Multimodal Affect Assessment in Aviation Training

Tianshu Li

Department of Educational and Counselling Psychology

McGill University, Montreal, Canada

July 2020

A thesis submitted to McGill University

in partial fulfillment of the requirements of the degree of

Master of Arts in Educational Psychology

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Dr. Susanne Lajoie, without whom this research would not be possible. Her unwavering support and professional guidance throughout my study have helped me grow as a researcher. Her mentorship and her way with people demonstrated to me how to be a constructive and respectable leader. Furthermore, I cannot thank her enough for the opportunity to study and work with an amazing team of skillful researchers and trainees, the ATLAS lab. I thank her for always being an inspiration and glue to our team. Many of the team members have helped me in various aspects of my research, namely Dr. Imene Jraidi, Dr. Maher Chaouachi, Alejandra Ruiz Segura, Leo Holton, Kirsten Nynych, Shan Li, Lingyun Huang, Jenny Zheng and Dr. Amanda Jarrell. I am grateful for all their help and support. I would like to thank all my participants for this study. Their time, effort and feedback are greatly appreciated.

Secondly, I thank all collaborators in the InLook Project (CAE Inc.; Marinvent Inc.; GlobVision Inc., the Design Lab, Concordia University; National Research Council; the Faubert Lab, University of Montreal) which this study is part of. I thank Alain Bourgon and Hugh Grenier (CAE Inc.) for proposing and configuring the X-plane environment and tasks. I thank our project manager, David Bowness and the principle investigator, Dr. Yong Zeng for the opportunity to work on this innovative and interdisciplinary project.

Thirdly, I thank my parents, my grandparents and friends for always supporting my passion for psychology. In particular, my parents always respect my choices in life and unwaveringly support me in what I do. I could not have started this journey in psychology and research without them. Their love and care always push me through difficult times.

Last but not least, this study is jointly funded by Natural Sciences and Engineering Research Council of Canada (514052-17), Consortium for Aerospace Research and Innovation in Canada (CARIC) and Quebec (CRIAQ) (OPR-1618). I sincerely thank the funding agencies for providing me and other like-minded researchers with valuable opportunities to learn, grow and contribute to the fields of our passion.

ABSTRACT

Affect has an impact on learning by influencing various cognitive, psychomotor and motivational processes. This research aims to understand the role of affect in aviation training. We examined participants' (N = 19) affect in simulated aviation training while they perform ten aviation tasks designed to their skill level. Performance was graded with a rubric provided by our collaborator, CAE Inc. Affective states were inferred from two biometric measurements (electrodermal activity, facial expression) and one 'ground truth' measurement (experiential self-report). Convergent results were found among all the data channels: Arousal from facial expression and electrodermal activity correlated positively with each other; Correlations between biometrically inferred affect and self-reported affect correlates (workload, fatigue, effort, perceived control and value) were found and were consistent with past research. These convergent findings support the validity of the measurements. Furthermore, we found that arousal (inferred from electrodermal activity) positively predicted performance in low difficulty task, and that mental workload (as measured from self-report) negatively predicted performance in medium and difficult tasks. We found that emotions did not vary significantly by how well a task was performed.

This research adds empirical evidence on the impact of affective states on aviation training performance. Furthermore it addresses a gap in the literature regarding mechanisms to demonstrate convergence between multimodal affect data. This research is a preliminary step to a comprehensive assessment of affect in aviation training, which could eventually help training instructors to allocate appropriate scaffolding to pilots where needed, thereby improving the overall training experience and pilot performance.

RÉSUMÉ

L'affect a un impact sur l'apprentissage en influençant divers processus cognitifs, psychomoteurs et motivationnels. Cette recherche vise à concevoir une évaluation multimodale de l'affect et à comprendre le rôle de l'affect dans la formation aéronautique. Nous avons examiné l'effet des participants (N = 19) dans une formation simulée en aviation alors qu'ils effectuent dix tâches aéronautiques conçues en fonction de leur niveau de compétence. La performance a été notée à l'aide de rubriques fournies par notre collaborateur, CAE Inc. Les états affectifs ont été déduits de deux mesures biométriques (activité électrodermique, expression faciale) et d'une mesure de «vérité fondée» (auto-évaluation expérientielle). Nous avons trouvé des résultats convergents parmi tous les canaux de données: l'excitation de l'expression faciale et l'activité électrodermique sont en corrélation positive les unes avec les autres; Des corrélations cohérentes avec les recherches antérieures se trouvent entre l'affect inféré biométriquement et les corrélats d'affect autodéclarés (charge de travail, fatigue, effort, contrôle perçu et valeur). Ces résultats confirment la validité des mesures. De plus, nous avons constaté que l'excitation (déduite de l'activité électrodermique) prédisait positivement les performances dans les tâches de faible difficulté, et que la charge de travail mental (mesurée à partir de l'auto-évaluation) prédisait négativement les performances dans les tâches moyennes et difficiles. Nous avons constaté que les émotions ne variaient pas de façon significative selon la qualité de l'exécution d'une tâche.

Cette recherche comble le manque dans la littérature concernant les mécanismes pour démontrer la convergence entre les données d'affect multimodales. Nous ajoutons des preuves empiriques sur l'impact des états affectifs sur la performance de la formation en aviation. Cette recherche est une étape empirique vers une évaluation complète de l'affect dans la formation

aéronautique, qui pourrait éventuellement aider les instructeurs à répartir les échafaudages appropriés en cas de besoin, améliorant ainsi l'expérience et l'efficacité globales de la formation.

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

Affect is a complex psychological construct that can be defined as emotional states, stress responses and moods (Gross, 2015). It has an important yet complex role in cognitive, psychomotor and motivational processes, influencing human thought and action (Mandler, 1989). Affect embodies the information of appraisal, which is one's valuation over the object of focus. Therefore, affect can alter the way we think about and perceive the object, leading to differences in cognitive, motivational and behavioral outcomes (Storbeck & Clore, 2008).

