if they showed insufficient loading on factors at either the within- or between-level.

The goodness of model fit was evaluated by the model χ^2 test (Satorra and Bentler 2001), Comparative Fit Index (CFI; Bentler 1990), Tucker–Lewis Index (TLI; Marsh, Balla, and McDonald 1988), Root Mean Square Error of Approximation (RMSEA; Steiger 1990) and Standardised Root Mean Square Residual (SRMR; Muthén 1994). Acceptable model fit was supported by a non-significant χ^2, CFI and TLI values greater than 0.95, as well as RMSEA and SRMR (both within- and between-level) values below 0.10 (Kline 2023).

All analyses were performed using the Mplus software version 8.10 (Muthén and Muthén 1998–2017). Observations on Likert scales were set as ordered categorical (ordinal) variables instead of continuous for factor analysis on both levels (Bolton et al. 2022).

To test concurrent validity of the Startle-VAS and Surprise-VAS, Spearman's 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 correlation of $\rho > 0.30$ ($p < 0.01$) was considered as significant to establish validity (Cohen 2013).

3. Results

3.1. Phase 1: Items set generation and content validity

The initial set of formulated items and percentage of indicating relevance from experts are shown in Table 1. Two items were removed based on the 50% relevance criterion: Item 11 "It made my jaw drop." and Item 14 "It made me feel wide-eyed.", both originally referring to surprise. Based on the experts' open comments, wordings were changed for Item 9 (previously: "I was confused about why it happened."), Item 16 (previously: "It made me suddenly feel scared or angry."), and Item 21 (previously: "It caused a quick burst of stress or frustration in me.").

3.2. Phase 2: Multilevel exploratory factor analysis

3.2.1. Two-way ANOVA and ICCs

The two-way ANOVA results, shown in Table 3, 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 (ICC; Johnson and Koch 2011) 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 3. 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; Reise et al. 2005). The differences between ICCs also hint at possible differences in the outcomes of the following within- and between-level exploratory factor analysis.

3.2.2. Multilevel exploratory factor analysis

The final ML-EFA solution with two within (video)-level factors and two between (video)-level factors is shown in Table 4. In the within-level analysis, Items 8 and 10 were removed as they loaded on more than one factor with loadings greater than 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 greater than 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. Items loading on Factors 1 and 2 at the within-level (i.e. Factor 2 and 1 at between-level) will henceforth be referred as the Startle Inventory (Startle-I) and Surprise Inventory (Surprise-I).

In this solution, the largest factor loading for each item at the within-level ranged from 0.689 to 0.970, and 0.935 to 1.013 at the between-level, suggesting meaningful and significant factor loadings. Further evidences for the goodness of model fit are the non-significant chi-square test ($\chi^2 = 51.508$; $df = 68$; $p = 0.932$), 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

Table 3. The two-way ANOVA results and estimated intraclass correlation coefficients (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	0.333
	Video	495.33	34.38	50.32	
2. It was consistent with my expectation.	Participant	156.99	9.73	1.45	0.391
	Video	592.49	36.72	54.85	
3. I was taken aback by it.	Participant	268.40	26.07	3.54	0.159
	Video	154.49	15.00	20.37	
4. I did not understand why it happened	Participant	367.59	29.83	4.18	0.207
	Video	161.54	13.11	18.37	
5. I predicted it beforehand.	Participant	140.06	7.79	1.07	0.367
	Video	613.99	34.13	47.02	
6. Initially, it made no sense to me.	Participant	345.51	28.93	4.13	0.156
	Video	180.17	15.09	21.56	
7. I did not see it coming.	Participant	162.11	10.12	1.38	0.347
	Video	497.84	31.08	42.29	
8. It astonished me.	Participant	289.33	26.88	3.63	0.176
	Video	150.09	13.94	18.85	
9. Initially, I was confused about it.	Participant	215.39	18.18	2.36	0.223
	Video	239.81	20.24	26.29	
10. It bewildered me.	Participant	309.90	35.30	4.90	0.087
	Video	61.85	7.05	9.78	
12. I was not mentally prepared for it.	Participant	296.10	27.09	3.57	0.135
	Video	133.19	12.19	16.05	
13. It was unexpected.	Participant	160.00	10.01	1.53	0.384
	Video	601.70	37.63	57.50	
Item set for startle					
15. It startled me.	Participant	183.28	12.77	2.89	0.497
	Video	744.79	51.91	117.58	
16. It immediately made me feel scared or angry.	Participant	178.25	19.09	3.02	0.329
	Video	283.90	30.40	48.15	
17. It shocked me.	Participant	277.65	27.34	4.56	0.246
	Video	251.41	24.75	41.33	
18. It stunned me.	Participant	334.10	35.11	5.74	0.177
	Video	152.30	16.00	26.19	
19. It made me physically flinch.	Participant	174.44	12.31	2.53	0.467
	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	0.414
	Video	550.00	42.75	86.01	
21. It immediately caused stress or frustration to me.	Participant	221.57	24.46	3.85	0.276
	Video	223.59	24.68	38.83	
Average	Participant		20.50		
	Video		26.84		

