Flight emotions unleashed: navigating training phases and difficulty levels in simulated flying

Ruiz-Segura, Alejandra; Law, Andrew; Jennings, Sion; Bourgon, Alain; Churchill, Ethan; Lajoie, Susanne

Publisher's version / Version de l'éditeur:
https://doi.org/10.1111/jcal.13037
Journal of Computer Assisted Learning, 2024-07-27

Received: 16 October 2023 | Revised: 15 June 2024 | Accepted: 23 June 2024

DOI: 10.1111/jcal.13037

ORIGINAL ARTICLE

Journal of Computer Assisted Learning WILEY

Flight emotions unleashed: Navigating training phases and difficulty levels in simulated flying

Alejandra Ruiz-Segura[1] | Andrew Law[2] | Sion Jennings[2] | Alain Bourgon[3] | Ethan Churchill[1] | Susanne Lajoie[1]

[1]Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada

[2]Human Factors Team at the Flight Research Laboratory, National Research Council Canada, Ottawa, Ontario, Canada

[3]Civil Aviation and Defense Security, CAE, Montreal, Quebec, Canada

Correspondence
Alejandra Ruiz-Segura, Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada.
Email: alejandra.ruizsegura@mail.mcgill.ca

Funding information
Fonds de Recherche du Québec-Société et Culture, Grant/Award Number: 286648; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: 514052-17; Social Sciences and Humanities Research Council of Canada, Grant/Award Number: 950-231835; Consortium de Recherche et d'innovation en Aérospatiale au Québec, Grant/Award Number: OPR-1618

Abstract

Background: Flying accuracy is influenced by pilots' affective reactions to task demands. A better understanding of task-related emotions and flying performance is needed to enhance pilot training.

Objective: Understand pilot trainees' performance and emotional dynamics (intensity, frequency and variability) based on training phase and difficulty level in a flight simulator.

Methods: Twenty-three volunteers performed basic flight manoeuvres. Trials were divided into three phases: Introduction (trials 1–7), session A (trials 8–15) and session B (trials 16–22). Three task difficulty levels were implemented (low, medium and high). Flying performance was evaluated using root mean square error (RMSE) and expert ratings. Emotional intensity was inferred from physiological (electrodermal activity) and behavioural (facial expressions) emotional responses. Emotional variability was calculated to understand fluctuations among multiple emotions. Emotional responses were mapped into task-relevant emotions, like sadness with boredom, and fear with anxiety.

Results and Conclusions: The most frequent facial expressions neutral, anger and surprise. Neutral and anger were interpreted as deep focus states. Surprise was likely a response to unexpected events. Flying performance and emotional dynamics varied across training phases and difficulty levels. During introduction, performance was less accurate, and emotions were less frequent. During session A, performance improved while participants experienced more physiological arousal and emotional variability. During session B, performance was the most accurate. In high-difficulty tasks, performance was the least accurate, participants expressed emotions with more frequency, more variability and higher physiological arousal. Future studies can use simulated flying tasks for trainees to familiarize with their emotional reactions to task demands expecting to improve training outcomes.

KEYWORDS
electrodermal activity, emotion, facial expression, flight, simulation, training

1 | INTRODUCTION

Aviation is a safety-critical and high-stakes profession in which errors can have significant detrimental consequences on flight safety (Baumann et al., 2011). The aviation industry invests great efforts into diminishing human error through the improved design of flight deck systems as well as effective pilot training (Code of Federal Regulations, 2023). Simulations in aviation have been a fundamental training technique since World War II (Hamman, 2004; Herrera-Aliaga & Estrada, 2022). Simulations are defined as 'a technique, not a technology, to replace or amplify real experiences with guided experiences that evoke or replicate substantial aspects of the real world in a fully interactive manner' (Gaba, 2004, pp. i2). In this text, the term 'simulator' will be used to refer to the devices that physically mimic airplanes (Gaba, 2004; Lajoie, 2021).

From an educational perspective, simulations allow learners to deliberately practice skills in a safe and authentic environment, allowing them to be prepared to transfer those skills to the real world (Azher et al., 2023; Ericsson, 2006; Lajoie, 2021). Previous research has confirmed that simulations evoke authentic emotions, like those experienced using real airplanes (Koglbauer et al., 2011; Skibniewski et al., 2015). Using simulations might be especially beneficial for pilot trainees to familiarize themselves with the stakes of flying an airplane, and habituate to their emotional reactions in a safe context and controlled environment (Vine et al., 2015). However, the traditional measurements of emotions are conflicting with the objectives of flight training. Emotions are traditionally measured using questionnaires (Duffy et al., 2016; Harley, 2016). But, a common critique of using questionnaires is that they can be distracting, and the reporting of the emotion is differed from the moment that the emotion was experienced (Harley, 2016). Moreover, the aviation industry is recognized for investing efforts into preventing errors that may be due to distractions, faulty decision making and intense emotions experienced when flying (Hamman, 2004). The objective of this study is to explore performance and emotional changes while performing manoeuvres in a flight simulator across training phases and difficulty levels.

Previous research in aviation has been interested in understanding the relationships between affective responses (i.e., mood, stress and emotions) on flying performance when using flight simulators. The most common theoretical frameworks used to understand affect in flight training include stress and coping theory (Lazarus & Folkman, 1984) (e.g., Sassenus et al., 2022; Tichon et al., 2014), situational awareness theory (Endsley, 2000) (e.g., Gray et al., 2016; Kinney & O'Hare, 2020) and the circumplex model of affect (Russell, 1980, 2003) (e.g., Rosa et al., 2021, 2022), the latter being the only of these frameworks that focuses on emotions. In this manuscript, moods are excluded since these have a low physiological arousal and longer lasting affective states, potentially persisting for days, and thus the relation to performance in a standard commercial flight might be unclear (lasting a maximum of 14–16 h) (Code of Federal Regulations, 2023; European Commission, 2013; Gross, 2015). Stress is excluded since it occurs in taxing situations, being experienced as displeasing, and it is not stimuli-specific; in this manuscript, we attempt to understand a range of positive and negative emotions (DeMaria et al., 2010; Landman et al., 2017). Since emotions are context and stimuli specific (Gross, 2015), this study proposes to use a guiding theory that focuses on emotions that occur during the training process. We follow Pekrun's (2019) control-value theory of achievement emotions, which has been used to understand training of pilots (Li & Lajoie, 2021) and other high-stakes professions like nursing (Azher et al., 2023; Harley et al., 2023) and medicine (Artino et al., 2012; Lajoie et al., 2023; Nomura et al., 2021).

1.1 | Emotions and achievement performance

Emotions are defined as quick-changing affective states that last from seconds to minutes and have a clear object focus (Harley, Pekrun et al., 2019). Emotions are directly connected to a specific stimulus, rather than being a broad feeling as is the case of mood (i.e., feeling down) or stress (Gross, 2015). Moods differ from emotions since they are longer lasting, sustaining potentially for days, and the feeling is overall less arousing (Gross, 2015). Stress is characterized as a displeasing feeling experienced during taxing situations, whereas emotions have a range of pleasantness (Gross, 2015). Therefore, in this manuscript, it is argued that emotions might provide more rich information about the range of pleasing and displeasing feelings that are triggered during a flying task.

Achievement emotions refer to those emotions that occur in situations that can lead to success or failure, such as learning how to fly a plane (Pekrun, 2019). Achievement emotions can be classified according to their valence and arousal (Pekrun, 2019). Valence represents the subjective feeling of pleasantness (positive or negative), and arousal refers to the physiological activation or deactivation associated to the emotion (Pekrun, 2019; Russell, 2003). Emotions can be grouped according to their combination of valence and arousal, with each group having a different relationship to performance (Harley, Pekrun et al., 2019; Pekrun, 2019). Positive-activating emotions (like joy) facilitate learning and correlate to more accurate performance by associating to more flexible thinking and motivation towards the task (Pekrun & Perry, 2014). Negative-deactivating emotions, like sadness and boredom, are less beneficial for success because they cause disengagement and lead to the avoidance of similar future situations due to the displeasing sensation (Pekrun & Perry, 2014). Negative-activating and positive-deactivating have a less clear connection to performance. Negative-activating emotions (i.e., anxiety, anger and confusion) might cause displeasing feelings towards the task, yet the physiological activation may trigger the person to act and seek solutions (Pekrun & Perry, 2014). Conversely, positive-deactivating emotions, like relief, lead learners to invest less energy on the task, yet the pleasant feeling will motivate learners to re-engage in similar future tasks (Pekrun & Perry, 2014).

We recognize the inherent co-occurrence of emotions and cognition, as shown in our guiding theory (Pekrun, 2019); however, we emphasize that this manuscript is focused on emotions. We see emotions as autonomous reactions that guide instinctive actions (Damasio, 2005). In the context of flight training, as pilots are highly concentrated in the task, we use autonomous physiological and behavioural reactions, namely emotions, to infer quick decision-making (Damasio, 2005).

1.2 | Emotions' dynamics

Emotions are dynamics processes (D'Mello & Graesser, 2012; Zheng et al., 2023). Emotions are conventionally studied according to their tendency like categories (e.g., joy, boredom) and duration (Li, Zheng, Lajoie, 2021). When learning a new skill, emotions might change constantly as the task evolves (Martins, 2016; Pekrun, 2019). Recent advances in learning sciences suggest the need to account for emotions' dynamics, defined as the patterns and regularities of fluctuations of emotions over time (Houben et al., 2015; Kuppens, 2015). Dynamic features of emotions are observed in frequency, intensity and variability (Bailen et al., 2019; Krone et al., 2018; Zheng et al., 2023). The frequency of emotions reflects how many times an emotion is experienced in a designated period of time (Bailen et al., 2019). Emotional intensity shows 'the strength and magnitude of emotional response' (Bailen et al., 2019, pp. 64). Emotional variability emphasizes the fluctuations among multiple emotions over time (Thompson et al., 2012; Trull et al., 2008). Notably, there emotional features are complementary and can overlap. For example, a participant can answer a question reporting that they experienced a positive emotion scoring on 4 out of 10 strengths, showing that they experienced an emotion (having a frequency of one) but the intensity was low. Similarly, emotional variability can account for fluctuations of multiple emotions as they occur (i.e., frequency) or filtering only emotions with an intense presence (Zheng et al., 2023). Emotion dynamics are becoming a phenomenon of interest in learning and educational context as technology to detect nuanced and quick changes of emotions advances (D'Mello & Grasser, 2012; Duffy et al., 2016). For this manuscript, to avoid confusion, we use arousal and intensity as different terms. Arousal refers solely to the physiological changes associated to emotions: physiological activation denotes increases in arousal, whereas physiological deactivation shows decrease in arousal (Pekrun, 2019). Commonly, arousal is interpreted as physiological intensity, for instance outstanding peaks of physiological arousal (i.e., skin conductance responses [SCRs]) (Harley, Pekrun et al., 2019). In this manuscript, we will also refer to intensity of behavioural changes such as the degree of presence of facial expression of emotions. Emotional intensity is used as an umbrella term to detect changes in the degree of emotional responses that could come from different measures, such as self-reports, physiology or behaviour.

Research exploring the use of flight simulations for training has mainly focused on the role detrimental impact of negative-activating affective states, like stress and anxiety (i.e., Allsop & Gray, 2014; Hart, 2006). In the case of research exploring learning, the focus is extended to include both negative and positive valence; however, emotions tend to be studied as discrete states, rather than pairing emotional fluctuations according to the task demands (Duffy et al., 2016; Pekrun & Perry, 2014). For that reason, in this manuscript, we argue that emotions dynamics can visualize the functionality of emotions as an adaptive response that go beyond emotional valence (i.e., (un)pleasantness) (Kashdan & Rottenberg, 2010). Particularly, this study contributes to understanding emotion dynamics in the context of pilot training using continuous and non-invasive measurements. It is expected that the results of this study identify the patterns of beginner pilot trainees' emotion dynamics and their potential function in parallel to performance accuracy changes.

This study is guided by a multimodal approach combining physiological and behavioural responses to infer single-variable emotions (Han et al., 2020; Harley et al., 2015). Emotional intensity (strength of single-variable emotional responses, Zheng et al., 2023) will be inferred from the SCRs of electrodermal activity (EDA) and single dominant facial expressions. Emotional variability will be inferred from fluctuations among multiple dominant emotions (Figure 1). This study focuses on authentic reactions during the learning process, and emotions are not artificially induced.

To inform our understanding of emotional changes across training phases and difficulty levels, the section below reviews and discusses findings of previous studies exploring affective changes and performance as trainees go through flying simulations.

1.3 | Emotional changes as flight progresses and according to difficulty levels

Other studies in flight training have been interested in understanding emotional as the flight evolves and comparing difficulty levels. However, to our knowledge, both dimensions (time and difficulty) have not been examined together in a same study, and performance is explored only on some occasions.

Previous studies have confirmed that emotions change as the flight progresses, having distinctive effects on flying performance. Anxiety can have contrasting effects on flying performance, it can associate to less optimal cognitive processing in secondary tasks, but anxiety might not impact performance significantly (Hidalgo-Munoz et al., 2018). Long flights related to participants reporting more negative-deactivating emotions, a decrease in positive-activating emotions, and an increase in heart rate variability (Rosa et al., 2021, 2022). However, positive emotions increase when pilots are provided with more opportunities to practice (Koglbauer et al., 2011), showcasing the need to evaluate emotions in context (Lajoie, 2021; Pekrun, 2019).

More studies reported increases in negative-activating emotions with increases in flight difficulty level. For instance, pilots and pilot trainees tend to have more displeasing and unstable affective experiences during difficult flying manoeuvres (Skibniewski et al., 2015; Tichon et al., 2014). However, the relationship between displeasing affect and flying performance has only been hypothesized to impact consequent performance as some studies had relied on pilots' perceptions of stereotypical scenarios, without assessing an actual flight (Hart & Bortolussi, 1984).