The complex interplay between affect and other cognitive, motivational and behavioral processes suggests that affect can have a strong impact on learning. Education researchers have demonstrated the influence of affect on learner's current and future learning outcomes. For instance, negative emotions experienced during mathematics homework sessions negatively predict students effort and mathematics achievement (Dettmers et al., 2011); similar functions of negative emotions are found in science problem solving, whereas positive emotions such as happiness facilitates problem-solving events (D'Mello, Lehman, Pekrun & Graesser, 2012; Lehman, D'Mello & Graesser, 2012). Furthermore, in second language learning, positive emotions and emotion regulation improves motivation and adaptive interaction with peers (Dewaele, 2011).

Although the examples above demonstrate consensus on the role of affect in different fields, verifying these findings in other contexts is important as affect is domain-specific (e.g. Mathematics) and task specific (e.g. algebra, geometry) (Pekrun & Perry, 2014). As students' valuation system and self-concept differ for each subject or task, their affective experience also differs (Storbeck & Clore, 2008). Affect is time-sensitive because it changes dynamically and thus can vary based on the moment of assessment (Pekrun & Perry, 2014). Therefore, it is important to understand students' task-related, domain specific affect in real time to determine the relationship between affect and learning.

A critical goal in educational psychology research is to find ways to support learning by accurately assessing the influence of affect in the learning process. Affect's role is especially pertinent in aviation training, during which trainees are required to execute proper psychomotor skills and make decisions under stressful situations (Kaempf & Klein, 2017). In current training programs, affect is only assessed subjectively by pilot instructors who are simultaneously tasked with giving appropriate instructions, scaffolding trainees and evaluating their performance. Hence, there is a great need for an appropriate methodology to measure affect dynamically during aviation training.

The recent surge of affective computing technologies brings alternatives to the traditional self-report measurement of affect, even providing the option to interpret affective states in real time. These affective computing technologies target different aspects of affect. One measurement could be more sensitive to a certain aspect of affect than another. For instance, facial expression analysis infers emotional states from behavioral changes manifested in facial expression, whereas electrodermal activity (EDA) and heart rate variability (HRV) simply assess the physiological arousal of the participant. Facial expression analysis is more sensitive to the valence dimension of affect, whereas EDA is more sensitive to the arousal aspect. However, affective states, particularly emotions are manifested through behavior, physiological changes and experiences (Gross, 2015). To choose only one of these methods would lead to partial and incomplete inferences of affective states. Multimodal measurements align with the definition of affect as it. However, multimodal measurements need to be carefully chosen to reduce costs and time incurred in data collection and analyses. Furthermore, multimodal data should be aligned and converged to generate reliable and interpretable results.

In this study, students' affect in a simulated aviation training setting was examined. Considering the multi-componential manifestations of affect and its situation-specificity (Gross, 2002, 2015), this research aims to develop a multimodal protocol for assessing affect in the context of simulated aviation training. This protocol includes measures of physiological, behavioral and experiential affective states in a simulated aviation context. Affect is operationalized as stress responses inferred from physiological measurement (electrodermal activity), basic emotional states inferred from facial expression analysis and experiential feedback measured through self-reports. Emotions and stress are of our research interest particularly because they are among the most common causes of fatal accidents in aviation, and they are fast-changing reaction to specific objects of focus (Jensen, 2017). Results from the various data channels are interpreted and synthesized through convergence analysis.

This research is conducted in collaboration with CAE Inc., a world leader in producing aviation simulation and training programs. We conducted an experiment using flying tasks in X-plane, a flight simulation software commonly employed for pilot training.

This research addresses the gap in the literature regarding mechanisms to demonstrate convergence between multimodal affect data. The protocol developed could be used for future aviation training. A comprehensive assessment of affect could help aviation training instructors to allocate appropriate scaffolding to pilots where needed, thereby improve the overall training experience and effectiveness.

CHAPTER 2: LITERATURE REVIEW

In an effort to ground the current research on affect in an aviation context, a literature review conducted in 3 parts: (a) theoretical background on affect; (b) methdological approaches to studying affect; and (c) exisitng research on affect in aviation.

Theoretical Background on Affect

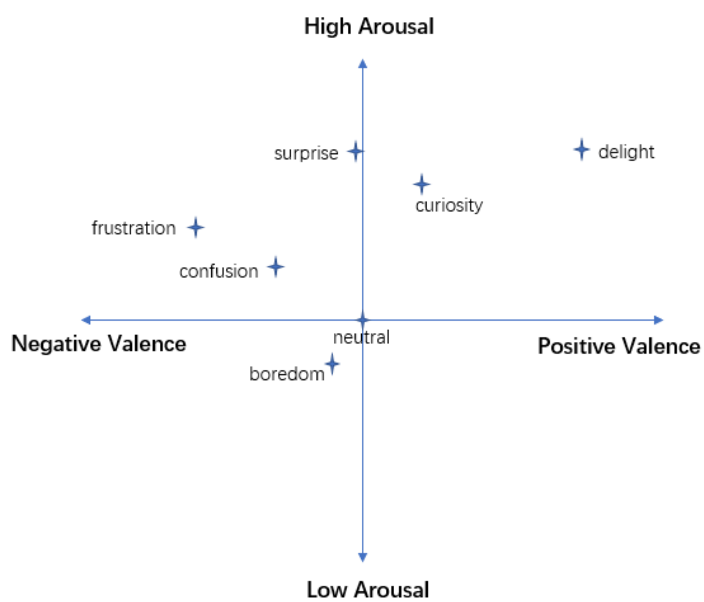
Following Gross (2015), this thesis will use affect as an umbrella term for stress responses, moods and emotional states. Emotional states and stress responses often involve whole-body reactions to a specific object of focus, while moods are the overarching affective 'climate' which do not necessarily have a specific object of focus (Gross, 2015). Emotional states can be described in two dimensions: valence and arousal. Valence is the positivity of the affective experience, ranging from negative, such as anger, to positive, such as happy. Positive valence, such as the experience of pleasantness, could lead to motivational and cognitive changes, eventually influencing learning experience positively (Pekrun & Linnenbrink-Garcia, 2014a). Arousal is the activation of the sympathetic nervous system. Coupled with various physiological changes, arousal has also been proven to influence psychomotor function, temporarily altering the limit of how fast and accurately one could react to stimuli. Therefore, psychomotor function could be compromised under excessive or insufficient arousal (Duffy, 1957, 1962). Stress responses are reaction to a highly-taxing circumstances. They are often considered as negative in valence (Gross, 2015). In particular, emotional states and stress responses have attracted much research interest in the field of experimental psychology because they are fast to vary with the changes in on the object of focus (Gross, 2015). Therefore, this thesis focuses on the discussion of emotions and stress among all affective states as they are among the most common causes of fatal accidents in aviation (Jensen, 2017).