Table 4. Final factor loadings of items in the multilevel exploratory factor analysis (ML-EFA).

	Within (video)-level		Between (video)-level	
Item	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.

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 ($\rho = -0.199$, $p = 0.637$).

Data from the 81 participants were also used to establish the reliability. Cronbach's α coefficients

suggested that items under each factor possess acceptable to excellent high internal consistency, $\alpha = 0.714$ to $\alpha = 0.929$ for Startle-I, and $\alpha = 0.843$ to $\alpha = 0.955$ for Surprise-I across video clips (Table 5).

Another ML-EFA was done using the data as continuous instead of ordinal under the same factor extraction criterion. The results wre 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.

3.3. Phase 3: the Visual Analogue Scales for startle and surprise

The VAS scores from one participant on Tumbler and Clouds were missing, resulting in one value less for these videos. Table 6 lists the Spearman's correlation coefficients between the Startle-VAS and Startle-I, and Surprise-VAS and Surprise-I for predicted startling and surprising stimuli, respectively. For startle, the ratings of Startle-VAS highly correlated with the scores from the Startle Inventory, ranging from $\rho = 0.778$ $\rho = 0.877$. All correlations were highly significant ($p < 0.001$). For surprise, high correlations were found between Surprise-VAS and Surprise-I, ranging from $\rho = 0.681$ to $\rho = 0.903$. All correlations were highly significant ($p < 0.001$).

Table 5. The Cronbach's α of the Startle-I and Surprise-I per video clip.

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

Table 6. Correlations between the Startle-VAS and Startle-I, and the 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).

3.4. Responses to the stimuli

For the inventories, the scores of all items in each inventory were averaged to obtain a total score ranging from 1 to 5. Responses of startle and surprise in the form of inventories and VASs over nine stimuli are shown in Figure 1 as pirate plots. These plots represent 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 averaged values (with standard deviations) for each video across participants are shown in Figure 2a based on the Startle and Surprise Inventories, and in Figure 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.

4. Discussion

The purpose of this study 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 (Blumenthal 2015; Koch 1999; Martin et al. 2015), 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 Phase2, only the Item 18 "It stunned me." was