In the context of flight training, only Gaetan et al. (2015) discussed the implications of emotional variability in a descriptive

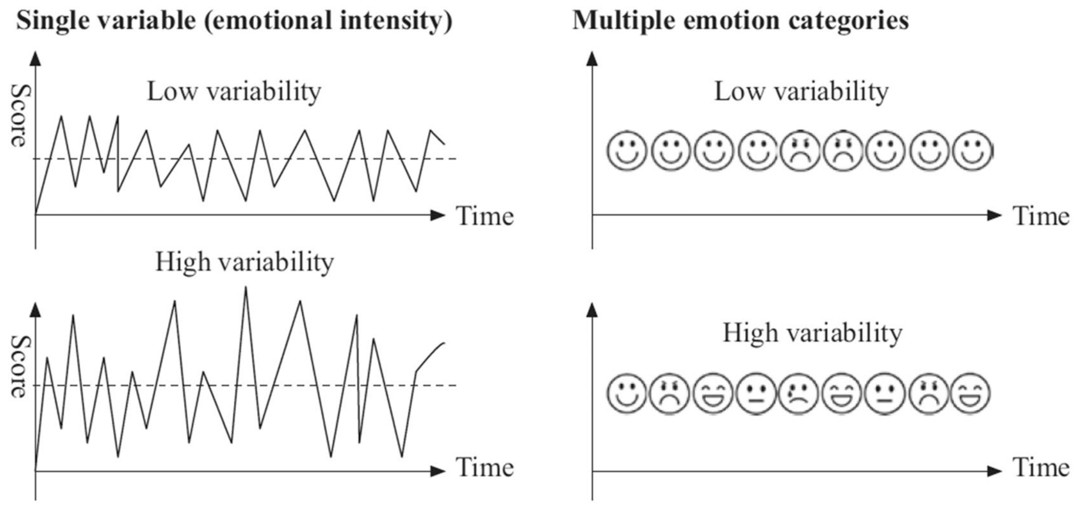

FIGURE 1 Illustration of emotional dynamics. From '*A review of measurements and techniques to study emotion dynamics in learning*' by Zheng et al. 2023 (p. 10). Copyright 2023 Springer. Reproduced with permission.

manner; however, to our knowledge, pilots' emotional variability has not been previously quantified. Gaetan et al. (2015) examined the interaction between emotions and difficulty level during simulated flying tasks by describing patterns of participants according to similarity to expert performance. They found that the physiological arousal (i.e., muscle tension and EDA) of expert-like pilots increased with difficulty levels, and these participants reported experiencing more positive emotions with low intensity tasks. On the contrary, novice-like participants had increases in physiological arousal aligned with perceived workload, and they reported experiencing positive and negative emotions with high intensity and high variability. Intermediate pilots were like novices in presenting increases in physiological arousal aligned with perceived workload but reported more positive emotions with low intensity, similar to expert reports. Although the authors did not report how they measured expertise, we believe that they pose an interesting new dimension to understanding the experience of pilots that can account for emotional variability.

Research in medical diagnostics has investigated the impact of emotional variability when diagnosing virtual patients (Li, Zheng, Lajoie, 2021; Li Zheng, Lajoie, Wiseman, 2021). Findings show that medical students experienced more emotional variability during high-difficulty tasks as compared with low-difficulty ones, and high performers, who reached a correct diagnosis, experienced lower emotional variability independently of task difficulty (Li, Zheng, Lajoie, Wiseman, 2021). Furthermore, better diagnostic performance, similar to an expert solution, negatively correlated to medical students' emotional variability (Li, Zheng, Lajoie, Wiseman, 2021).

1.4 | Current study

This study attempts to examine the evolution of pilot trainees' performance and emotion dynamics according to training phase and difficulty level in a simulated flying task. We expect to contribute to the literature by exploring flying performance relative to behavioural and physiological emotional dynamic changes across time and difficulty level. Previous aviation research has explored emotional dynamics focusing on unique emotional expressions, such as intensity inferred from frequency of emotions and physiological responses (Li & Lajoie, 2021; Zheng et al., 2023). To our knowledge, this would be the first study in aviation to quantitatively explore fluctuations among multiple emotions (i.e., emotional variability). We pose the following research questions and hypotheses. Figure 2 shows a diagram summarizing our approach.

1. How did (a) flying performance, (b) intense single expressions of emotions, inferred from behavioural (facial expression) and physiological responses (SCR); and (c) emotion variability (inferred from facial expression) differ across training phases?
2. How did (a) flying performance, (b) intense single expressions of emotions, inferred from behavioural (facial expression) and physiological responses (SCR); and (c) emotional variability (inferred from facial expression) differ across difficulty levels?

1.4.1 | Hypotheses

1. Across training phases: Training phases were consecutive and divided as introduction (feedback provided), sessions A and B (independent execution without feedback).
    a. Performance and emotions will vary across training phases. Performance will be better in the final practice (session B) compared with introduction and session A.
    b. Positive emotions (i.e., happy) will be more frequent during introduction compared with session A and session B (Rosa et al., 2021, 2022) since participants might be more energetic and motivated at the beginning of the task. Negative-activating

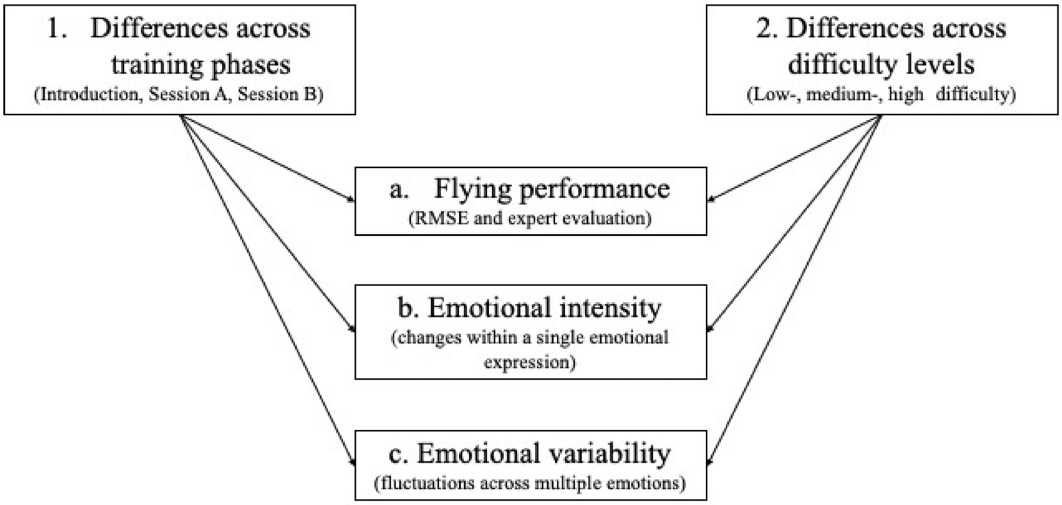

FIGURE 2 Visual representation of hypotheses and research questions. RMSE, root mean square error.

emotions (i.e., anger, disgust and fear) and SCR will be more frequent in introduction, since participants are expected to be more engaged at the beginning of the task, and session A because participants are expected to have more difficulties solving the tasks independently (without feedback), compared with session B, when participants will have had sufficient practice and might be more tired (Hidalgo-Munoz et al., 2018; Koglbauer et al., 2011). Negative-deactivating emotion (i.e., sad/ boredom) will be more frequent in session B compared with introduction and practice A because participants might be more tired at the end of the task (Rosa et al., 2021, 2022). Surprise is known to occur after an unexpected event while flying (Kinney & O'Hare, 2020), therefore, we expect surprise to be higher during introduction phase due to the novelty of the task (Landman et al., 2020).

c. Emotional variability (i.e., fluctuations among multiple emotions) will be higher during introduction and first session A, as we expect participant to be less emotionally engaged in session B. Practice and familiarity might imply that the task is less stimulating, causing less emotional fluctuations (Rosa et al., 2022).

2. Across difficulty levels: Performance and emotions will vary across task difficulty.
   a. Performance will be poorer in high difficulty, as compared with low- and medium-difficulty levels (Li, Zheng, Lajoie, Wiseman, 2021).
   b. Positive emotions will be higher in low-difficulty as compared with medium and difficult levels (D'Mello & Graesser, 2012; Pekrun, 2019). Negative-activating emotions will be higher in difficult tasks, compared with low and medium levels (Hart & Bortolussi, 1984; Skibniewski et al., 2015; Tichon et al., 2014). Negative-deactivating emotion will be higher during low-difficulty tasks, compared with medium and difficult tasks since low-difficulty will require less cognitive engagement (D'Mello & Graesser, 2012). We expect that difficult tasks will cause more cognitive dissonance from expectations, compared with easy and medium-difficulty tasks, causing participants to express more surprise during difficult tasks (D'Mello & Graesser, 2012).
   c. Emotional intensity (single-variable emotional reactions), inferred from SCRs and emotional variability (fluctuations among multiple emotions) will increase with difficulty level, being higher in high-difficult tasks compared with low and medium difficulty (Gaetan et al., 2015; Li, Zheng, Lajoie, 2021; Skibniewski et al., 2015).

2 | METHODS

2.1 | Participants

This study was part of a larger project with the objective of measuring cognitive and affective processes of ab initio pilot training using flight simulations. Only relevant methods are presented. Volunteers were recruited to mimic junior pilot trainees with little to no experience flying airplanes (Marques et al., 2023). The only requirements for applying to become a student pilot in North America is being 14 years of age or older, being able to read, write, speak and understand English and passing a medical evaluation (FAA Department of Transportation, 2003; Transport Canada, 2019). These same requirements were used to recruit participants for this project. The larger project implied using multiple physiological sensors (i.e., electroencephalogram, heart rate and EDA) and participating in a two-day data collection, for an approximate total of 8 h for the experiment, in addition to commuting. For those reasons, the recruitment was performed by convenience, targeting potential volunteers who had the time flexibility and willingness to participate with knowledge of the experiment demands.

Following ethics approval, 23 volunteers ($M_{age} = 28.96$, $SD = 4.68$) were recruited from a large North American city; 12 self-identified as females (52.2%) and 11 as males (47.8%). Participants had diverse educational backgrounds, including high school degree currently studying to obtain college and a bachelor's degree ($n = 2$), bachelors ($n = 9$), and masters ($n = 12$) degrees. The participants who had University degrees were mostly from science, technology, engineering, and mathematics STEM fields ($n = 11$), followed by studies in finance and accounting ($n = 3$), psychology-related domains ($n = 3$), nutrition ($n = 2$) and neuroscience ($n = 1$). One participant graduated from aviation school obtaining an US Federal Aviation Agency (FAA) commercial pilot rating, the prerequisite to obtaining a commercial pilot certificate (FAA Department of Transportation, 2003) with experience using Microsoft Flight Sim and X-Plane. Other five participants reported having experience using flight simulations: four of them had used a flight simulation one time, and one did not specify. The facial expression of the participant with the commercial pilot rating was not recorded, thus, not included in the analyses using facial expression. Participants who were screened into the study reported to be free from any medical condition that could limit their participation. Participants signed a consent form informing them of the purpose of the study, emphasizing voluntary participation.

Due to the heterogeneity of the participants, non-parametric statistical analyses were conducted to identify if the distribution of the variables differed according to participants' backgrounds. Independent-samples Kruskal–Wallis' test show that the distribution of the four performance variables (heading and altitude RMSEs and heading and altitude scoring) had the same distribution across educational categories (i.e., completed high-school, bachelor's and master's degrees). Similarly, Mann–Whitney *U* tests demonstrated that the performance variables had the same distribution regardless of previous experience using flight simulations.

G*power software application was used to conduct a power analysis a priori to define the sample size for conducting within subjects repeated measures ANOVAs. For comparing three groups (i.e., three training phases or three difficulty levels) and 13 measures (i.e., four measures of performance, seven emotions, one measure of skin conductance and one measure of emotional variability), and their corresponding post hoc analyses, revealed that a sample size of 12 participants is needed to achieve a power of 0.80, assuming a significance level of 0.05 and a medium effect size of 0.25 (Cohen, 1988; Faul et al., 2007). Using the same statistical parameters for comparing the changes in facial expression of the seven emotions across each training phase and each difficulty level, G*power analysis revealed that a sample of 17 participants would be need.

2.2 | Apparatus

The experiment consisted of training ab initio pilots to perform flying twizzles (i.e., basic flying manoeuvres) in a fixed-base simulator designed and operated by Marinvent Corporation. The cockpit included a control yoke, throttle and pedals, and a screen, which showed an aircraft primary flight display (see Figure 3). The throttle and pedals were operated by an autopilot. Participants used the yoke to control aircraft roll, by turning the yoke left and right, and aircraft pitch, by moving the yoke forward and backwards. Aircraft bank angle

FIGURE 3 Flying simulator cockpit.

primarily affects aircraft turn rate and is used to control heading, while the aircraft pitch angle is used to set climb/descent rates and thus control altitude. The primary flight display was rendered using X-Plane 11, which is a flight simulation software package designed to reflect the behaviour of real aircraft (Figure 4a, Laminar Research, 2022). The investigating team installed a camera on top of the primary flight display to record participants' facial expression. A tablet was set up on the left side of the control wheel for participants to read instructions and answer questionnaires (see Figure 4b).

2.3 | Flying tasks

The twizzles were designed by an expert pilot instructor. Participants were required to perform changes in direction (i.e., turns), and altitude (climbs and descents) while only referring to instrument indications with no outside visual cues. Flight manoeuvres are dynamic processes, where changes in one metric will affect another. For example, increasing altitude will decrease airspeed. Therefore, in this experiment, airspeed was set to autopilot programmed to maintain a steady 250 knots.

The flying manoeuvres had three difficulty levels: low, medium and high. Low difficulty manoeuvres required a change in one axis (roll or pitch) to achieve a target (heading or altitude) and then maintenance of the new heading or altitude. Medium-difficulty manoeuvres required a change in one axis (heading or altitude) while maintaining the other axis, followed by a reversal to the original flight condition (i.e., starting heading and altitude), and maintenance of the original conditions. High difficulty manoeuvres required simultaneous changes in two axes: a change with reversal in one axis and a change without reversal in the second axis. See Table 1 for example manoeuvre instructions.