Some researchers have studied affect in the valence and arousal dimensions separately or emphasized one over another (Lazarus, 1991; Ortony et al., 1988), while others argue for the importance of incorporating both to understand affect more fully (Barrett, 1998; Russell et al., 1989). Although positive affect is pleasant to experience, valence of affect alone does not determine the desirability of its impact. For instance, confusion, which has a negative valence, could motivate thinking and problem solving (D'Mello & Graesser, 2012a). On the other hand, relief and relaxation are pleasant states to experience, but their effect on motivation and learning could be complex and ambivalent (Pekrun & Linnenbrink-Garcia, 2014a).

Theories on discrete emotions offer the frameworks to examine affective states along two dimensions: discrete emotions are distinguishable by their embodied valence and arousal, and hence can be mapped on a two-dimensional scale, as demonstrated below in Figure 1.

Figure 1

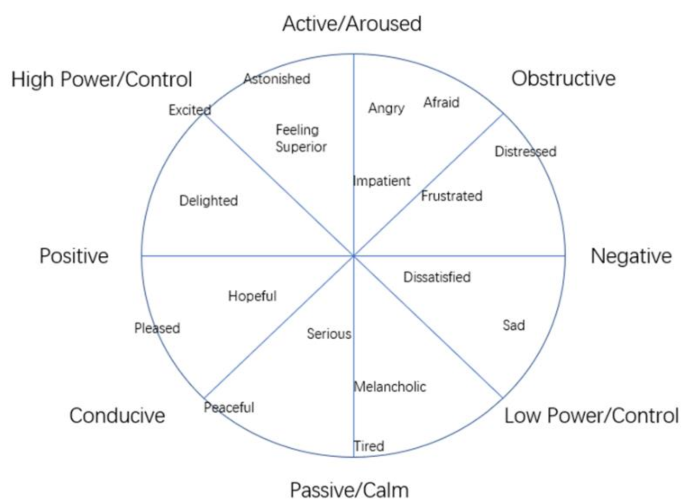
Discrete Emotions Mapped on Valence and Arousal Scales (Hussain et al., 2011)



Other theorists describe discrete emotions as part of more complex systems that carry additional information on appraisal. For instance, as demonstrated in Figure 2, the Geneva Emotion Wheel (GEW) incorporates appraisal to the structure of discrete emotions, resulting in two additional dimensions: coping potential (power or control) and goal conduciveness (whether the situation help or hinder goal attainment) (Scherer, 2005). The GEW also emphasizes the importance of evaluating emotions by their intensity.

Figure 2

Geneva Emotion Wheel (Scherer, 2005)



In the next section, we further review the role of the appraisal process in experiencing emotions.

Appraisal Theories of Affect

Emotions arise from an appraisal process, which is an evaluation process of the context and how it relates to one's own goals, values, interests and needs (Arnold, 1960; Jarrell, 2015). As discussed previously, the appraisal process provides valuable information about the function of affect, potentially addressing why individuals could react differently to the same situation. Decades of research have established abundant empirical evidence on the strong reciprocal relationship between appraisal, affect and performance (Davis et al., 2008; Hembree, 1988; Zeidner, 1998). The control-value theory of acheievement emotions (Pekrun & Perry, 2014) provides explanatory power for how appraisal and affect relate to performance and learning.

Control-Value Theory of Achievement Emotions

The control-value theory (CVT) is a prominent framework of affect and appraisal in the context of learning (Pekrun & Perry, 2014). It explains how affective responses can be activated by achievement activities, which relate to performance, competency and evaluation. These affective responses are categorized as achievement emotions. Enjoyment, pride, anxiety, boredom and anger are some examples of achievement emotions. The CVT provides a conceptual structure of the origin, situational specificity, functions and regulation of achievement-related affective states. In addition to the two dimensions of emotion (valence and arousal) as mentioned previously, the CVT proposes a three-dimensional taxonomy of emotions by integrating the dimension of object of focus (prospective, activity, retrospective). For instance, anxiety is a prospective negative activating emotion instigated by the expectation of low controllability of an activity that one values highly. The perceived controllability and value of the situation are components of appraisal (one's subjective perception of the object of focus). This CVT proposes that perceived control and value are of primary importance in understanding the antecedent of affect because other factors of appraisal (e.g., goal structure, gender, culture) indirectly influence affect by influencing control and value first.

The CVT addresses how affect influences achievement or performance through various cognitive and motivational processes. Specifically, enjoyment of the achievement activity helps to focus attention, support motivation and prevents forgetting of coherent material and promotes flexible and deep cognitive strategies (Csikszentmihalyi, 1997); whereas anxiety and boredom reduces attention on the activity, undermines motivation and promotes rigid memory retrieval and superficial information processing (Bäuml & Kuhbandner, 2007; Pekrun et al., 2010; Pekrun & Perry, 2014).

The CVT acknowledges that the relationship across environment, appraisal, emotion and achievement is reciprocal. Therefore, performance feedbacks strengthen the relationship across appraisal, affect and performance due to the cumulative effect of appraisal. The CVT is pertinent to the analysis of aviation training and performance as aviation tasks are achievement activities. During a flying task, pilots in-command execute actions based on their skill, knowledge and judgement to achieve well-defined goals (e.g. land at the destination airport safely). Pilots constantly receive feedback from the aircraft system (the altitude meter, the compass etc.) about whether their actions contribute to achieving their goals (Endsley, 1995). For instance, when pilots are required to perform a sharp turn, they might experience a low sense of control if they expect the situation to be difficult to manage. A feeling of anxiety and stress could be induced by low perceived control, and will be intensified if the outcome is of high value (e.g. safety concerns for themselves and others). Anxiety, is a negative activating emotion that could exacerbate the situation by reducing pilots' deep-thinking ability and hampering their psychomotor functions (Panganiban et al., 2011b; Barrett et al., 2019; Duffy, 1962). Thus, trainees' affect should be interpreted from an appraisal perspective to promote training efficiency and wellbeing in aviation.