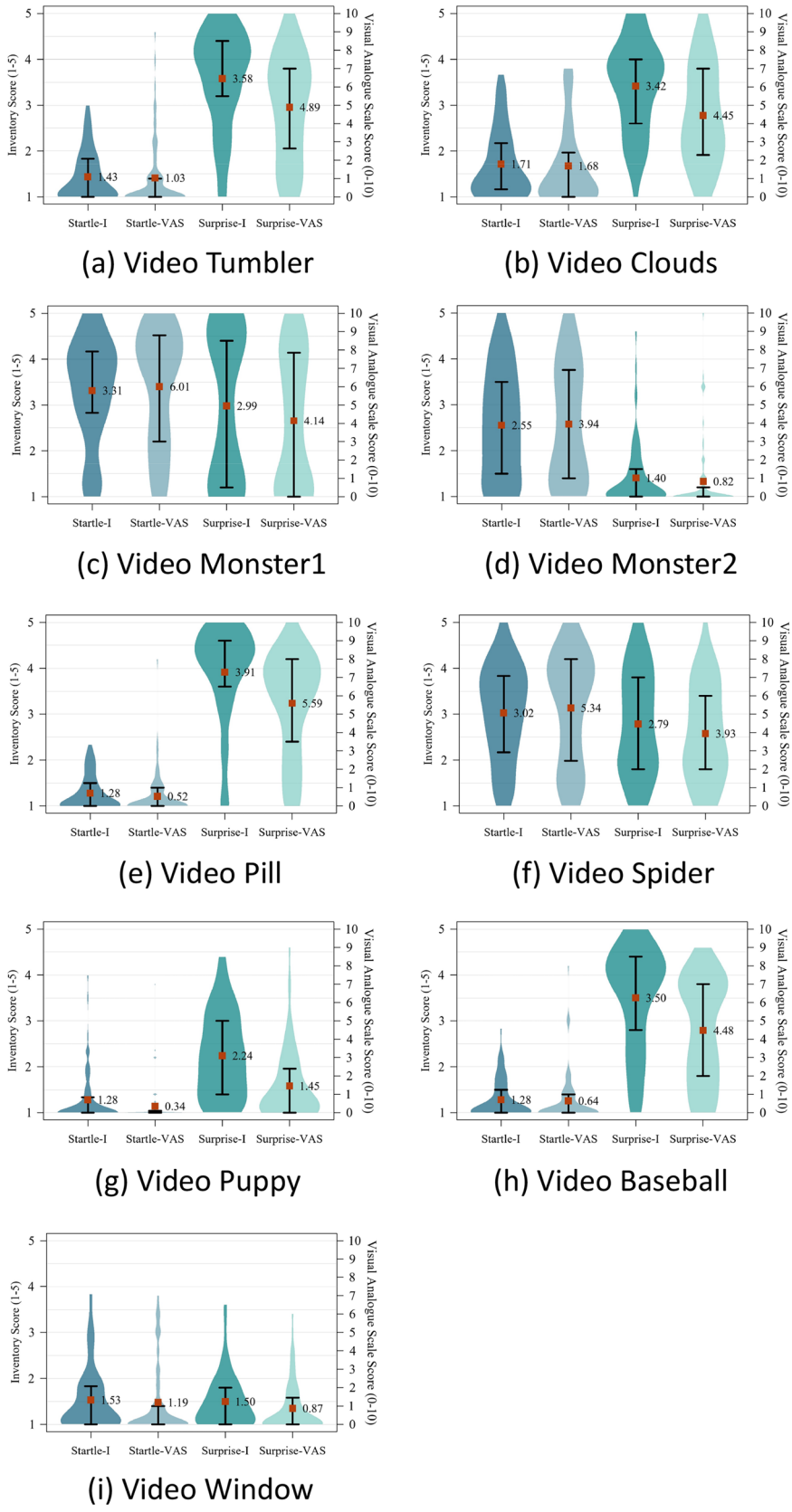
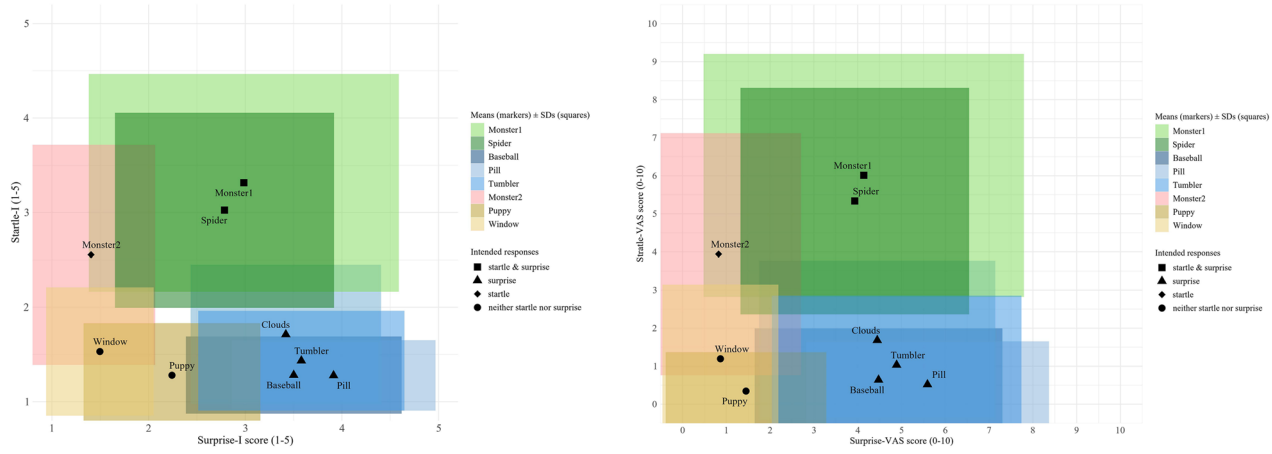