2.4 | Procedures

Upon arrival, the goals of the experiment were explained to the participants before they signed the consent form. Participants filled out questionnaires regarding demographic information and relevant

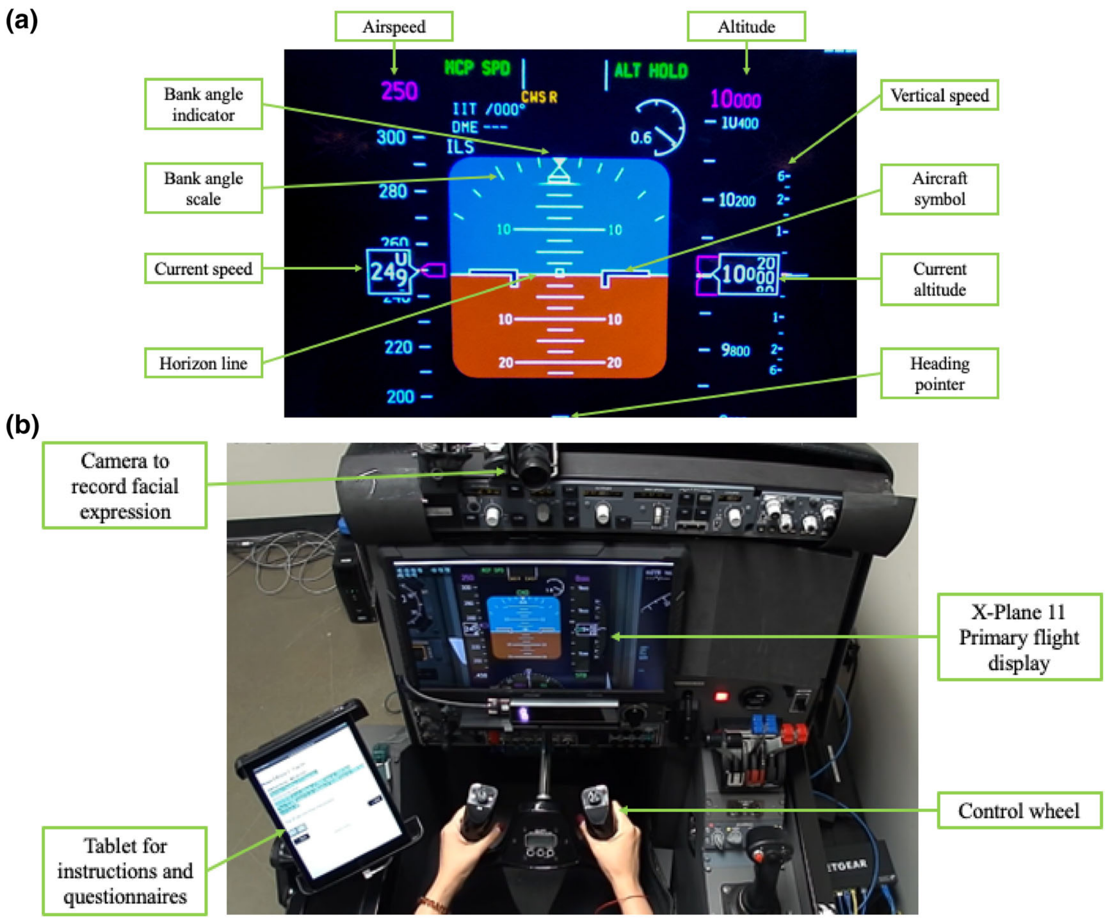

FIGURE 4 X-Plane 11 primary flight display.

TABLE 1 Sample instructions for manoeuvres according to difficulty levels.

Difficulty level	Instructions
Low	Maintain altitude at 10,000 feet At the same time: Turn LEFT at 30° AOB to a heading of 240° and roll out on a steady heading
Medium	Maintain altitude at 10,000 feet At the same time: Turn LEFT at 30° AOB to a heading of 300°; then turn RIGHT at 30° AOB back to a heading of 0° and roll out on a steady heading
High	DESCEND to an altitude of 9000 feet at 1000 fpm and level off At the same time: Turn LEFT at 30° AOB to a heading of 300°; then turn RIGHT at 30° AOB back to a heading of 0° and roll out on a steady heading

previous experience. Participants received instructions on how to use the cockpit and perform flying manoeuvres, emphasizing aviation-related language. Afterwards, researchers set up electrodes to detect EDA. The experiment was divided into three main phases: introduction (seven trials), session A (eight trials) and session B (seven trials), for a total of 22 trials. Each trial had a 30 s baseline ('maintain straight and level' at 10,000 feet altitude and a heading of 0°), followed by a 90 s twizzle manoeuvre. The X-Plane 11 instrument display and aircraft state were re-set before every trial: participants would start from 10,000 feet altitude, 250 kts (automatically controlled) airspeed, 0° bank angle and 0° heading. The introduction phase was the same for all participants, with a sequence of increasing difficulty: four low-difficulty, two medium-difficulty and one high-difficulty manoeuvres. During the introduction phase, participants received feedback from a trained researcher to confirm that participants understood the instructions and how to use the control wheel and displays to complete the task. During sessions A and B, the task difficulty order was randomized, and participants completed five trials for each difficulty level. During sessions A and B, participants no longer received feedback from a researcher. Participants could read the twizzle manoeuvre instructions at their own pace and start the task when they were ready. Participants were allowed to take breaks between each phase. The full procedure, see Table 2, took around 6 h.

2.5 | Measures

2.5.1 | Flying performance

Flying performance was assessed from aircraft state data contained in the X-Plane 11. The metrics of altitude and heading were selected to

TABLE 2 Sample procedures of experiment.

Sign consent					
Demographics					
Video-training					
Guided hands-on familiarization					
Training phase		**First practice**		**Second practice**	
1.	Low	8.	Medium	16.	Medium
2.	Low	9.	Low	17.	Low
3.	Low	10.	High	18.	High
4.	Low	11.	Low	19.	Medium
5.	Medium	12.	High	20.	High
6.	Medium	13.	Medium	21.	Medium
7.	High	14.	Low	22.	Low
		15.	High		

evaluate task performance in pitch and roll axes. Performance was measured using two metrics: (1) flying error and (2) expert rating.

Flying error

RMSE is the most common measure to assess flying performance in aviation (Allsop & Gray, 2014; Gray et al., 2016). RMSE is calculated by subtracting actual altitude or heading from the target altitude or heading of each time point, squaring the error and computing the square root of mean error. Since the tasks in this experiment are continuous and, on some occasions require reversal paths, sinusoidal target functions were used to connect the starting, middle and endpoints for instructed heading and altitude across the 90 s allocated to each twizzle (Jennings et al., 2024). Thus, the target metric accounted for expected continuous changes for an idealized flight path; see Figure 5 for a visual representation. RMSE reflects error regarding the deviation between ideal flight and actual performance: a larger number reflects more deviation from ideal flight, thus more error.

Expert rating

Expert aviation instructors rated participants performance after observing graphs with their flying trajectories (see Figure 6). Scores ranged from one to four, one represented very low accuracy compared with the instruction, and four would represent high accuracy compared the target metric. Zero was set when the task was not carried out. Consequently, a higher score would represent higher performance. This kind of scoring is commonly used for rating trainees' performance by the industrial partner in this project. The scoring is usually done with the instructor observing trainees' performance in the simulation. However, this study uses an innovative approach whereby instructors would examine and score graphs that summarized the individual participant's flying path. Preliminary results of this new approach (evaluating flying path in graphs) demonstrate alignment with state data of the aircraft, and more studies are being performed to confirm its reliability (Jennings et al., 2024).

2.5.2 | Emotional intensity inferred through skin conductance responses

Electrodermal activity shows changes in the autonomous nervous system from changes in skin conductance (Braithwaite et al., 2015). The main features of EDA include skin conductance levels (SCLs) and SCRs. SCL shows continuous changes of skin conductance and general autonomic arousal (Braithwaite et al., 2015). SCR shows rapid and intense changes (significantly higher than the individual's baseline) in physiological arousal, creating peaks in skin conductance (Braithwaite et al., 2015). This study focuses on SCR as a meaningful indicator of physiological arousal to infer emotional intensity, and engagement towards learning and solving the task (Hardy et al., 2013). SCL, as a general autonomic response, might not provide additional information about emotional and cognitive engagement in the task (Boucsein et al., 2012).

EDA was recorded using BioSemi. Electrodes were positioned on the right hand, one at the hypothermal region of the hand and the second at the wrist, to detect changes in sweat to infer skin conductivity (see Figure 7). EDA was processed using NeuroKit2, an open-source python toolbox designed for neurophysiological signals processing (Makowski et al., 2021). In the case of EDA, raw data is used as the input, and NeuroKit2 returns the filter signal, phasic components, SCR onsets, indexes and amplitudes (Makowski et al., 2021). NeuroKit2 includes the convex optimisation approach (cvxEDA) accounting for white Gaussian noise for controlling for artefacts and errors (Greco et al., 2016). Since the raw data is used, a personalized baseline is automatically created for each participant and filters out components of EDA, including phasic components, and SCR onsets, indexes and amplitudes (Makowski et al., 2021). We concentrated on SCRs since they are significantly higher peaks that can be interpreted as rapid and meaningful changes resulting from psychological engagement, and thus interpreted a measure of emotional intensity (Boucsein et al., 2012; Harley et al., 2013; Harley, Jarrell et al., 2019). See Figure 8 for a visual representation of Neurokit analysis and extraction of SCR.

2.5.3 | Emotional intensity inferred through facial expression of emotions

Facial expressions of participants were recorded throughout the experiment using a camera mounted on top of the screen showing the primary flight display. Participants who used corrective lenses were required to use contact lenses to better identify facial expression. For identifying emotions, videos were processed using FaceReader 6.0 (Loijens et al., 2015). FaceReader is a software application trained to detect and analyse facial expressions (Loijens & Krips, 2021). FaceReader classifies emotions based in three main stages (Loijens & Krips, 2021). First, FaceReader detects the face using the Viola–Jones algorithm, a deep-learning algorithm (Viola & Jones, 2004; Zafeiriou et al., 2015). In the second stage, FaceReader uses an active appearance model, based on deep neural networks, to map and record 500 key points of the face (Cootes & Taylor, 2001;

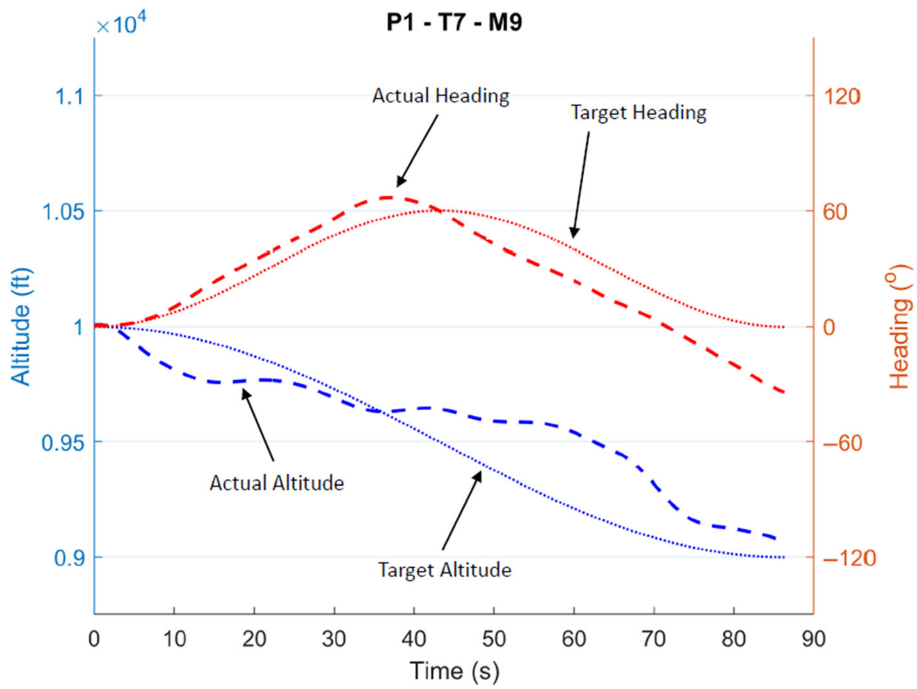

FIGURE 5 Example of root mean square error calculation. The instruction of this example was as follows: 'descend to an altitude of 9000 feet at 1000 fpm and level off. At the same time: turn right at 30° angle of bank to a heading of 60°. Then, turn left at 30° angle of bank back to a heading of 0° and roll out on a steady heading'.

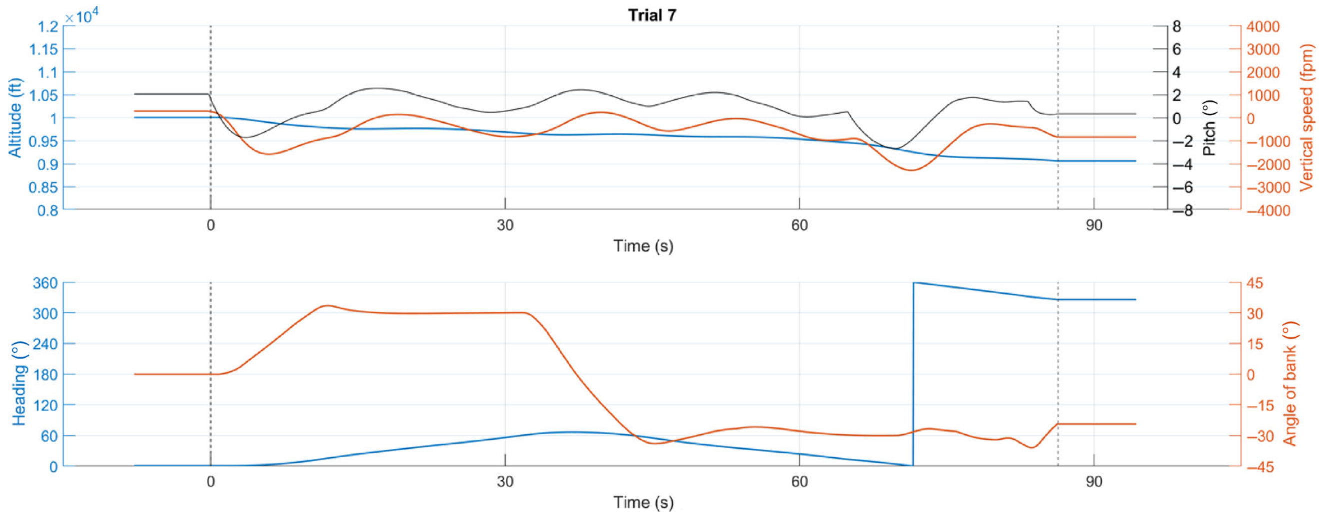

FIGURE 6 Sample graphs for expert analysis. The instruction of this example was as follows: 'descend to an altitude of 9000 feet at 1000 fpm and level off. At the same time: turn right at 30° angle of bank to a heading of 60°. Then turn left at 30° angle of bank back to a heading of 0° and roll out on a steady heading'. Expert rated altitude in a score of 4 and heading with a score of 3.