Theories and definitions of affect have demonstrated that reliable affect measurements are needed to account for the multi-facetted nature of affect. Information on appraisal contributes to making appropriate inferences about affect, which is especially important when conducting multimodal experiments (Arnold, 1960; Jarrell, 2015). In the next section, we review the various methodologies to measure affect to inform our design of a multimodal affect assessment in aviation.

Methodological Approaches to Studying Affect

Affect is a multi-faceted construct coupled with behavioral changes, physiological responses and subjective experience (Gross, 2002, 2015). Measurement of affect could be operationalized by measuring its multiple manifestations. As the field of computer science advances, various affective computing technologies have been invented to infer affect through its manifestations. However, measuring an affect from a single facet will result in partial inferences, as cues manifested in other ways would be overlooked. An impartial measure of affect should account for affective cues expressed in all three dimensions (behavioral, physiological and experiential). A valid and trustworthy inference of affect should be a result of convergence analysis and triangulation across multiple sources. The following sections review the design of several common measurements of affect and data convergence procedures.

Behavioral Measurement of Affect: Facial Expression

Facial expression analysis is a common behavioral assessment technique for affect detection (Cohn, 2006; Niedenthal et al., 2000). There are cross-cultural similarities in some emotional facial expressions either involuntarily elicited by emotions or deliberately displayed to reflect emotions (Ekman, 1997). These facial expressions are associated with six universal emotions: fear, anger, disgust, sadness, enjoyment or joy, surprise or startle. Ekman postulated an empirically validated guide for coding universal facial expressions: Facial Action Coding System (FACS) (1997). The FACS describes visually distinguishable facial expressions in terms of 44 action units. An action unit is the contraction and relaxation of one or a group of facial muscles, e.g. inner-brow raiser. The FACS also provides Ekman's interpretation of some facial expressions, which indicates six basic emotions. This comprehensive guide allows coders to distinguish facial expressions by manually grading 44 action units on 5-point intensity scales. This guide also distinguishes some involuntary and voluntarily posed facial expression, e.g. a sincere smile of joy versus a fake smile. The FACS has been widely used in many applied contexts for both clinical and experimental purposes. It also pushed the development of affective computing software to automate facial expression recognition. For instance, FaceReader (Den Uyl & Van Kuilenburg, 2005) is a commercial software made to automatically detect facial action units and produce interpretations of basic emotions. It has been validated with other affect measurements such as electromyography and has been used in various empirical studies (D'Arcey, 2013; Lewinski et al., 2014; Terzis et al., 2010).

Physiological Measurement of Affect: Electrodermal Activity

Although facial expression has become a common indicator of affect, not all affective states of all intensity are distinguishable visually. Ekman et al. (1997) found that the sub-visible emotions are detectable physiologically through electromyography (EMG). This finding is consistent with the theoretical definition of affect and its physiological manifestation. The arousal dimension of affect can be reflected physiologically. The physiological component of affect has been measured through EMG, heart rate variability (HRV) and electro-dermal activity (EDA). However, EMG and HRV are rather laborious to obtain and intrusive to measure due to the sensor placements. Furthermore, muscle movements and parasympathetic nervous system could corrupt the inference of the reaction for the sympathetic nervous system reactions.

Electrodermal activity (EDA) is a common, non-invasive method to measure the physiological expression of arousal. Changes in arousal is manifested in the change of the moisture level of the skin. The change in moisture leads to fluctuations of skin’s electrical conductivity. EDA recording tracks the magnitude of the electrical current between two points on the skin (Braithwaite, Watson et al. 2013, Harley 2016). Hence, one's arousal state could be inferred by analyzing the EDA data. This method is superior to other physiological measures of arousal such as heart rate variability for its short latency time and independence from contamination of parasympathetic nervous system activities (Braithwaite, Watson et al. 2013).

To eliminate the confounding effect of individual and environmental differences on skin conductivity, EDA could be processed into two elements: skin conductance level (SCL) and skin conductance response (SCR). Skin conductance level is a smooth, ever-changing baseline of skin conductivity, whereas SCR is the rapid changes in skin conductivity that is most probably caused by fluctuation in arousal level. Each significant increase of skin conductivity from SCL is a SCR peak. SCR peak frequency and amplitude are both common indicators of arousal level (Braithwaite, Watson et al. 2013). More frequent SCR peaks are observed when arousal level increases. An SCR peak with high amplitude indicates a sudden surge of stress.

Experiential Measurement of Affect: Self-reports as the 'Ground truth'

Self-reports assess affect in the experiential dimension, which reflects the subjective experience of the participant. Experiential measures include questionnaires and interviews. Some questionnaires may assess affect by directly asking the participant to rate his or her experience of every affective states of interest. The Achievement Emotion Questionnaire is one of the common affect questionnaires used in educational psychology. For instance, enjoyment is assessed through items such as 'I enjoy acquiring knowledge' (Pekrun et al., 2011).

Other relevant self-report measures of affect could be questionnaires on the correlates of affect, such as appraisal and mental workload. Appraisal has a direct, reciprocal relationship with affective experience (Pekrun & Perry, 2014). As discussed previously in Appraisal Theories of

Affect, affect embodies information of the valuation system and the perceived control over the object of focus. Perceived control and perceived value are two components of appraisal, which could be measured by the Motivated Strategies for Learning Questionnaire (Pintrich, 1991). It measures three types of value: utility, relevance and interest. For instance, an item to assess interest value is 'I feel interested in this activity'.

In addition to capturing the experiential dimension of affect, self-reports are often employed for data convergence and alignment in multimodal studies which involves multiple data channels. The following section explains the importance and the procedure of data convergence.