Figure 1. The responses of startle and surprise measured by inventories and VASs across nine video stimuli.



(a) The Startle & Surprise Inventories.

(b) The Visual Analogue Scales for Startle & Surprise.

Figure 2. The averaged (±SD) startle and surprise ratings of the nine video clips across participants.

removed. Items in the construct of Startle possessed high internal consistency, $\alpha_{Startle}$ = 0.714 to $\alpha_{Startle}$ = 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 3. 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, $\alpha_{Surprise}$ = 0.843 to $\alpha_{Surprise}$ = 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 characterised by a (brief) moment of confusion and requirement to reframe (Klein et al. 2007; Landman et al. 2017a; Noordewier et al. 2016). 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. We therefore recommend using real (non-scripted) events in future research on surprise 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 of the Startle-I and Surprise-I. High consistencies were found between the Startle-VAS and Startle-I, as well as Surprise-VAS and Surprise-I. Note that without further study, it is unlikely to consider Visual Analogue Scales for Startle and Surprise to be less or more accurate compared to Startle and Surprise Inventories. 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 (see Figure 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 study, we tested responses on the items set to multiple stimuli with varying degrees of startling and surprising, 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 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 (Rivera et al. 2014). In addition, the measures are useful to further test and explain some of the (ambiguous) findings in the literature. Detailed instructions for administering the Startle and Surprise Inventories as well as the Visual Analogue Scales for Startle and Surprise can be found in the instruction manual (Chen et al. 2025).

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; Ryffel et al. 2019) or behavioural markers (e.g. reaction time, micro-expressions). Additionally, operational relevance should be further tested by stimuli presented in a more ecologically-valid environment. 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.

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.

Authors contributions

All authors meet the criteria for authorship as defined by the International Committee of Medical Journal Editors (ICMJE). Jiayu Chen, Annemarie Landman, Olaf Stroosma, M. M. (René) van Paassen and Max Mulder conceived the research and designed the research methodology. Jiayu Chen was responsible for data collection. Jiayu Chen and Alexis Derumigny contributed to the data analysis and interpretation of the results. Jiayu Chen and Annemarie Landman drafted the manuscript. All authors (Jiayu Chen, Annemarie Landman, Alexis Derumigny, Olaf Stroosma, M. M. (René) van Paassen and Max Mulder) critically revised the manuscript for intellectual content. All authors approved the final version to be published and agree to be accountable for all aspects of the work.

Ethics statement

The study was approved by the Research Ethics Committee of Delft University of Technology (No.2718) and written informed consents were obtained from all participants.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received.