Loijens et al., 2015). Such key points are later merged using Principal Component Analysis to calculate a vector comparing the dimensionality of the face with a model (Loijens et al., 2015). In the last stage, the vector is used as the input in an artificial neural network (ANN) to identify the emotions classification (Loijens et al., 2015). FaceReader's ANN was trained using 10,000 manually annotated images following the Facial Action Coding system (Ekman & Rosenberg, 2005). The ANN is trained to classify facial expression of seven basic emotions: anger, disgust, happy, neutral, sad, scared and surprised with an accuracy of 90% (Ekman, 1992; Loijens et al., 2015). Due to individual differences in facial expression, FaceReader was calibrated for each participant by selecting a representative neutral facial expression before performing the analysis.

The sample rate used for FaceReader was 30 samples per second (Loijens et al., 2015). For each time point, the intensity of each emotion is recorded in a scale of 0 to 1, in which 0 indicates that the emotion is not present, and 1 implies that the emotion is fully detected (Loijens et al., 2015). For the current study, we used the frequency of dominant emotions, using the state log output. Dominant emotions are recorded in the log if they meet the following criteria: (1) emotions are sustained for more than 0.5 s and (2) emotions have a higher intensity (or presence) compared with the other emotions

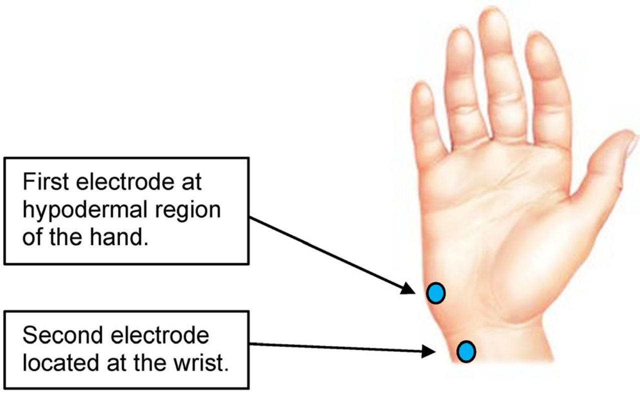

FIGURE 7 Sample location of electrodes to detect electrodermal activity.

(Loijens et al., 2015). For that reason, we argue that the dominant emotions output represents intense and meaningful emotions.

The frequency of dominant emotions was counted per training phase and per difficulty level. Such frequency was used in a first analysis to identify difference among type of emotions expressed across training phases and difficulty levels. Additionally, the frequency of emotions was used in a second analysis to calculate emotional variability.

Notably, our guiding theory is the control-value theory of emotions (Pekrun, 2019); however, FaceReader, categorizes emotions based on Ekman's (1992) basic emotions theory. Therefore, we rely on Harley et al. (2015) empirical pairing of facial expression of achievement and basic emotions to interpret our results: frustration is aligned with anger, confusion with disgust, joy with happiness, anxiety with fear and boredom with sadness. Surprise cannot be paired directly as an achievement emotion; however, it is included for its relevance in flight training and performance (Landman et al., 2017). Although the definition of neutral as an emotion is debatable (Lajoie et al., 2021; Russell, 2003), this study will follow Harley et al. (2012) proposal of accounting neutral as a baseline state, indicating that trainees are not emotionally distracted and can learn and, neutral will facilitate the identification of fluctuation of emotions.

2.6 | Emotional variability using facial expressions

Emotion variability refers to the fluctuations in emotional states and provides a dynamic insight into emotions above and beyond the frequency of emotions. In this manuscript, we calculate the variability, or degree of randomness of different occurring emotions, of intense (i.e., dominant) emotions using Shannon's (1948) entropy formula (Jack et al., 2014; Li, Zheng, Lajoie, 2021):

$$h(p_1,\ldots,p_a) = -\sum_{i=1}^{a} \text{pi} \log_2(p_i)$$

In which, p_i is the probability of emotion i occurring in a sequence of emotions. In this case, the sequence includes the expression of the seven basic emotions. The binary logarithm ($\log_2$) identifies the times the number is multiplied by itself to obtain p_i, therefore, the sum of the $\log_2$ of the probabilities of the seven categories of emotions would show the degree of randomness (Karaca & Moonis, 2022). Specifically, the minimal value of entropy is zero, implying that the person expressed the same emotion throughout the task, with only one emotion having the highest probability of occurring. The maximum value of entropy was 2.8 ($\log_2(7)$), implying that the person experienced the seven emotions with equal probability, thus having high emotional variability (Zheng et al., 2023).

Notably, zero does not equal neutral, but rather shows persistence of only one of the seven emotions. Neutral is treated as a dominant emotion, and its frequency was transformed to calculate its probability of occurring. Thus, having neutral in the dominant emotions log would imply that this expression was sustained for more than 0.5 s and that its presence was higher than the other categories (Loijens et al., 2015).

3 | RESULTS

3.1 | Data screening

Within-subjects repeated-measures ANOVAs (RM-ANOVAs) were conducted in IBM® SPSS® version 29 to address each of our research questions. Python's package Matplotlib was used to create graphs (Hunter, 2007). Assumptions of independence and sphericity were met for all variables. Normality assumption was assessed with Shapiro–Wilk test (Meyers et al., 2013), showing a violation of normality for facial expression of disgust, fear, happy, sad and surprise across the three training phases and three difficulty levels. Therefore, we conducted non-parametric tests for identifying changes in emotional intensity inferred from facial expression. First, we conducted a Friedman test to identify if there were differences, followed by a Wilcoxon Signed-Rank test for pairwise comparisons, using the Bonferroni method to control for Type I error, in which the expected alpha value to identify significance (i.e., $p = 0.05$) was divided by the number of repeated measures (i.e., seven emotions for a cut-off value of $p = 0.007$).

For the rest of the variables, RM-ANOVAs were conducted to detect significant differences, after pairwise comparisons were conducted to evaluate the differences among the individual scores of the variables, using the Bonferroni method to control for Type I error, in which the expected alpha value to identify significance (i.e., $p = 0.05$) was divided by the number of repeated measures (three training phases or three difficulty levels) for a cut-off value of $p = 0.017$.

Two clarifications are provided here regarding the data used for analysis. First, during the experimental process, the facial expression of the participant with FAA instruction rating could not be analysed because the participant wore a face mask throughout the experiment, thus analyses regarding frequency of emotions and emotional variability were performed only with 22 participants and flying performance and SCR analyses were conducted with 23 participants. Second,

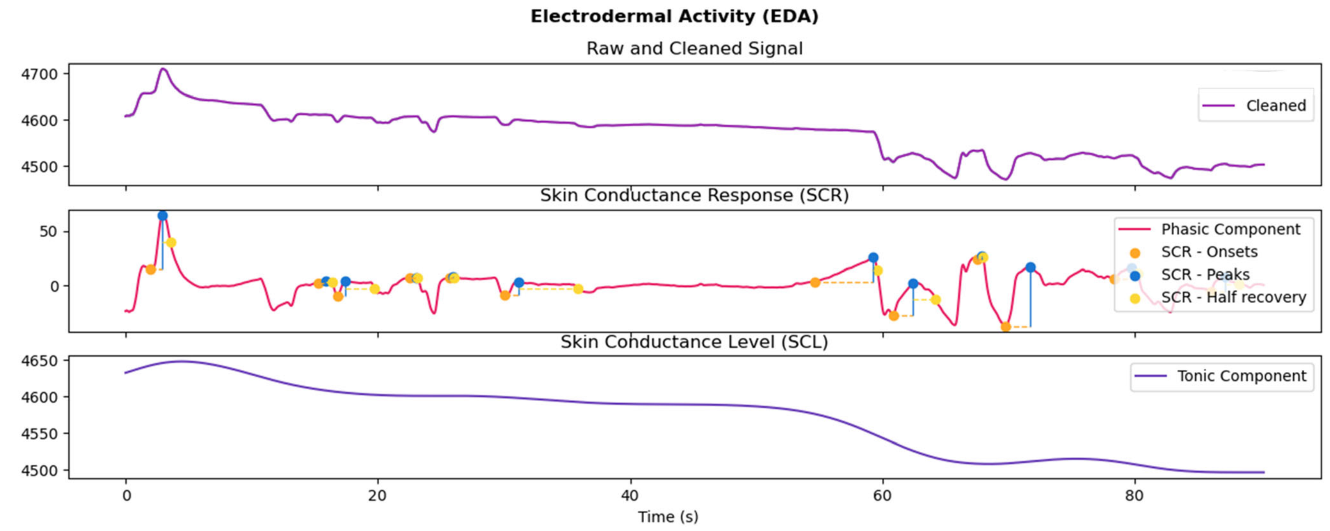

FIGURE 8 Sample of clean electrodermal activity (EDA) signal. The trial of this example was task two during introduction, with low difficulty. The instruction was as follows: 'Maintain altitude at 10,000 feet. At the same time: turn right at 30° AOB to a heading of 120° and roll out in a steady heading'.

TABLE 3 Flying performance across training phases.

Performance measure	Practice phase	Mean	SD	MD	F	p	η^2
Heading RMSE	Introduction	10.12	4.64		0.079	0.930	0.003
	Session A	10.21	3.33				
	Session B	10.48	3.81				
Altitude RMSE	Introduction	329.98	161.97	I > SA*	9.566	0.002	0.303
	Session A	262.57	113.01				
	Session B	233.42	83.20	I > SB**			
Heading expert	Introduction	3.67	0.33		1.077	0.350	0.047
	Session A	3.73	0.36				
	Session B	3.76	0.33				
Altitude expert	Introduction	3.09	0.60	I < SA**	15.79	<0.001	0.418
	Session A	3.31	0.48	SA < SB**			
	Session B	3.49	0.46	I < SB**			

Abbreviations: I, introduction; MD, mean difference; RMSE, root mean square error; SA, session A; SB, session B; SD, standard deviation.
*$p = 0.017$ after Bonferroni correction; **$p < 0.001$.

analyses exploring differences across difficulty levels only included trials in sessions A and B, such that there was an equal number of trials (i.e., five) per difficulty level.

3.2 | Differences across training phases

3.2.1 | Flying performance

Training phase had a significant effect on altitude performance measures, and it did not have a significant effect on heading metrics, see Table 3. Altitude RMSE was significantly higher in the introduction phase compared with session B ($p < 0.001$), and it was marginally higher in introduction compared with session A ($p = 0.024$). Expert altitude scorings also increased across phases: the introduction phase had a lower score than session A ($p = 0.006$) and session B ($p < 0.001$); and session A had a lower score than session B ($p = 0.006$). Overall, altitude error was lower at the end of the experiment (session B), compared with the introduction phase. See trends in Figure 9.

3.2.2 | Emotional intensity—physiological arousal

Training phases had a significant effect on the frequency of SCR, descriptive statistics in Table 4. Pairwise comparisons showed that participants significantly experienced more SCRs during session A compared with introduction ($p < 0.001$) and session B ($p = 0.011$), see Figure 10.

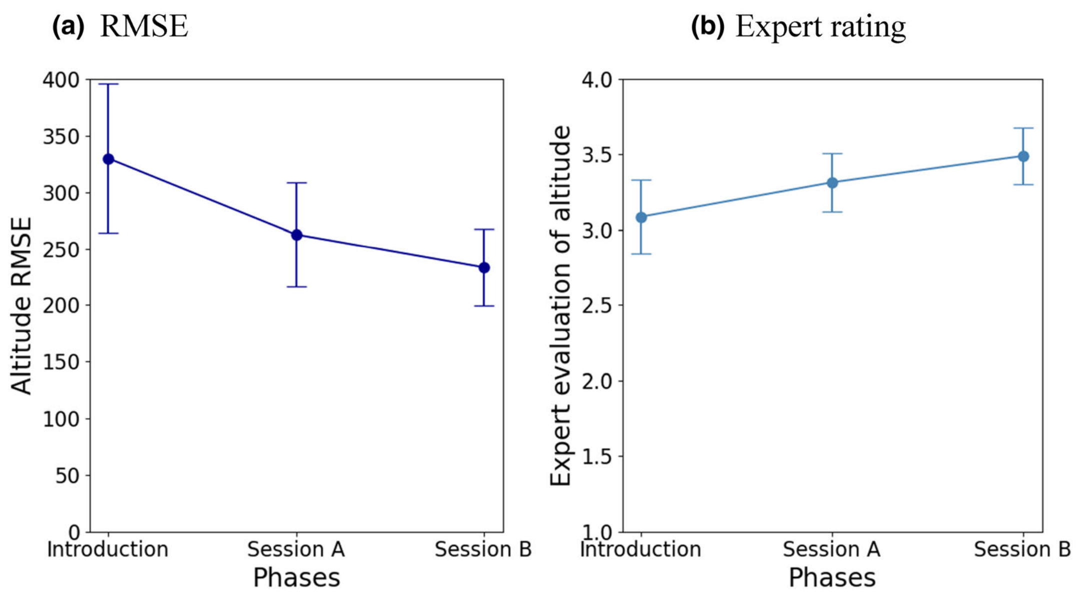

FIGURE 9 Altitude performance across training phases. Error bars represent confidence intervals. RMSE, root mean square error.

TABLE 4 SCR across training phases.