Multimodal Data Convergence

Although behavioral and physiological measurements of affect complement each other in capturing the biometrics cues of affect, it is challenging to empirically converge them and generate a valid understanding of the affective states at the time of measurement. As each data channel reflects a single manifestation of affect, our inference on affect only becomes impartial through triangulation and convergence among data channels. For instance, by simply observing raw multimodal data, we might find that the facial expression of anger occurs in absence of any concurrent rise of raw EDA. This observation is inconsistent with theories of affect which suggest anger as an activating emotion should be coupled with an increase in physiological arousal. To generate accurate inference and resolve conflicts between data channels, several steps are involved in converging multiple sources of affective data. First, data need to be accurately aligned based on time. Latency time should be considered for each mode of measurement. Facial expression is considered as an instant-onset behavioral cue of affect. However, latency of electro-dermal activities is close to one second. This means that a change in

EDA only happens approximately one second after the onset of the stimulus. Aligning facial expression inference and EDA without consideration of latency will result in conflicts and inaccurate interpretation of the affective states at the time point.

Second, the object of focus should be the same among all modes of measurements to converge the data. During experimentation, it is essential to control the experimental condition to help participants focus on the stimuli or the experimental task. We can then assume that the object of focus of the affective states we detect are regarding the concurrent feeling of performing our experimental tasks. Some researchers may choose to increase the difficulty or mental demand of the task to enhance engagement (Jensen, 1965; Kane et al., 2007). Such experiment designs include incrementing requirements (e.g. an inclination versus an inclination with a fixed pitch-up angle) and setting time limit. Other studies may use immersive technologies and set up the experiment in an immersive, simulated environment with little distraction (Freina & Ott, 2015). For instance, trainees may be more focused when they are immersed in a full simulation lab that creates an illusion or a real flight than in a classroom. These two steps could reduce the presence of noise in affective data.

After the process of data convergence, internal validity is created if there are few conflicts among the data sources and the affective states. For instance, by definition, anger is an activating emotion, hence it is expected to be associated with higher arousal and higher SCR. If multimodal data shows that anger as observed in facial expression and the data align with low SCR, then a conflict is found across data sources. To practice caution in interpretation, such conflicts would require examination of the protocol, the video recording of the experiment and the ‘ground truth’.

A ‘ground truth’ is crucial in multimodal assessments to validate the other modalities of data, especially in the phase of protocol or data analysis. Ground truth refers to a chosen source of data, typically video recording or self-report. Other sources of data in a multimodal study is often compared to the ‘ground truth’ for convergence assessment and conflict resolution. A ground truth could be instrumental to interpret and resolve conflicts occurred during data convergence. As human affect experience could be complex, even neatly controlled experiments could have unforeseen conflicts during convergence of data sources. Therefore, a ground truth could provide the additional context needed for researcher to interpret affect data. Self-reports, interviews and qualitative video analysis are often chosen as ‘ground truths’ in multimodal studies (Harley, 2016).

In sum, the process of data convergence is an essential step in multi-channel data analysis. Data alignment enhances accuracy of affect interpretation in comparison to single-sourced experiment. The integration of ‘ground truth’ in multimodal experiments provides resolutions to conflicts across data channels and validates measurements in novel or exploratory experiment protocols. In the next section, we review the extent to wheich multimodal protocol design and existing theoretical frameworks and assessments of affect, have been conducted in aviation contexts.

Affect in Aviation

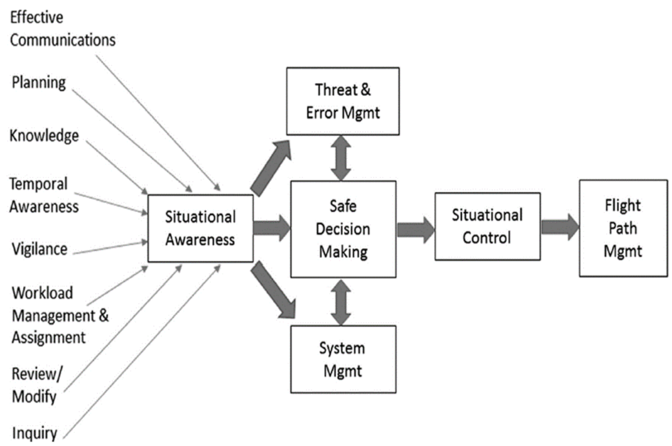
In the field of aviation, negative affect can have a detrimental effect on performance. Stress and emotions are among the six main causes for fatal accidents in aviation (Jensen, 2017). There is a consensus that to safely control a dynamic environment in and around the aircraft, pilots need to be aware of the situation and execute effective decisions with minimal risk (Jensen, 1997; Murray & Martin, 2012). Such pilot-related factors have been studied in the context of aviation through human factor (ergonomics) perspectives, which is the application of psychological or physiological theories to the analysis of products, processes or systems (Wickens et al., 2009). Two of the most influential models that examine essential human factors for flight safety are the Model of Situation Control and the Aeronautic Decision-Making Model. These models are reviewed to examine how the role of affect is defined in the aviation decision making and pilot performance.

Model of Situation Control in Aviation

In aviation, pilots in command must be aware of the environment, the goal and the dilemma presented to them. The model of situation control (Figure 3) outlines the essential intrinsic factors for pilots in command to manage a flight path safely and effectively (Murray & Martin, 2012). Pilots need to be aware of the situation, which includes the aircraft environment, the outer environment (wind, weather etc.), the goal (targeted flight path) and any dilemma or challenge (e.g. encountering unexpected turbulence during landing). In addition to sufficient technical skills, pilots also need to engage their metacognitive skills (planning) and interpersonal skills to communicate with the flight crew and the Air Traffic Control Center. As shown in Figure 3, situation awareness is a prerequisite for further steps to achieve flight path management. Therefore situation awareness influences threat management, decision making and flight system management. Although affect management and regulation is not explicitly listed as a factor of situation awareness, it could be interpreted that affect and affect regulation has an overarching effect on many factors in the Model of Situation Control: it is implicit that stress management is needed for flight management, as stress has a strong impact on workload and vigilance (Warm et al., 2018). Furthermore, affective states such as boredom and anxiety can influence situation awareness and threat management by impacting workload management and vigilance (Panganiban et al., 2011a; Sawin & Scerbo, 1995).