References

Airbus. 2025. "A Statistical Analysis of Commercial Aviation Accidents 1958–2024." https://accidentstats.airbus.com/wp-content/uploads/2025/02/20241325_A-Statistical-analysis-of-commercial-aviation-accidents-2025-links.pdf

Antony, J.W., T.H. Hartshorne, K. Pomeroy, T.M. Gureckis, U. Hasson, S.D. McDougle, and K.A. Norman. 2021. "Behavioral, Physiological, and Neural Signatures of Surprise during

Naturalistic Sports Viewing." *Neuron 109* (2): 377–390.e7. doi:10.1016/j.neuron.2020.10.029.

Banks, V., C. K. Allison, K. Parnell, K. Plant, and N. A. Stanton. 2022. "Predicting and Mitigating Failures on the Flight Deck: An Aircraft Engine Bird Strike Scenario." *Ergonomics* 65 (12): 1672–1695. doi:10.1080/00140139.2022.2048897.

Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin 107* (2): 238–246. doi: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. doi: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. doi:10.1111/psyp.12506.

Blumenthal, T. D., B. N. Cuthbert, D. L. Filion, S. Hackley, O. V. Lipp, and A. Van Boxtel. 2005. "Committee Report: Guidelines for Human Startle Eyeblink Electromyographic Studies." *Psychophysiology 42* (1): 1–15. doi:10.1111/j.1469-8986.2005.00271.x.

Bolton, M.L., E. Biltekoff, J. Wei, and L. Humphrey. 2022. "On the Level of Measurement of Subjective Psychometric Ratings." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 66, 80–84. Los Angeles, CA: SAGE Publications Sage CA.

Bradley, M. M., and P. J. Lang. 2000. "Affective Reactions to Acoustic Stimuli." *Psychophysiology 37* (2): 204–215. doi:10.1111/1469-8986.3720204.

Bradley, M. M., B. Moulder, and P. J. Lang. 2005. "When Good Things Go Bad: The Reflex Physiology of Defense." *Psychological Science 16* (6): 468–473. doi:10.1111/j.0956-7976.2005.01558.x.

Casner, S. M., R. W. Geven, and K. T. Williams. 2013. "The Effectiveness of Airline Pilot Training for Abnormal Events." *Human Factors 55* (3): 477–485. doi:10.1177/0018720812466893.

Chen, J., A. Landman, O. Stroosma, M. M. Van Paassen, and M. Mulder. 2025. *Manual for the Startle and Surprise Inventories and Visual Analogue Scales*. Delft, The Netherlands: Delft University of Technology. doi: 10.4233/uuid:4aa39791-4d21-4427-b86a-628f52d17fbe.

Civil Aviation Authority 2016. *Flight-Crew Human Factors Handbook CAP 737*. Crawley, UK: Civil Aviation Authority.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge.

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

Dreissen, Y. E. M., M. J. Bakker, J. H. T. M. Koelman, and M. A. J. Tijssen. 2012. "Exaggerated Startle Reactions." *Clinical Neurophysiology: official Journal of the International Federation of Clinical Neurophysiology 123* (1): 34–44. doi:10.1016/j.clinph.2011.09.022.

Evinger, C., and K. A. Manning. 1993. "Pattern of Extraocular Muscle Activation during Reflex Blinking." *Experimental Brain Research 92* (3): 502–506. doi:10.1007/BF00229039.

Eysenck, M. W., and M. G. Calvo. 1992. "Anxiety and Performance: The Processing Efficiency Theory." *Cognition & Emotion 6* (6): 409–434. doi:10.1080/02699939208409696.

Eysenck, M.W., N. Derakshan, R. Santos, and M.G. Calvo. 2007. "Anxiety and Cognitive Performance: Attentional Control Theory." *Emotion (Washington, D.C.) 7* (2): 336–353. doi:10.1037/1528-3542.7.2.336.

Federal Aviation Administration. 2015. *Advisory Circular 120-111: Upset Prevention and Recovery Training*. Washington: Federal Aviation Administration.

Flaten, M. A., and T. D. Blumenthal. 1998. "A Parametric Study of the Separate Contributions of the Tactile and Acoustic Components of Airpuffs to the Blink Reflex." *Biological Psychology 48* (3): 227–234. doi:10.1016/s0301-0511(98)00018-0.