Performance measure	Training phases	Mean	SD	MD	*F*	*p*	η^2
SCR	Introduction	107.69	25.57	I < SA**	7.98	0.001	0.266
	Session A	127.61	30.69	SA > SB*			
	Session B	114.43	27.44				

Abbreviations: I, introduction; MD, mean difference; SA, session A; SB, session B; SCR, skin conductance response; SD, standard deviation.
*$p < 0.017$ after Bonferroni correction; **$p < 0.001$.

3.2.3 | Emotional intensity—facial expressions

Three nonparametric Friedman tests were conducted, one for each phase (introduction, session A and session B), comparing the frequency of the seven dominant emotions (i.e., anger, neutral, happy, disgust, fear, sad and surprise); descriptive statistics in Table 5. We used frequency of facial expressions, rather than proportions, since the duration of the tasks was equal for all participants (Lajoie et al., 2021). Frequency of facial expression of emotions had statistically significant differences across the three phases ($X^2_{\text{introduction}} = 66.80$, $X^2_{\text{Session A}} = 64.92$, $X^2_{\text{Session B}} = 72.51$; $p < 0.001$), see Figure 11. Neutral, anger and surprise were more frequent than fear, and disgust across the three training phases. Happiness was less frequent than neutral and anger across the three training phases, and less frequent than surprise in session B; however, it was more frequent than disgust in sessions A and B. Sadness was less frequent than neutral across the three phases, less frequent than anger during introduction and session A.

As a follow-up, nonparametric Friedman tests were conducted to understand if single emotions differed across training phases. Results did not show a significant effect of training phases within emotions.

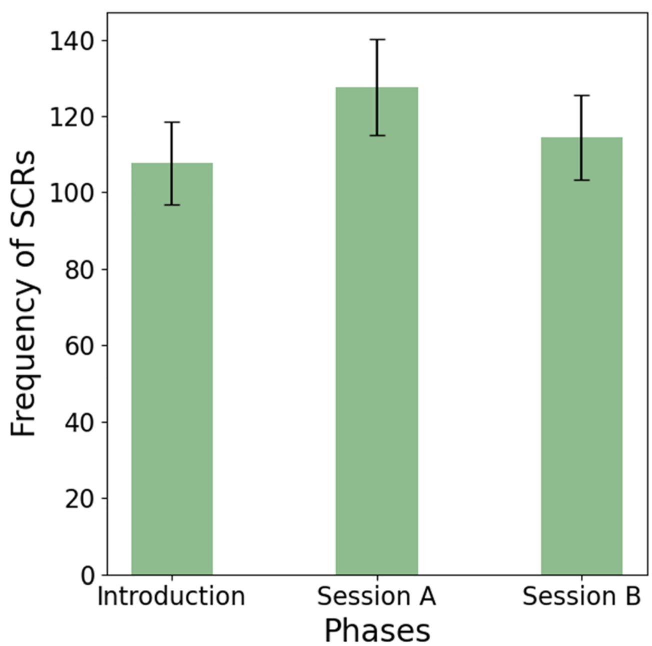

FIGURE 10 Skin conductance responses (SCRs) across training phases. Error bars represent confidence intervals.

TABLE 5 Descriptive statistics of frequency of facial expressions across training phases.

	Introduction		Session A		Session B	
	M	*SD*	*M*	*SD*	*M*	*SD*
Neutral	29	22.18	31.18	16.09	33.36	21.27
Anger	26.23	18.71	23.36	15.53	21.82	17.52
Surprise	14.04	14.73	15.64	20.08	16.23	17.78
Sad	6.14	9.37	8.59	15.34	10.91	16.90
Happy	3.54	4.75	5.64	7.74	2.5	3.26
Scared	1.23	1.92	2.68	5.48	2.82	6.06
Disgusted	1	1.90	0.77	1.54	0.45	0.67

3.2.4 | Emotional variability—facial expressions fluctuations

Training phases did not have a significant effect on emotional variability. Descriptively participants experienced more emotional variability during session A, compared with introduction and session B (see Table 6 and Figure 12).

3.3 | Differences across difficulty levels

3.3.1 | Flying performance

Difficulty level had a significant effect on performance during session A and B; descriptive statistics in Table 7. Pairwise comparisons showed performance had statistically significant differences across difficulty levels.

Heading RMSE was significantly higher for high-difficulty manoeuvres, compared with low- ($p < 0.001$) and medium-difficulty manoeuvres ($p < 0.001$). Participants had a lower heading score in high-difficult manoeuvres compared with medium difficulty, with a marginally significant difference ($p = 0.02$). See trends in Figure 13.

Altitude RMSE was higher for high difficulty manoeuvres compared with low ($p < 0.001$) and medium-difficulty levels ($p < 0.001$). Participants had lower altitude scores for high difficulty manoeuvres compared with low- and medium-difficulty manoeuvres ($p < 0.001$). Considering that higher RMSE reflects more error, and a higher expert rating reflect more accuracy, the results show that performance was generally worse in high-difficulty tasks, compared with low- and medium-difficulty levels, see trends in Figure 14.

3.3.2 | Emotional intensity—physiological arousal

Difficulty levels had a significant effect on the frequency of SCR (Table 8). Participants experienced significantly more SCRs in high-difficulty compared with low-difficulty manoeuvres ($p = 0.002$). See trends in Figure 15.

3.3.3 | Emotional intensity—facial expressions

Three Friedman nonparametric tests were conducted, one for each difficulty level (low-, medium- and high-difficulty) comparing the frequency of the seven dominant emotions (i.e., anger, neutral, happy, disgust, fear, sad and surprise); descriptive statistics in Table 9. Frequency of facial expression of emotions had a statistically significant difference across the three difficulty levels ($X^2_{low} = 61.15$; $X^2_{Medium\ A} = 64.92$, $X^2_{High} = 65.94$; $p < 0.001$), see Figure 16. Participants showed more neutral, anger and surprise facial expressions compared with happiness, fear and disgust, and happiness was more frequent than disgust for all three difficulty levels. Moreover, expression of neutral was more frequent than sadness across the three difficulty levels.

As a follow up, nonparametric Friedman's tests were conducted comparing single discrete emotions across difficulty levels. Results did not show a significant effect of difficulty level within emotions.

3.3.4 | Emotional variability—fluctuation of facial expressions

Difficulty levels had a significant effect on emotional variability, descriptive statistics in Table 10. Pairwise comparisons revealed a marginally significant difference in emotional variability when comparing medium and high-difficulty tasks, showing a higher emotional variability in high-difficulty tasks compared with medium-difficulty tasks ($p = 0.041$). See trends in Figure 17.

4 | DISCUSSION

4.1 | Differences across training phases

Our results showed that performance changed across training phases. Aligned with our hypothesis, altitude control improved over training phases such that by the end of the experiment, participants were more accurate, requiring less amplitude corrections and were granted a better score by the expert. However, heading error and expert rating did not have a statistically significant difference across phases. A potential explanation is that the yoke movements to control heading are similar to performing car turns using a steering wheel. Thus, participants were likely more familiar with the movement to control heading, causing less error and less differences across training phases. In contrast, since participants were mostly brand-new beginners, they might had focused on altitude, as an unknown metric and had less intuitive control. This is a known pattern among novices who tend to concentrate on one aspect of the task when it is not familiar, compared with experts who have the ability to visualize problems from a global perspective and have perfected technical methods and cross-check techniques (Ericsson, 2006; Lajoie & Gube, 2018).

When analysing changes on high intensity (single-variable emotions), we found significant changes in SCRs. Our expectations

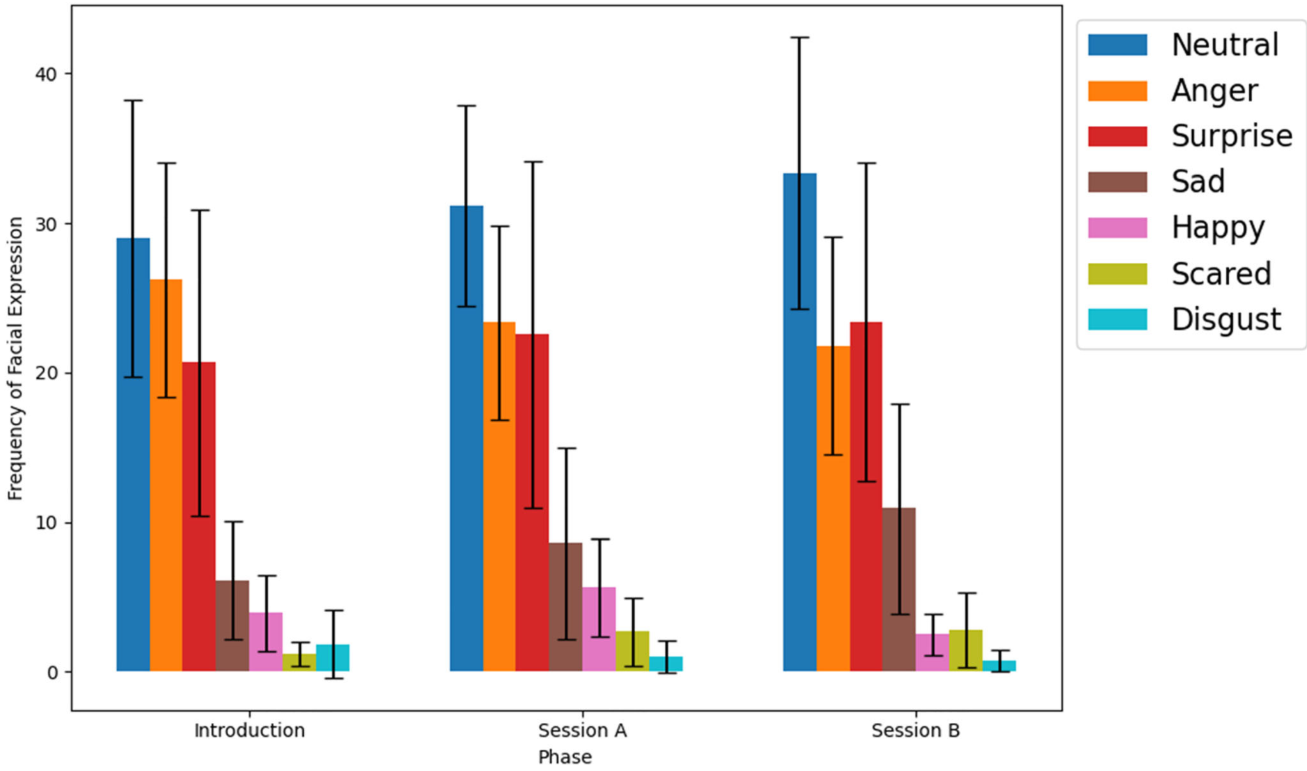

FIGURE 11 Frequency of intense facial expressions across training phases. Pairwise comparisons: NE > SA, NE > HA, NE > SC, NE > DI, AN>HA, AN>SC, AN>DI in introduction**, sessions A** and B**. SU > HA in session B**. SU > SC in sessions A* and B*. HA > DI in sessions A* and B*. AN>SA in introduction** and session A*. *p < 0.007 after Bonferroni correction, **p < 0.001. Error bars represent confidence intervals. AN, angry; DI, disgust; NE, neutral; SA, sad; SC, scared; SU, surprise.

TABLE 6 Emotional variability across training phases.

Performance measure	Training phases	Mean	SD	MD	F	p	η^2
Emotional variability	Introduction	1.58	0.55		0.312	0.734	0.015
	Session A	1.65	0.36				
	Session B	1.58	0.37				

Abbreviations: I, introduction; MD, mean difference; SA, session A; SB, session B; SD, standard deviation.

regarding changes in physiological arousal were partially fulfilled. Aligned with our hypothesis, participants experienced more SCRs in session A compared with session B. Based on this finding, we infer that participants were most engaged during session A, which aligns with experiencing more boredom (and less engagement) at the end of the task (Goetz & Hall, 2014; Hidalgo-Munoz et al., 2018; Koglbauer et al., 2011). However, in our hypothesis, we did not clarify expectations regarding differences between introduction and session A, finding that participants experienced more SCRs in session A. A potential explanation of this pattern can be that feedback was removed as the practice phases begun. In our experimental design, researchers provided only constructive feedback during the training phase (Krahenbuhl et al., 1981). The positive feedback served as a learning scaffold for participants to fill the gap between current and potential knowledge (Lajoie, 2017). Therefore, participants were likely to be less aroused when receiving feedback (during introduction) compared with when they performed the task independently for the first time (session A). As participants habituated to the task and its expectations (during session B), there was a decrease in physiological arousal and emotional variability (Hidalgo-Munoz et al., 2018; Krahenbuhl et al., 1981).

When analysing differences in intense emotions, inferred from facial expressions, we found that emotions had a binary separation between high versus low frequency. Namely, neutral, anger and surprise were significantly more frequent across the three training phases compared with happiness, fear and disgust, in that order. As expected, neutral served as an emotional baseline, thus, it is likely that trainees were not emotionally distracted during the task and were able to engage their cognitive efforts to perform the flying manoeuvres (Harley et al., 2012), which at the same time can explain the improvement in altitude performance by the end of the task. Moreover, surprise is recognized as a frequent emotion in pilots as they constantly react to unexpected changes when flying (Landman et al., 2020). However, we believe that surprise should be researched in more detail as its immediate effect in performance might be negative, by distracting participants, yet if the impasse is solved successfully, the increase in physiological arousal might stimulate learners to solve the unexpected change (D'Mello & Graesser, 2012; Landman et al., 2020).

However, explaining the frequency of anger was less straightforward. Anger, paired with frustration (Harley et al., 2015; Pekrun, 2019), is a negative-deactivating emotion that occurs when the learner perceives the task demands as unreasonable or controlled by external factors (Pekrun & Perry, 2014). In a simple explanation, participants might had perceived the task as unnecessarily complicated, however, they were determined (aroused) to finish the task. However, the current analysis lacks information as of the locus of control of participants. Therefore, we propose a second explanation. Facial expression of anger, specially lowering browns and tightening the eye lids (Farnsworth, 2022), has been previously interpreted as deep-focus in a flying task, especially when looking at a screen with a 'frowning face' (Li & Lajoie, 2021). Due to the task characteristics, since participants are facing a screen and it is a low-stakes task using a simulator, we argue that the second interpretation might be more pertinent.