Figure 3

Situation Control Model (Murray & Martin, 2012)



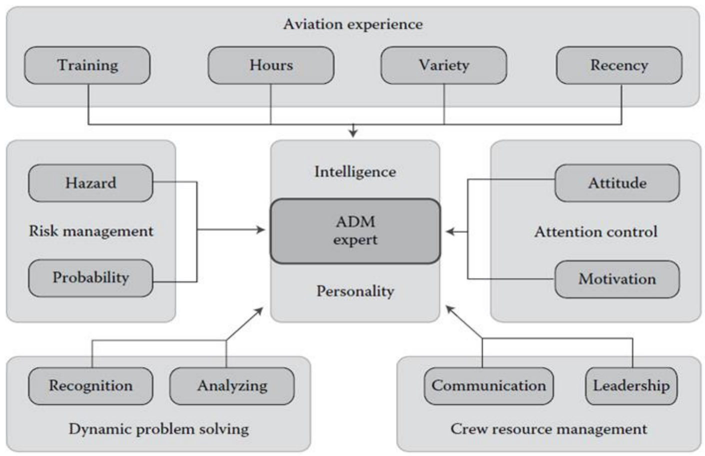
The model of Situation Control focuses on the factors influencing situation awareness as it stems from the Situation Awareness Model. The concept of situation control builds on situation awareness by integrating other factors such as decision making and threat management. In particular, decision making is a complex process that require further dissemination (Starcke & Brand, 2012). The Aeronautic Decision-Making Model is reviewed in the following section to examine the role of affect in pilots' decision-making process.

Aeronautic Decision-Making Model

Factors influencing the decision-making process of a pilot in command is structured in the Aeronautic Decision-Making Model (Figure 4). The expert aeronautic decision-making process involves five elements, stable intrinsic characters (intelligence and personality), aviation experience, risk management, dynamic problem-solving, crew resource management and attention control to make an aeronautic decision (see Figure 4) (Kochan et al., 1997). This model addresses some affective correlates such as motivation and attitude in terms of how these factors influence attention control. However, like the Model of Situation Control, affect and affect regulation are not explicitly discussed in the model.

Figure 4

Aeronautic Decision-Making Model (Jensen, 1997)



A common criticism of human factor frameworks in aviation is that they operate with the assumption that the expert-like decision making process should be independent of affect (Soll et al., 2016). On the other hand, the broader literature in psychology stipulates that affect could arise in various processes of decision making (Lerner et al., 2015). The ability and willingness to regulate these affective states are essential to maximize decision making efficiency. For instance, in problem solving, maladaptive affect often couples with reduced cognitive resources and task disengagement, which lead to failure in problem solving (D'Mello & Graesser, 2012a). In crew-resource management, maintaining a positive social-emotional environment is crucial for the team to work towards the same goal effectively (Bakhtiar et al., 2018). A theoretical framework that reconsiders the role of affect in the aviation context is lacking. There is a demand for systematic studies on affect in aviation to build a better understanding of how affect influences pilots' expertise. To empirically study the effect of affective states in aviation, a valid and feasible measure of affect is needed. The design and selection of affect measures should also be compatible with the aviation context. As aviation training takes place in simulators of various fidelity levels, often involving whole body motion in tight spaces, affect measures need to be non-invasive with minimal effect on the aviation training context, and non-disruptive of the training process (Lee, 2017). The next section presents relevant literature on the use of biometrics measures of affect in aviation contexts.

Biometrics Measures of Affect in Aviation

Biometrics are commonly referred to as the technology to identify people by measuring physiological and behavioral traits (Li, 2009). Popular biometrics technologies include facial expression analysis, voice analysis, fingerprint matching etc. Many of the biometrics technologies have been adapted to use in affective computing. In comparison to traditional self-report measure of affect measurements, biometrics measurements have the advantage of being non-disruptive of task performance (Conati et al., 2003; Jones & Troen, 2007). Therefore, researchers could use biometrics measures of affect concurrently with task performance. To determine which biometrics measures of affect would be feasible and effective in aviation training setting, the existing uses of biometrics measurements in aviation context are reviewed.

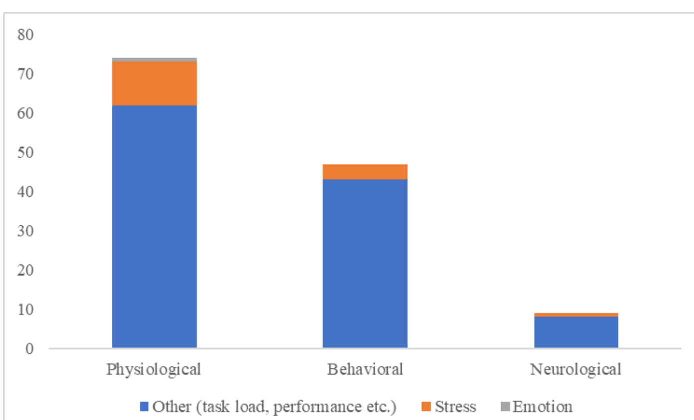
Due to the inter-disciplinary nature of this project, this part of the literature review was performed using two databases, Scopus (Burnham, 2006) and PsycINFO, to thoroughly examine the relevant research conducted both within and beyond the field of psychology. Scopus is an interdisciplinary database providing access to over 14,000 journal articles from over 4,000 publishers. The literature review in the Scopus database was conducted to examine the interdisciplinary research that used biometrics assessments in aviation contexts. This literature review was done to inform the experiment design and measurement selection. Keywords were chosen in the three dimensions of interest for this project. First, for the dimension of aviation context, phrases ‘aircraft pilot’ and ‘airplane pilot’ were entered in the search. For the purpose of reviewing research in similar task context, the word ‘driver’ was included, although eventually very few papers were identified with this keyword. Secondly, for the multimodal dimension of the project, search phrases ‘Face reader’, ‘facial expression analysis’, ‘voice analysis’, ‘galvanic skin response’, ‘electrodermal assessment’, ‘heart rate’, ‘eye fixation’, ‘visual attention’, ‘gaze behavior’, ‘visual tracking’, eye tracking’, ‘pupil dilation’, ‘posture analysis’ were included in the search.