Gehricke, J.-G., E. M. Ornitz, and P. Siddarth. 2002. "Differentiating between Reflex and Spontaneous Blinks Using Simultaneous Recording of the Orbicularis Oculi Electromyogram and the ElectroOculogram in Startle Research." *International Journal of Psychophysiology: official Journal of the International Organization of Psychophysiology 44* (3): 261–268. doi:10.1016/s0167-8760(02)00008-9.

Grillon, C., and J. Baas. 2003. "A Review of the Modulation of the Startle Reflex by Affective States and Its Application in Psychiatry." *Clinical Neurophysiology: official Journal of the International Federation of Clinical Neurophysiology 114* (9): 1557–1579. doi:10.1016/s1388-2457(03)00202-5.

Hajcak, G., and D. Foti. 2008. "Errors Are Aversive: Defensive Motivation and the Error-Related Negativity." *Psychological Science 19* (2): 103–108. doi:10.1111/j.1467-9280.2008.02053.x.

Hansen, J., and S. Topolinski. 2011. "An Exploratory Mindset Reduces Preference for Prototypes and Increases Preference for Novel Exemplars." *Cognition & Emotion 25* (4): 709–716. doi:10.1080/02699931.2010.496994.

Henckens, M. J. A. G., E. J. Hermans, Z. Pu, M. Joëls, and G. Fernández. 2009. "Stressed Memories:' How Acute Stress Affects Memory Formation in Humans." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 29* (32): 10111–10119. doi:10.1523/JNEUROSCI.1184-09.2009.

Hinkelmann, K., and O. Kempthorne. 2007. "Latin Square Type Designs." In *Design and Analysis of Experiments: Introduction to Experimental Design*, 373–417. Chichester, UK: John Wiley & Sons, Ltd.

Hoffman, H. S., R. R. Marsh, and C. L. Stitt. 1980. "Tests of a Principle of Reflex Modification: Modification of the Human Eyeblink-Reflex Is Independent of the Intensity of the Reflex-Eliciting Stimulus." *Animal Learning & Behavior 8* (1): 81–84. doi:10.3758/BF03209733.

Holand, S., A. Girard, D. Laude, C. Meyer-Bisch, and J.-L. Elghozi. 1999. "Effects of an Auditory Startle Stimulus on Blood Pressure and Heart Rate in Humans." *Journal of Hypertension 17* (12 Pt 2): 1893–1897. doi:10.1097/00004872-199917121-00018.

Horstmann, G. 2006. "Latency and Duration of the Action Interruption in Surprise." *Cognition & Emotion 20* (2): 242–273. doi:10.1080/02699930500262878.

International Air Transport Association 2019. *Loss of Control In-Flight Accident Analysis Report Edition 2019 Guidance Material and Best Practices*. Montreal, Canada: International Air Transport Association.

Izard, C. E., D. Z. Libero, P. Putnam, and O. Haynes. 1993. "Stability of Emotion Experiences and Their Relations to Traits of Personality." *Journal of Personality and Social Psychology 64* (5): 847–860. doi:10.1037/0022-3514.64.5.847.

Jansen, A. S., X. V. Nguyen, V. Karpitskiy, T. C. Mettenleiter, and A. D. Loewy. 1995. "Central Command Neurons of the Sympathetic Nervous System: Basis of the Fight-or-Flight Response." *Science (New York, N.Y.) 270* (5236): 644–646. doi:10.1126/science.270.5236.644.

Johnson, W. D., and G. G. Koch. 2011. "Intraclass Correlation Coefficient." In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, 685–687. Berlin Heidelberg: Springer.

Kaiser, H. F. 1958. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika 23* (3): 187–200. doi:10.1007/BF02289233.

Kinney, Lana, and David O'Hare. 2020. "Responding to an Unexpected In-Flight Event: Physiological Arousal, Information Processing, and Performance." *Human Factors 62* (5): 737–750. doi:10.1177/0018720819854830.