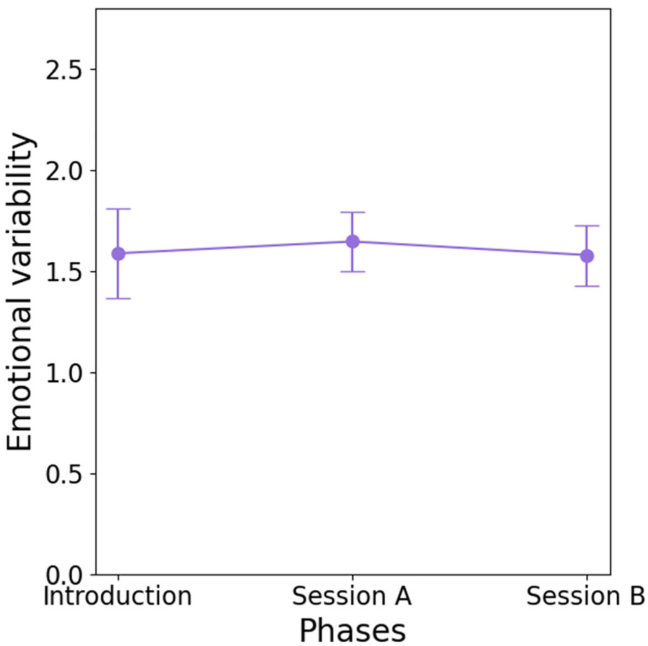

FIGURE 12 Emotional variability across training phases. Error bars represent confidence intervals.

Moreover, results show that happiness was more frequent than disgust in training sessions A and B, as well as across all difficulty levels. Previous studies have explored undergraduate students' emotions when interacting with intelligent tutoring systems, similarly finding that neutral is one of the most frequent emotions that learners experience (D'Mello & Graesser, 2012; Harley et al., 2013). Contrary to our results, these studies found that confusion (paired with disgust) and happiness were identified as the most frequent emotions (accordingly D'Mello & Graesser, 2012; Harley et al., 2013). Yet, in our case, these two emotions came as part of the least frequent emotions, but still happiness was more present than disgust. Since this was the pattern in both practice sessions, it might be that participants became more familiar with the task, thus experiencing less confusion and more joy. However, as both emotions are not too frequent, more studies are needed to explore if the frequency of emotions is aligned to the profession (i.e., pilots) and/or the task (i.e., simulated flying).

Results did not show a statistically significant difference in emotional variability across training phases, but a trend showed that emotional variability was higher during session A than during introduction and session B training phases.

The interpretations of our results are linked to our measurements. FaceReader 6.0 is built as an objective measure detecting facial action units that are then merged and results are labelled according to basic emotions theory and the circumplex model of affect (Ekman, 1992,

TABLE 7 Flying performance across difficulty levels.

Performance measure	Difficulty level	Mean	SD	MD	F	p	η^2
Heading RMSE	Low	8.40	3.57	L < H**	31.11	<0.001	0.586
	Medium	7.77	3.02	M < H**			
	High	14.92	5.35				
Altitude RMSE	Low	186.55	81.65	L < H**	24.28	<0.001	0.525
	Medium	183.07	87.13	M < H**			
	High	377.54	188.45				
Heading expert	Low	3.84	0.30		3.88	0.028	0.150
	Medium	3.83	0.31	M > H			
	High	3.55	0.65				
Altitude expert	Low	3.68	0.41	L > H**	32.71	<0.001	0.598
	Medium	3.56	0.47	M > H**			
	High	2.95	0.65				

Abbreviations: H, high; L, low; M, medium; MD, mean difference; SD, standard deviation.
**p < 0.001.

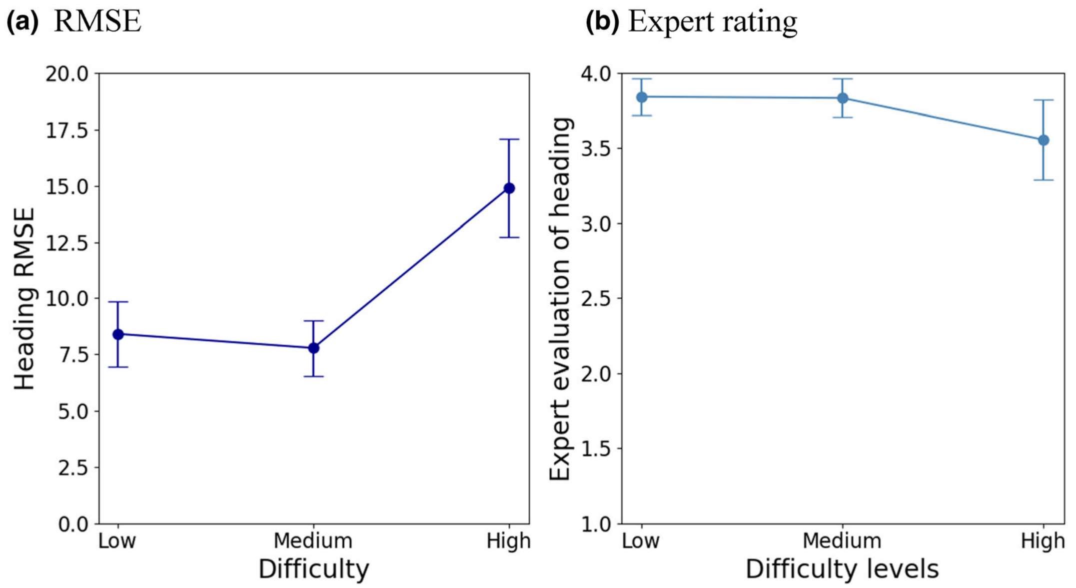

FIGURE 13 Heading performance across difficulty levels. Error bars represent confidence intervals. RMSE, root mean square error.

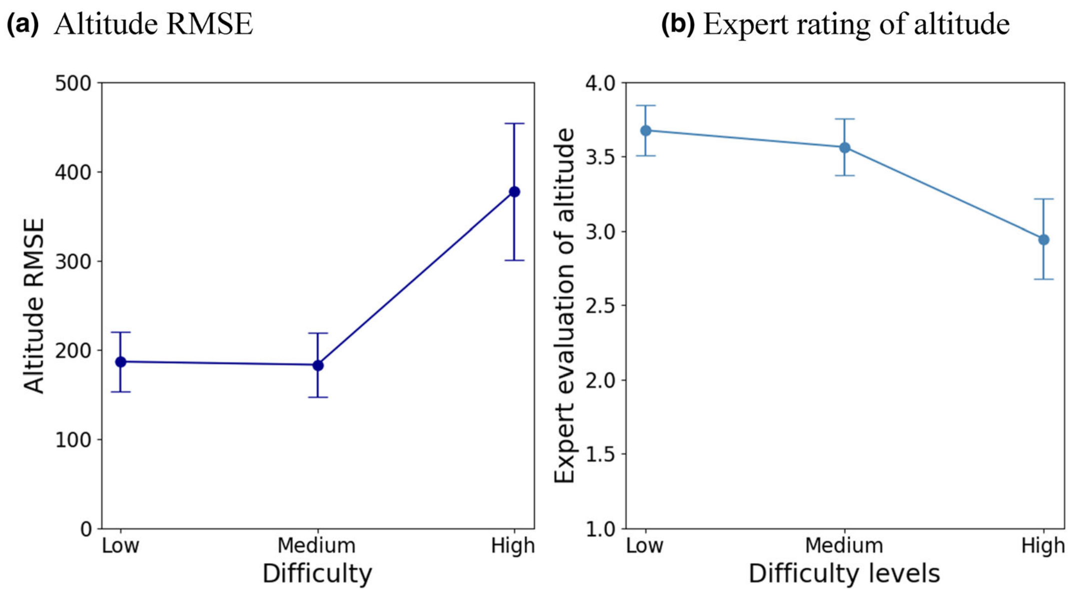

FIGURE 14 Altitude performance across difficulty levels. Error bars represent confidence intervals. RMSE, root mean square error.

den Uyl et al., 2023; Russell, 2003). FaceReader's manufacturer, Noldus, has recently changed its approach, encouraging users to interpret facial expressions according to their needs and theoretical standing (den Uyl et al., 2023). Therefore, a limitation of using FaceReader version 6.0 is that it does not have personalization features to label groups of action units, as in our case anger or deep focus, leading us to interpret a frowning face according to our context (den Uyl et al., 2023; Loijens et al., 2015). We recognize this limitation; however, our results invite future research to use more flexible facial-detection software, such as FaceReader 9.0 or OpenFace, to label emotions according to each study theoretical framework.

We recognize a few limitations of this study, and we provide some recommendations for future directions. For instance, most of the participants were not pilot students, but rather represent ab initio pilot trainees (i.e., brand-new trainees) with limited experience using flight simulators or aircraft. Moreover, participants' educational background was heterogeneous, potentially influencing the results. We suggest future studies to replicate or extend the test methodology and include student pilots recently enrolled in aviation school and more advanced pilots. Moreover, it is suggested that the analysis is reproduced with a larger sample size to increase statistical power. We recommend that future studies record and control for instructors'

TABLE 8 SCR across difficulty levels.

Emotion	Difficulty level	Mean	SD	MD	*F*	*p*	η^2
SCR	Low	74.61	18.56	L < H*	6.71	0.003	0.234
	Medium	80.48	20.41				
	High	86.96	21.40				

Abbreviations: H, high; L, low; M, medium; MD, mean difference; SCR, skin conductance response; SD, standard deviation.
**p* = 0 .017 after Bonferroni correction.

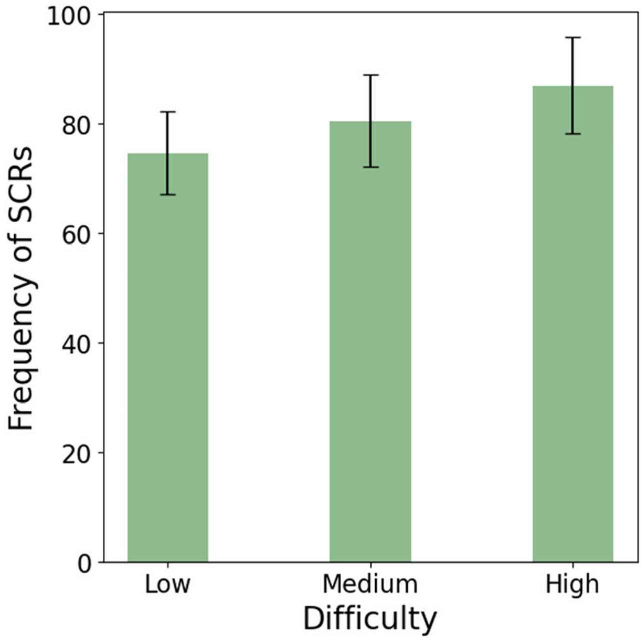

FIGURE 15 Skin conductance responses (SCRs) across difficulty levels. Error bars represent confidence intervals.

TABLE 9 Descriptive statistics frequency of facial expressions across difficulty levels.

	Low		Medium		Difficult	
	M	SD	*M*	SD	*M*	SD
Neutral	19.36	12.31	22.73	14.77	22.45	10.50
Anger	13.18	9.62	15.23	13.16	16.77	11.24
Surprise	14.04	14.73	15.64	20.08	16.23	17.78
Sad	5.81	9.14	6.95	11.45	6.73	11.98
Happy	2.14	3.67	1.77	2.56	4.23	6.13
Scared	2.18	4.10	1.45	3.33	1.86	4.05
Disgusted	0.41	0.73	0.41	1.30	0.95	2.63

feedback to understand how advice aligns with students' expectations and emotional responses (Naismith & Lajoie, 2018; Shute, 2008).

Regarding emotional intensity, we recognize a few limitations. The non-normality of the facial expression of the non-frequent emotions (i.e., sad, fear, disgust, happiness and surprise) may impact the generalizability of our findings, and emotional variability did not show statistically significant differences and a low effect size (Cohen, 1988).

In summary, the introduction phase was characterized as having poorer altitude performance and lower frequency of emotions overall, with participants experiencing less sadness and less SCRs. During session A, performance improved being significantly better than introduction but not different from session B; additionally, participants experienced more SCRs, and we observed (descriptively) more emotional variability. Last, during session B performance was more accurate and participants had a higher frequency of neutral and sad expressions across the three phases.

4.2 | Differences across difficulty levels

Our results confirmed that flying performance and emotions varied across difficulty levels. The trend in performance was clear, showing that flying performance was less precise in high-difficulty tasks compared with low and medium-difficulty levels.

Aligned with our expectations, participants experienced more emotional intensity, inferred from more SCRs, during high-difficulty manoeuvres than low-difficulty manoeuvres. These findings align with previous research showing that pilot trainees' peaks of physiological arousal increase with task difficulty (Gaetan et al., 2015; Skibniewski et al., 2015). Moreover, more SCRs indicates that participants were more emotionally and cognitively engaged when attempting to solve the task (Braithwaite et al., 2015; Harley, Jarrell et al., 2019).

When analysing the differences of emotional intensity across difficulty levels, neutral anger and surprise remain as the most frequently expressed emotions, being significantly more frequent than happiness, fear and disgust. Notably, only neutral is more frequent than sadness confirming that neutral serves as a baseline state and shows that participants are generally stable during the flying task (Harley et al., 2012).