Although Scopus provides access to articles in various domains, ranging from health sciences to social sciences, it does not guarantee complete inclusion of all articles published in these fields. Articles from certain publishers may be included in more specialized databases instead. Therefore, the same search is conducted in a database specialized in psychology research, PsycINFO. PsycINFO database is a robust academic database for psychology related literature (Leininger, 2000). The joint search yielded 228 peer-reviewed papers from scopus and 71 from PsycINFO. Due to the small amount of search results, the search results were not narrowed down to papers reporting affect-related research. Alternatively, the title and abstract of these articles were manually scanned to screen for relevant papers that used biometrics measures to infer affect and its correlates (attitude, mental state or performance) in the aviation or driving contexts. Finally, 120 papers were selected from Scopus and 22 are selected from PsycINFO. The review affirms that research on affect in aviation is still in its infancy and that affective computing and biometrics measurement has yet to be more integrated more fully into research that sheds light on the role of affect in aviation training.

The review of selected papers from the two databases revealed that physiological measures such as heart rate variability, electro-dermal activity and blood pressure, are the most frequently used measures in aviation research on affect: (a) 74 papers (69 from Scopus, 4 from PsycINFO) employed physiological measures (e.g. Hormeño-Holgado & Clemente-Suárez, 2019; Mansikka et al., 2018), (b) 47 papers (39 from Scopus, 8 from PsycINFO) used behavioral measures including speech analysis, eye-tracking and posture analysis (e.g. Kuroda et al., 1976; Tole et al., 1982) and; (c) 9 papers used the neurological measure, electro-encephalogram (EEG) (e.g. Saproo et al., 2016) . However, of all these papers only 17 of these papers (10 from Scopus, 7 from PsycINFO) examined affect through biometrics measures (e.g. (Hannula et al., 2008; Kuroda et al., 1976)). Among them, 16 papers measured stress (10 from Scopus and 6 from PsycINFO) (e.g. Doorey et al., 2011; Main et al., 2017; Regula et al., 2014). These statistics inform us that research on the use of biometrics measures to infer affective states in pilot training is still rare, especially in the measurement of emotional states that are relevant in learning context (see Figure 5). Furthermore, there is a lack of exploration using multimodal protocols to address the changes of multiple components of affect (behavioral, physiological and experiential) in aviation tasks.

Figure 5

Diagram of Comparison of the Number of Peer-reviewed Articles on Biometrics Measures in Aviation Context



However, from the selected measures, cognitive load is a construct that is assessed frequently in the aviation context. Thirty papers (29 from Scopus, 1 from PsycINFO) out of 142 measured task load or cognitive load (e.g. Diaz-Piedra et al., 2019; Mansikka et al., 2018). This construct is considered as one of the most important human factors in aviation. The research demonstrates that cognitive overload is associated with poor performance and learning of the aviation tasks as insufficient cognitive resources are available for the respective tasks (van Erp et al., 2007). Although in the aviation context, several terms such as task load, cognitive workload, engagement and stress seem to be taken as equivalent or interchangeable (Flin et al., 2003), there is a commonly used measure for them: NASA task load index (NASA TLX) (Hart, 2006). NASA TLX is a validated indicator of cognitive load, physical workload and temporal workload in aviation tasks. It is frequently used in aviation research that pertains to pilot decision making and performance. As reviewed previously, workload has a strong association with affective states, particularly stress (Warm et al., 2018). Consequently, NASA TLX could be a valuable measurement to include in multimodal assessment of affect as it is a validated indicator of an affect correlate.

In conclusion, this chapter reviewed the recent findings on the importance of multimodal assessment for affect and the need for affect assessment in aviation context. This literature review serves to inform the design of the current multimodal protocol, to define the research questions and to establish evidence-based hypotheses.

Research Questions and Hypotheses

This thesis addresses the gap in both psychology and affective computing literature on how to measure affect through multiple data channels. To acknowledge the demand on theoretical and empirical understanding of the role of affect in aircraft pilot training, this study aims to use multimodal assessment to examine affect in aviation training context and to understand the role of affect in aviation training performance. This thesis examines the role of affect in terms of the valence-arousal dimensions and the occurrence of discrete emotions. A series of three research questions are answered regarding the validity of the protocol, the predictive value of affect for aviation performance, and the differences between high and low performers in terms of their discrete emotion profiles. Each question is outlined below with its corresponding hypotheses.

Research Question 1

The first research question is: can affective inferences obtained from physiological indicators, behavioral cues and subjective self-report align with each other in an aviation training context? In other words, do these multimodal indicators of affect show convergence? This research question was answered by exploring the data collected using a multimodal protocol designed based on the multi-faceted definition of affect. In this protocol, the physiological aspect of affect was measured through electrodermal activity. The behavioral dimension was inferred from facial expression analysis. The affective experience was examined through self-report questionnaires on affect correlates (workload, effort, fatigue, perceived control and value). Three sub-questions were posed to determine if the protocol described above demonstrates converging evidence of affect inferences in aviation training. These sub-questions pertain to the correlations among arousal inference from EDA, discrete emotion inferred from facial expression and affect correlates inferred from self-reports. The three sub-questions and their corresponding hypotheses are described in the following sections.

Research Question 1A

The following research question is posed: does arousal inferred from EDA correlate with arousal analyzed from facial expression? This question aims to investigate if EDA and facial expression analysis generate agreeing inferences on arousal responses during aviation training. Agreement across data channels with overlapping constructs of interest is one of the indicators of data convergence (Harley et al., 2015). In the current multimodal protocol, the arousal dimension of affect could be inferred both physiologically through EDA and behaviorally through facial expression. Therefore, a correlation analysis was conducted between arousal inferred from EDA and that inferred from facial expression.