Klein, G., J. K. Phillips, E. L. Rall, and D. A. Peluso. 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*. New Jersey, USA: Lawrence Erlbaum Associates Publishers.

Kline, R. B. 2023. *Principles and Practice of Structural Equation Modeling* (fifth ed.). New York: Guilford Publications.

Kloosterman, N. A., T. Meindertsma, A. M. van Loon, V. A. F. Lamme, Y. S. Bonneh, and T. H. Donner. 2015. "Pupil Size Tracks Perceptual Content and Surprise." *The European Journal of Neuroscience 41* (8): 1068–1078. doi:10.1111/ejn.12859.

Koch, M. 1999. "The Neurobiology of Startle." *Progress in Neurobiology 59* (2): 107–128. doi:10.1016/s0301-0082(98)00098-7.

Kotsch, W. E., D. W. Gerbing, and L. E. Schwartz. 1982. "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*, edited by C. E. Izard, Vol. 1, 251–278. Cambridge, UK: Cambridge University Press.

Landman, A., E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder. 2017a. "Dealing With Unexpected Events on the Flight Deck: A Conceptual Model of Startle and Surprise." *Human Factors 59* (8): 1161–1172. doi:10.1177/0018720817723428.

Landman, A., E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder. 2017b. "The Influence of Surprise on Upset Recovery Performance in Airline Pilots." *The International Journal of Aerospace Psychology 27* (1-2): 2–14. doi:10.1080/10508414.2017.1365610.

Lang, P. J., M. M. Bradley, and B. N. Cuthbert. 1990. "Emotion, Attention, and the Startle Reflex." *Psychological Review 97* (3): 377–395. doi:10.1037/0033-295X.97.3.377.

Marsh, H. W., J. R. Balla, and R. P. McDonald. 1988. "Goodness-Of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size." *Psychological Bulletin 103* (3): 391–410. doi:10.1037/0033-2909.103.3.391.

Martin, W. L., P. S. Murray, P. R. Bates, and P. S. Y. Lee. 2015. "Fear-Potentiated Startle: A Review from an Aviation Perspective." *The International Journal of Aviation Psychology 25* (2): 97–107. doi:10.1080/10508414.2015.1128293.

Martin, W. L., P. S. Murray, and P. R. Bates. 2012. "The Effects of Startle on Pilots During Critical Events: A Case Study Analysis." In *Proceedings of 30th EAAP Conference: Aviation Psychology & Applied Human Factors*, pp. 387–394.

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

Meyer, W.-U., M. Niepel, U. Rudolph, and A. Schützwohl. 1991. "An Experimental Analysis of Surprise." *Cognition & Emotion 5* (4): 295–311. doi:10.1080/02699939108411042.

Meyer, W.-U., R. Reisenzein, and A. Schützwohl. 1997. "Toward a Process Analysis of Emotions: The Case of Surprise." *Motivation and Emotion 21* (3): 251–274. doi:10.1023/A:1024422330338.

Muthén, B. O. 1994. "Multilevel Covariance Structure Analysis." *Sociological Methods & Research 22* (3): 376–398. doi:10.1177/0049124194022003006.

Muthén, L. K., and B. O. Muthén. 1998–2017. *Mplus Statistical Analysis with Latent Variables User's Guide*. 8th ed. Muthén & Muthén. Los Angeles, CA, USA.

Nakagawara, V.B., R.W. Montgomery, A. Dillard, L. McLin, and C.W. Connor. 2004. "DOT/FAA/AM03/12: The Effects of Laser Illumination on Operational and Visual Performance of Pilots during Final Approach." U.S. Department of Transportation, Federal Aviation Administration. https://www.faa.gov/sites/faa.gov/files/data_research/research/med_humanfacs/oamtechreports/0312.pdf

Neisser, U. 2014. *Cognitive Psychology: Classic Edition*. New York: Psychology Press.