Emotional variability was higher during high-difficulty tasks compared with medium-difficulty levels, similar to results found with medical students (Li, Zheng, Lajoie, 2021). Difficult tasks required more changes between aircraft metrics, and these constant changes might imply constant fluctuations in object focus, triggering different emotions and causing less stability (Pekrun, 2019). However, results show that emotional variability was (descriptively) higher during low than medium-difficulty tasks. Our results show a U-shaped pattern between emotional variability and difficulty level, in which emotional variability was higher in low and difficult tasks, compared with

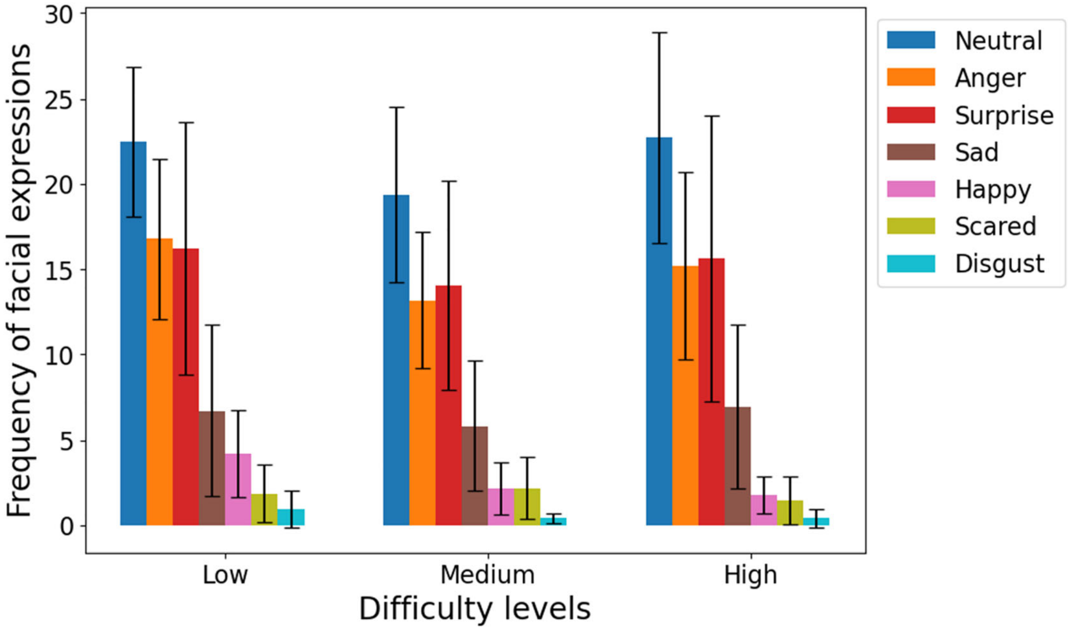

FIGURE 16 Frequency of facial expressions across difficulty levels. Pairwise comparisons: NE > SA, NE > HA, NE > SC, NE > DI, AN>HA, AN>SC, AN>DI, SU > DI in low**, medium** and high** difficulty. SU > HA in low*, medium* and high* difficulty. HA > DI in low*, medium* and high* difficulty. NE > AN in low* and medium* difficulty. AN > SA in low* and high* difficulty. SA > DI in low* and medium* difficulty. SU > SA in high difficulty*. *$p < 0.01$, **$p < 0.001$. Error bars represent confidence intervals. AN, angry; DI, disgust; NE, neutral; SA, sad; SC, scared; SU, surprise.

TABLE 10 Emotional variability across difficulty levels.

Performance measure	Difficulty level	Mean	SD	MD	F	p	η^2
Emotional variability	Low	1.68	0.31		3.37	0.044	0.138
	Medium	1.53	0.51	M < H*			
	High	1.68	0.36				

Abbreviations: H, high; L, low; M, medium; MD, mean difference; SD, standard deviation.
*$p = 0.017$ after Bonferroni correction (Meyers et al., 2013).

medium difficulty; yet performance was increasingly worse as difficulty level increased. Emotional variability might be following the Yerkes–Dodson law (1908), in which an optimal level of emotional fluctuations can improve performance, but trespassing that threshold (as in the high-difficulty tasks) might imply an impairment in flying accuracy. This threshold might be inferred by physiological arousal, which was higher in high than low-difficulty tasks.

It is recognized that our analyses had limitations that could guide future research. For instance, when analysing changes of single facial expressions across difficulty levels there were no statistically significant differences for any of the seven emotions. In that sense, we recognize a common critique of using behavioural measures to evaluate emotion: behavioural responses should be triangulated with people's subjective experience, such as questionnaires or interviews (Harley, 2016). Nevertheless, this was a calculated risk to understand the feasibility of using a non-invasive and non-distracting measure of emotions, which is particularly relevant in the context of aviation training where pilots require full concentration on the task, rather than responding to questionnaires concurrently during the flying tasks.

We encourage future directions according to our results. Future studies could explore the correlation of neutral expressions as a positive correlation with flying performance. Moreover, we suggest exploring emotional variability of pilot trainees in a larger sample and adding (non-interrupting) self-report measures to understand trainees' subjective experience to identify the interaction between emotional variability and flying performance. In the current study, we explored behavioural and physiological expression of emotions during the same time period; however, we suggest that future research explores the co-occurrence of behavioural and physiological expression of emotions, such as checking time points in which the significant facial expressions co-occur with SCRs in key moments of the task (i.e., sudden decrease of altitude) (Ruiz Segura, 2020). Such combinations could be automatically detected to create interventions for trainees to up or down-regulate their emotions, resulting in optimal performance (Gross, 2015).

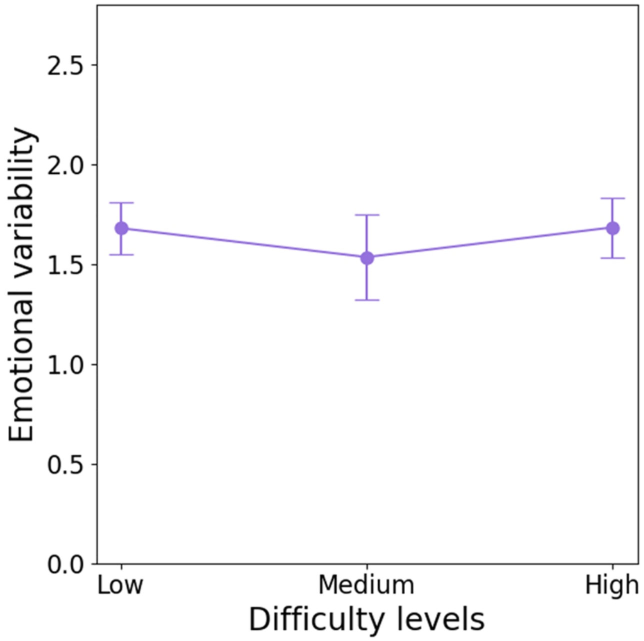

FIGURE 17 Emotional entropy across difficulty levels. Error bars represent confidence intervals.

In summary, in low- and medium-difficulty levels, performance was significantly better than in difficult tasks, and participants had more neutral, angry and surprise than happy, fear and disgusted expressions. In low-difficulty tasks, participants had less SCRs than in difficult tasks. Last, high-difficulty tasks had worse performance than low and medium-difficulty levels, participants expressed overall more emotions, had more SCRs and had more emotional variability.

5 | CONCLUSION

In this study, we argue that a mean to improve pilot trainees experience and performance is to account for their emotional experiences when flying (Martins, 2016). This study shows that pilot trainees experience dynamic changes among multiple emotions. Our results show that emotion dynamics, such as intensity, frequency and variability, and flying performance change according to training phases and difficulty levels (Bailen et al., 2019; Zheng et al., 2023).

This manuscript serves as a baseline to understand emotional experiences of beginner pilot trainees, with the subsequent objective of informing the development of interventions to improve flight performance. To reduce human error, pilot trainees can learn emotion regulation techniques as part of their curriculum. Particularly, future training can use simulated flying tasks to assist trainees to familiarize themselves with the emotional reactions they have as training advances or encountering difficult tasks. More awareness about their emotions might help pilot trainees recognize key moments of emotional fluctuations or high physiological arousal, identify how those emotional changes relate to their performance for allowing trainees to self-regulate and modify their behaviour or strategy, in addition to permitting instructors to intervene in a timely fashion and reduce likelihood of errors when flying an airplane (Dismukes, 2010).

AUTHOR CONTRIBUTIONS

Alejandra Ruiz-Segura: Conceptualization; investigation; writing – original draft; methodology; writing – review and editing; visualization; formal analysis; software; data curation. **Andrew Law:** Project administration; methodology; validation; visualization; writing – review and editing; software; formal analysis. **Sion Jennings:** Project administration; supervision; methodology; writing – review and editing. **Alain Bourgon:** Methodology; validation; formal analysis; writing – review and editing. **Ethan Churchill:** Validation; methodology. **Susanne Lajoie:** Supervision; resources; funding acquisition; writing – review and editing; data curation.

ACKNOWLEDGEMENTS

This work was supported in part by the Fonds de recherche du Québec – Société et Culture Doctoral training scholarships awarded to Alejandra Ruiz-Segura (286648), and the Social Sciences and Humanities Research Council (SSHRC) Grant of Canada 895–2011-1006 awarded to Professor Susanne Lajoie and the Learning Across the Disciplines and Canada Research Chair funding on Advanced Technologies in Authentic Settings 950-231835 and largely supported by a jointly funded Natural Sciences and Engineering Research Council of Canada (514052-17) and Consortium for Aerospace Research and Innovation in Canada (CARIC) and Quebec (CRIAQ) (OPR-1618) grants, which had multiple collaborators in the InLook Project (CAE; Marinvent Corporation; the Design Lab, Concordia University; the Faubert Lab, University of Montreal; Flight Research Laboratory, National Research Council). We thank all our collaborators from the InLook project for their roles in experimental set-up and data collection. Special thanks to Hugh Grenier and Alain Bourgon (CAE) as the subject-matter experts for designing the flying tasks, training researchers and creating the video training for participants, MSc. David Bowes, Alexander Ireland and Jack Russ (CAE) for managing this project, Joao Araujo and Dr. Sujoy Ghosh Hajra for the system integration and experimental data collection and Master of Applied Science (M.A.Sc.) Daniel Kyrollos for serving as NRC internal reviewer.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/10.1111/jcal.13037

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

All procedures performed in human participants were in accordance with the ethical standards from Concordia University Research Ethics Board and was accepted by McGill University based on the CREPUQ interinstitutional agreement.

INFORMED CONSENT

Informed consent was obtained from all individual participants included in the study.

ORCID

Alejandra Ruiz-Segura https://orcid.org/0000-0003-4061-4931
Andrew Law https://orcid.org/0000-0002-7604-0506
Sion Jennings https://orcid.org/0009-0008-5490-9996
Susanne Lajoie https://orcid.org/0000-0003-2814-3962

REFERENCES

Allsop, J., & Gray, R. (2014). Flying under pressure: Effects of anxiety on attention and gaze behavior in aviation. *Journal of Applied Research in Memory and Cognition*, *3*(2), 63–71. https://doi.org/10.1016/j.jarmac.2014.04.010

Artino, A. R., Holmboe, E. S., & Durning, S. J. (2012). Control-value theory: Using achievement emotions to improve understanding of motivation, learning, and performance in medical education: AMEE guide no. 64. *Medical Teacher*, *34*(3), e148–e160. https://doi.org/10.3109/0142159X.2012.651515

Azher, S., Cervantes, A., Marchionni, C., Grewal, K., Marchand, H., & Harley, J. M. (2023). Virtual simulation in nursing education: Headset virtual reality and screen-based virtual simulation offer a comparable experience. *Clinical Simulation in Nursing*, *79*, 61–74. https://doi.org/10.1016/j.ecns.2023.02.009

Bailen, N. H., Green, L. M., & Thompson, R. J. (2019). Understanding emotion in adolescents: A review of emotional frequency, intensity, instability, and clarity. *Emotion Review*, *11*(1), 63–73. https://doi.org/10.1177/1754073918768878

Baumann, M. R., Gohm, C. L., & Bonner, B. L. (2011). Phased training for high-reliability occupations: Live-fire exercises for civilian firefighters. *Human Factors*, *53*(5), 548–557. https://doi.org/10.1177/0018720811418224

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Walton, T. R., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). A guide for analysing electrodermal activity (EDA) and skin conductance responses (SCRs) for psychological experiments. SAAL, Behavioural Brain Sciences Centre, University of Birminghan. https://www.birmingham.ac.uk/Documents/college-les/psych/saal/guide-electrodermal-activity.pdf

Code of Federal Regulations. (2023, April 19). 14 CFR 91.1059—Flight time limitations and rest requirements: One or two pilot crews. Electronic code of federal regulations. https://www.ecfr.gov/current/title-14/chapter-I/subchapter-F/part-91/subpart-K/subject-group-ECFRc17623c0e0be17e/section-91.1059

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cootes, T. F., & Taylor, C. J. (2001). Statistical models of appearance for medical image analysis and computer vision. *Medical Imaging 2001: Image Processing*, *4322*, 236–248. https://doi.org/10.1117/12.431093

Damasio, A. R. (2005). *Descartes error: Emotion, reason, and the human brain*. Penguin Books.

den Uyl., T., Lewinski, P., & Gudi, A. (2023). *FaceReader and different scientific theories on emotion. Behavioral Research Blog by Noldus Information Technology*. Retrieved 28 February 2024: https://www.noldus.com/blog/scientific-theories-on-emotion

Dismukes, R. K. (2010). Understanding and analyzing human error in real-world operations. In E. Salas & D. Maurino (Eds.), *Human factors in aviation* (2nd ed., pp. 335–374). Academic Press. https://doi.org/10.1016/B978-0-12-374518-7.00011-0

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Duffy, M. C., Lajoie, S., & Lachapelle, K. (2016). Measuring emotions in medical education: Methodological and technological advances within authentic medical learning environments. In S. Bridges, L. K. Chan, & C. E. Hmelo-Silver (Eds.), *Educational technologies in medical and health sciences education* (pp. 181–213). Springer. https://doi.org/10.1007/978-3-319-08275-2_10

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 3–28). Lawrence Erlbaum Associates.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–704). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.038

European Commission. (2013, October 4). *Pilot and crew fatigue—Frequently asked questions*. Directorate-General for Communication, European Commission. https://ec.europa.eu/commission/presscorner/detail/en/MEMO_13_854

FAA Department of Transportation. (2003, September 17). 14 CFR 91.1059—Flight time limitations and rest requirements: One or two pilot crews. Electronic Code of Federal Regulations. https://www.ecfr.gov/current/title-14/chapter-I/subchapter-F/part-91/subpart-K/subject-group-ECFRc17623c0e0be17e/section-91.1059

Farnsworth, B. (2022, October 18). Facial action coding system (FACS)—A visual guidebook—iMotions. https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Gaba, D. M. (2004). The future vision of simulation in health care. *Quality & Safety in Health Care*, *13*(Suppl 1), i2–i10. https://doi.org/10.1136/qhc.13.suppl_1.i2

Gaetan, S., Dousset, E., Marqueste, T., Bringoux, L., Bourdin, C., Vercher, J.-L., & Besson, P. (2015). Cognitive workload and psychophysiological parameters during multitask activity in helicopter pilots. *Aerospace Medicine and Human Performance*, *86*(12), 1052–1057. https://doi.org/10.3357/AMHP.4228.2015

Goetz, T., & Hall, N. C. (2014). Academic boredom. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 311–330). Routledge.