Noordewier, M. K., S. Topolinski, and E. Van Dijk. 2016. "The Temporal Dynamics of Surprise." *Social and Personality Psychology Compass 10* (3): 136–149. doi:10.1111/spc3.12242.

Papadimitriou, A., and K. N. Priftis. 2009. "Regulation of the Hypothalamic-Pituitary-Adrenal Axis." *Neuroimmunomodulation 16* (5): 265–271. doi:10.1159/000216184.

Reise, S. P., J. Ventura, K. H. Nuechterlein, and K. H. Kim. 2005. "An Illustration of Multilevel Factor Analysis." *Journal of Personality Assessment 84* (2): 126–136. doi:10.1207/s15327752jpa8402_02.

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

Ricard-St-Aubin, J.-S., F. L. Philippe, G. Beaulieu-Pelletier, and S. Lecours. IV IV). 2010. "Validation Francophone de L'échelle Des Émotions Différentielles IV (EED-IV) [French-speaking Validation of the Differential Emotions Scale IV (EED-IV)]." *European Review of Applied Psychology 60* (1): 41–53. doi:10.1016/j.erap.2009.05.001.

Rivera, J., A. B. Talone, C. T. Boesser, F. Jentsch, and M. Yeh. 2014. "Startle and Surprise on the Flight Deck: Similarities, Differences, and Prevalence." Proceedings of the Human Factors and Ergonomics Society Annual Meeting *58*: 1047–1051.

Rossato, C. J. L., M. A. Uphill, J. Swain, and D. A. Coleman. 2018. "The Development and Preliminary Validation of the Challenge and Threat in Sport (CAT-Sport) Scale." *International Journal of Sport and Exercise Psychology 16* (2): 164–177. doi:10.1080/1612197X.2016.1182571.

Ryffel, Chiara P., Celine M. Muehlethaler, Sandro M. Huber, and Achim Elfering. 2019. "Eye Tracking As a Debriefing Tool in Upset Prevention and Recovery Training (UPRT) for General Aviation Pilots." *Ergonomics 62* (2): 319–329. doi:10.1080/00140139.2018.1501093.

Satorra, A., and P. M. Bentler. 2001. "A Scaled Difference Chi-Square Test Statistic for Moment Structure Analysis." *Psychometrika 66* (4): 507–514. doi:10.1007/BF02296192.

Stanton, N. A., W.-C. Li, and D. Harris. 2019. "Editorial: Ergonomics and Human Factors in Aviation." *Ergonomics 62* (2): 131–137. doi:10.1080/00140139.2019.1564589.

Steiger, J. H. 1990. "Structural Model Evaluation and Modification: An Interval Estimation Approach." *Multivariate Behavioral Research 25* (2): 173–180. doi:10.1207/s15327906mbr2502_4.

Stokes, A., and K. Kite. 2017. *Flight Stress: Stress, Fatigue and Performance in Aviation*. 1st ed. London: Routledge.

Streiner, D. L. 1994. "Figuring Out Factors: The Use and Misuse of Factor Analysis." *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie 39* (3): 135–140. doi:10.1177/070674379403900303.

Tabachnick, B. G., L. S. Fidell, and J. B. Ullman. 2013. *Using Multivariate Statistics*. 7th ed. New York: Pearson.

Thackray, R. I., and R. M. Touchstone. 1983. *Rate of Initial Recovery and Subsequent Radar Monitoring Performance Following a Simulated Emergency Involving Startle (No. FAA-AM-83-13)*. Washington: Federal Aviation Administration.

Tran, T. Q., R. L. Boring, D. D. Dudenhoeffer, B. P. Hallbert, M. Keller, and T. M. Anderson. 2007. "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,* 259–263. Monterey, CA: IEEE.

Watson, D., and L. A. Clark. 1994. *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form*. Iowa, USA: The University of Iowa.

Watson, D., and J. Vaidya. 2003. "Mood Measurement: Current Status and Future Directions." *Handbook of Psychology: Research Methods in Psychology*, *2*, 351–375. Hoboken, NJ, USA: John Wiley & Sons, Inc.