Gray, R., Gaska, J., & Winterbottom, M. (2016). Relationship between sustained, orientated, divided, and selective attention and simulated aviation performance: Training & pressure effects. *Journal of Applied Research in Memory and Cognition*, *5*(1), 34–42. https://doi.org/10.1016/j.jarmac.2015.11.005

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Hamman, W. R. (2004). Will simulation fly in medicine as it has in aviation? *Quality & Safety in Health Care*, *13*(5), 397–399. https://doi.org/10.1136/qhc.13.5.397

Han, S.-Y., Kwak, N.-S., Oh, T., & Lee, S.-W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering*, *40*(1), 324–336. https://doi.org/10.1016/j.bbe.2019.12.002

Hardy, M., Wiebe, E. N., Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2013). Physiological responses to events during training: Use of skin conductance to inform future adaptive learning systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 2101–2105. https://doi.org/10.1177/1541931213571468

Harley, J. M. (2016). Measuring emotions: A survey of cutting edge methodologies used in computer-based learning environment research. In S. Y. Tettegah & M. Gartmeier (Eds.), *Emotions, technology, design, and learning* (pp. 89–114). Academic Press. https://doi.org/10.1016/B978-0-12-801856-9.00005-0

Harley, J. M., Bilgic, E., Lau, C. H. H., Gorgy, A., Marchand, H., Lajoie, S. P., Lavoie-Tremblay, M., & Fried, G. M. (2023). Nursing students reported more positive emotions about training during coronavirus disease 2019 (COVID-19) after using a virtual simulation paired with an in-person simulation. *Clinical Simulation in Nursing*, *81*, 101420. https://doi.org/10.1016/j.ecns.2023.04.006

Harley, J. M., Bouchet, F., & Azevedo, R. (2012). Measuring learners' co-occurring emotional responses during their interaction with a pedagogical agent in MetaTutor. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (pp. 40–45). Springer. https://doi.org/10.1007/978-3-642-30950-2_5

Harley, J. M., Bouchet, F., & Azevedo, R. (2013). Aligning and comparing data on emotions experienced during learning with MetaTutor. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 61–70). Springer. https://doi.org/10.1007/978-3-642-39112-5_7

Harley, J. M., Bouchet, F., Hussain, M. S., Azevedo, R., & Calvo, R. (2015). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, *48*, 615–625. https://doi.org/10.1016/j.chb.2015.02.013

Harley, J. M., Jarrell, A., & Lajoie, S. P. (2019). Emotion regulation tendencies, achievement emotions, and physiological arousal in a medical diagnostic reasoning simulation. *Instructional Science*, *47*(2), 151–180. https://doi.org/10.1007/s11251-018-09480-z

Harley, J. M., Pekrun, R., Taxer, J. L., & Gross, J. J. (2019). Emotion regulation in achievement situations: An integrated model. *Educational Psychologist*, *54*(2), 106–126. https://doi.org/10.1080/00461520.2019.1587297

Hart, S. G., & Bortolussi, M. R. (1984). Pilot errors as a source of workload. *Special Issue: Aviation Psychology*, *26*(5), 545–556.

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX): 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Herrera-Aliaga, E., & Estrada, L. D. (2022). Trends and innovations of simulation for twenty first century medical education. *Frontiers in Public Health*, *10*, 619769. https://doi.org/10.3389/fpubh.2022.619769

Hidalgo-Munoz, A. R., Mouratille, D., Matton, N., Causse, M., Rouillard, Y., & El-Yagoubi, R. (2018). Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation. *International Journal of Psychophysiology*, *128*, 62–69. https://doi.org/10.1016/j.ijpsycho.2018.04.002

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. Psychological Bulletin, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology: CB*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Jennings, S., Law, A., Bourgon, A., & Grenier, H. (2024, January). *Novel methodology for comparing student pilot flight performance with instructor pilot expert ratings*. AIAA SCITECH 2024 Forum.

Karaca, Y., & Moonis, M. (2022). Shannon entropy-based complexity quantification of nonlinear stochastic process: Diagnostic and predictive spatiotemporal uncertainty of multiple sclerosis subgroups. In Y. Karaca, D. Baleanu, Y.-D. Zhang, O. Gervasi, & M. Moonis (Eds.), *Multichaos, fractal and multi-fractional artificial intelligence of different complex systems* (pp. 231–245). Academic Press. https://doi.org/10.1016/B978-0-323-90032-4.00018-3

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*(7), 865–878. https://doi.org/10.1016/j.cpr.2010.03.001

Kinney, L., & O'Hare, D. (2020). Responding to an unexpected in-flight event: Physiological arousal, information processing, and performance. *Human Factors*, *62*(5), 737–750. https://doi.org/10.1177/0018720819854830

Koglbauer, I., Kallus, K., Braunstingl, R., & Boucsein, W. (2011). Recovery training in simulator improves performance and psychophysiological state of pilots during simulated and real visual flight rules flight. *The International Journal of Aviation Psychology*, *21*(4), 307–324. https://doi.org/10.1080/10508414.2011.606741

Krahenbuhl, G. S., Darst, P. W., Marett, J. R., Reuther, L. C., Constable, S. H., Swinford, M. E., & Reid, G. B. (1981). Instructor pilot teaching behavior and student pilot stress in flight training. *Aviation, Space, and Environmental Medicine*, *52*(10), 594–597.

Krone, T., Albers, C. J., Kuppens, P., & Timmerman, M. E. (2018). A multivariate statistical model for emotion dynamics. *Emotion*, *18*(5), 739–754. https://doi.org/10.1037/emo0000384

Kuppens, P. (2015). It's About Time: A Special Section on Affect Dynamics. Emotion Review, *7*(4), 297–300. https://doi.org/10.1177/1754073915590947

Lajoie, S. P. (2017). Learning science applications for research in medicine. In L. Lin & J. M. Spector (Eds.), *The sciences of learning and instructional design* (pp. 108–118). Routledge. https://doi.org/10.4324/9781315684444-8

Lajoie, S. P. (2021). Multimedia learning with simulations. In L. Fiorella & R. E. Mayer (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 461–471). Cambridge University Press. https://doi.org/10.1017/9781108894333.048

Lajoie, S. P., & Gube, M. (2018). Adaptive expertise in medical education: Accelerating learning trajectories by fostering self-regulated learning. *Medical Teacher*, *40*(8), 809–812. https://doi.org/10.1080/0142159X.2018.1485886

Lajoie, S. P., Li, S., & Zheng, J. (2023). The functional roles of metacognitive judgement and emotion in predicting clinical reasoning performance with a computer simulated environment. *Interactive Learning Environments*, *31*(6), 3464–3475. https://doi.org/10.1080/10494820.2021.1931347

Lajoie, S. P., Zheng, J., Li, S., Jarrell, A., & Gube, M. (2021). Examining the interplay of affect and self-regulation in the context of clinical reasoning. *Learning and Instruction*, *72*, 101219. https://doi.org/10.1016/j.learninstruc.2019.101219

Landman, A., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., Mulder, M., & Mulder, M. (2017). 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., van Middelaar, S. H., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2020). The effectiveness of a mnemonic-type startle and surprise management procedure for pilots. *The International Journal of Aerospace Psychology*, *30*(3–4), 104–118. https://doi.org/10.1080/24721840.2020.1763798

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer. https://doi.org/10.1007/978-1-4419-1005-9_215

Li, S., Zheng, J., & Lajoie, S. P. (2021). The frequency of emotions and emotion variability in self-regulated learning: What matters to task performance? *Frontline Learning Research*, *9*(4), 76–91. https://doi.org/10.14786/flr.v9i4.901

Li, S., Zheng, J., Lajoie, S. P., & Wiseman, J. (2021). Examining the relationship between emotion variability, self-regulated learning, and task performance in an intelligent tutoring system. *Educational Technology Research and Development*, *69*(2), 673–692. https://doi.org/10.1007/s11423-021-09980-9

Li, T., & Lajoie, S. (2021). Predicting aviation training performance with multimodal affective inferences. *International Journal of Training and Development*, *25*(3), 301–315. https://doi.org/10.1111/ijtd.12232

Loijens, L., & Krips, O. (2021). FaceReader methodology note. Noldus Information Technology.

Loijens, L., Krips, O., van Kuilenburg, H., den Uyl, M., & Ivan, P. (2015). FaceReader version 6.1 Reference manual. Noldus Information Technology.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Marques, E., Carim, G., Campbell, C., & Lohmann, G. (2023). Ab initio flight training: A systematic literature review. *The International Journal of Aerospace Psychology*, *33*(2), 99–119. https://doi.org/10.1080/24721840.2022.2162405

Martins, A. P. G. (2016). A review of important cognitive concepts in aviation. *Aviation*, *20*(2), 65–84. https://doi.org/10.3846/16487788.2016.1196559

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd Ed.). Sage.

Naismith, L. M., & Lajoie, S. P. (2018). Motivation and emotion predict medical students' attention to computer-based feedback. *Advances in Health Sciences Education*, *23*(3), 465–485. https://doi.org/10.1007/s10459-017-9806-x

Nomura, O., Wiseman, J., Sunohara, M., Akatsu, H., & Lajoie, S. P. (2021). Japanese medical learners' achievement emotions: Accounting for culture in translating western medical educational theories and instruments into an Asian context. *Advances in Health Sciences Education*, *26*(4), 1255–1276. https://doi.org/10.1007/s10459-021-10048-9

Pekrun, R. (2019). Achievement emotions: A control-value theory perspective. In R. Patulny, A. Bellocchi, R. Olson, S. Khorana, J. McKenzie, & M. Peterie (Eds.), *Emotions in late modernity* (pp. 142–157). Routledge.

Pekrun, R., & Perry, R. P. (2014). Control-value theory of achievement emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 120–141). Routledge.

Rosa, E., Gronkvist, M., Kolegard, R., Dahlstrom, N., Knez, I., Ljung, R., & Willander, J. (2021). Fatigue, emotion, and cognitive performance in simulated long-duration, single-piloted flight missions. *Aerospace Medicine and Human Performance*, *92*(9), 710–719. https://doi.org/10.3357/AMHP.5798.2021

Rosa, E., Lyskov, E., Gronkvist, M., Kolegard, R., Dahlstrom, N., Knez, I., Ljung, R., & Willander, J. (2022). Cognitive performance, fatigue, emotional, and physiological strains in simulated long-duration flight missions. *Military Psychology*, *34*(2), 224–236. https://doi.org/10.1080/08995605.2021.1989236

Ruiz Segura, A. (2020). Examining the interaction between clinical reasoning and emotion on medical students' diagnostic efficiency [Master's thesis, McGill University]. https://escholarship.mcgill.ca/concern/theses/5h73q1579

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Sassenus, S., Van den Bossche, P., & Poels, K. (2022). When stress becomes shared: Exploring the emergence of team stress. *Cognition, Technology & Work*, *24*, 537–556. https://doi.org/10.1007/s10111-022-00698-z

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Skibniewski, F. W., Dziuda, L., Baran, P. M., Krej, M. K., Guzowski, S., Piotrowski, M. A., & Truszczynski, O. E. (2015). Preliminary results of the LF/HF ratio as an indicator for estimating difficulty level of flight tasks. *Aerospace Medicine and Human Performance*, *86*(6), 518–523. https://doi.org/10.3357/AMHP.4087.2015

Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2012). The everyday emotional experience of adults with major depressive disorder: Examining emotional instability, inertia, and reactivity. *Journal of Abnormal Psychology*, *121*(4), 819–829. https://doi.org/10.1037/a0027978

Tichon, J. G., Wallis, G., Riek, S., & Mavin, T. (2014). Physiological measurement of anxiety to evaluate performance in simulation training. *Cognition, Technology & Work*, *16*(2), 203–210.

Transport Canada. (2019, November 4). *Standard 421—Flight Crew Permits, Licences and Ratings—Canadian Aviation Regulations* (*CARs*). AARBH 15472583; AARBH. https://tc.canada.ca/en/corporate-services/acts-regulations/list-regulations/canadian-aviation-regulations-sor-96-433/standards/standard-421-flight-crew-permits-licences-ratings-canadian-aviation-regulations-cars

Trull, T. J., Solhan, M. B., Tragesser, S. L., Jahng, S., Wood, P. K., Piasecki, T. M., & Watson, D. (2008). Affective instability: Measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of Abnormal Psychology*, *117*(3), 647–661. https://doi.org/10.1037/a0012532

Vine, S. J., Uiga, L., Lavric, A., Moore, L. J., Tsaneva-Atanasova, K., & Wilson, M. R. (2015). Individual reactions to stress predict performance during a critical aviation incident. *Anxiety, Stress & Coping: An International Journal*, *28*(4), 467–477.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology & Psychology*, *18*, 459–482. https://doi.org/10.1002/cne.920180503

Zafeiriou, S., Zhang, C., & Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, *138*, 1–24. https://doi.org/10.1016/j.cviu.2015.03.015

Zheng, J., Li, S., & Lajoie, S. P. (2023). A review of measurements and techniques to study emotion dynamics in learning. In V. Kovanovic, R. Azevedo, D. C. Gibson, & D. Ifenthaler (Eds.), *Unobtrusive observations of learning in digital environments* (pp. 7–29). Springer. https://doi.org/10.1007/978-3-031-30992-2_2

How to cite this article: Ruiz-Segura, A., Law, A., Jennings, S., Bourgon, A., Churchill, E., & Lajoie, S. (2024). Flight emotions unleashed: Navigating training phases and difficulty levels in simulated flying. *Journal of Computer Assisted Learning*, 1–22. https://doi.org/10.1111/jcal.13037