Hypothesis 1A

As previously mentioned in Chapter 2, high SCR (high SCR peak amplitude, high SCR peak frequency and high phasic EDA) indicates elevated arousal. From facial expression, arousal level could also be inferred from the arousal dimension of the discrete emotions identified in facial expressions (i.e. facial expressions of anger indicate higher arousal than facial expressions of sadness). Therefore, the hypothesis is that tasks with high SCR will demonstrate activating emotions through facial expression. The opposite result is expected for tasks demonstrating low SCR.

Research Question 1B

The second sub-question is posed: is there correlation between the self-reported affect correlates (workload, effort, fatigue, perceived control and value) and physiological arousal inferred from SCR during aviation tasks? Self-reported affect correlates serve as the ‘ground truth' of this multimodal protocol. A 'ground truth' measurement is often included in multimodal protocols to assess data convergence and alignment (Harley, 2016). As the 'ground truth' is a validated measurement, other measurements could be validated if they provide results that are consistent with existing theories or empirical research and align with the 'ground truth'.

Hypothesis 1B

According to previous research on task load in aviation, arousal increase is coupled with workload, fatigue and effort increases (Nittala et al., 2018; Shiomi et al., 2012). Therefore, it is expected that arousal inferred from SCR increases as the subjective ratings on the aforementioned affect correlates increase.

Research Question 1C

The third sub-question is: are there correlations between the discrete emotions inferred from facial expression (happy, surprised, angry, sad, scared, disgusted) and the self-reported affect correlates (workload, effort, fatigue, perceived control and value)? This question examines the data convergence between facial expression analysis and the 'ground truth'. We also aim to provide evidence on affect data convergence along the valence dimension inferred from facial expression analysis of discrete emotions.

Hypothesis 1C

Based on previous research, it is expected that facial expressions indicating negative, activating emotions appear more frequently as self-reported workload, fatigue and effort increases (Nittala et al., 2018; Shiomi et al., 2012). A negative relationship is expected between negative activating emotions and perceived control and value. Opposite trends are expected for positive emotions (Causse et al., 2013; D'Mello & Graesser, 2012b; Pekrun & Linnenbrink-Garcia, 2014b). It is expected that the facial expression of the negative deactivating emotion in this study (sad) becomes more frequent as perceived control decreases (Pekrun & Perry, 2014). I take an exploratory stand on the relationship between negative deactivating emotion and workload, fatigue and effort as it is not systematically studied in the aviation context.

Research Question 2

The second research question asks: How do affective valence, arousal and workload during aviation training together and separately influence performance?

The role of affect in learning and performance outcomes (e.g. examination grades) have been explored in various contexts (D'Mello & Graesser, 2012a; Dettmers et al., 2011). Theories suggest that emotions should be studied along two dimensions: valence (positive to negative) and arousal (activating to deactivating) (Lazarus, 1991; Ortony et al., 1988; Russell et al., 1989). This research question is posed to examine if the existing theoretical and empirical evidence on the functions of affect is generalizable to the aviation training context. Moreover, this research aims to add empirical evidence regarding the relationship between aviation training performance and workload, an important determinant of performance based on previous research in aviation psychology (Hart, 2006; Mansikka et al., 2018; Roscoe, 1978; Wickens et al., 2002).

Hypothesis 2

Based on previous research (D'Mello & Graesser, 2012a; Dettmers et al., 2011), it is expected that positivity of overall affective valence during the task is associated with performance excellence. An exploratory stand is taken on the relationship between arousal and performance as there are mixed findings on the directionality of this relationship in the aviation psychology literature (Stokes & Kite, 2017). According to existing evidence on the hampering effect of cognitive overload on performance (Mansikka et al., 2018; Roscoe, 1978; Wickens et al., 2002), it is expected that workload predicts performance negatively.

Research Question 3

The third research question is: are there differences between the basic emotions expressed (happy, sad, angry, disgusted, surprised, scared) during high performance and that of low performance?

According to educational psychology theories and human factors models in aviation (Knecht et al., 2016; Pekrun & Linnenbrink-Garcia, 2014a; Pekrun et al., 2017), knowledge, skill execution as well as emotion competency and regulation should be considered in tandem to determine their effects on learning and performance. This research question is posed with the long-term goal to help instructors to identify disruptive emotional profiles in pilot trainees and allocate sufficient emotional regulation training and support for these trainees. Affective states can be structured along dimensions of valence and arousal, but they also embody intricate cognitive information such as appraisal, resulting in different discrete emotions (i.e. anger, sadness). Each discrete emotion has a unique set of parameters that pertain to arousal, valence and appraisal (Barrett, 1998; Hussain et al., 2011; Scherer, 2005). Therefore, in addition to studying the role of affect by its arousal and valence dimension, this research aims to examine if discrete emotions are associated with differences in performance.

Hypothesis 3

Based on previous research on affect in other learning contexts, it is expected that happy occurs more in high performance, and sad, angry, disgusted, scared are expressed more often in low performance (D'Mello & Graesser, 2012a; Dettmers et al., 2011; Pekrun & Linnenbrink-Garcia, 2014a). I take an exploratory stand on surprise due to the mixed findings on its function in various context (Mauss & Robinson, 2009; Muis et al., 2015).

CHAPTER 3: METHODOLOGY

The experimental environment and tasks were designed in collaboration with subject-matter experts from CAE Inc. We conducted an experiment using flight tasks in X-Plane ®, flight simulation software that is commonly employed for pilot training. X-Plane aviation tasks were adapted to the beginner pilot trainee level by subject-matter experts from CAE Inc. to fit the skill level of our participants. X-Plane was loaded on Dell computers. Our data collection protocol included three-modality features: (a) physiological arousal was inferred from electro-dermal activity (EDA) recorded with BioNomadix® system; (b) The behavioral cues of affect were collected using Microsoft Lifecam HD cameras mounted on the computer screens to capture facial expressions during training and analyzed through facial expression inferences from FaceReader 6.0®; and (c) The experiential manifestation of affect was assessed through a self-report questionnaire administrated on a laptop. Self-reports also served as the ‘ground truth’ measure for the convergence analysis. These measures are described in the Measurements section. This study received IRB approval from Concordia University and was accepted by McGill University based on the CREPUQ inter-institutional agreement.

Participants

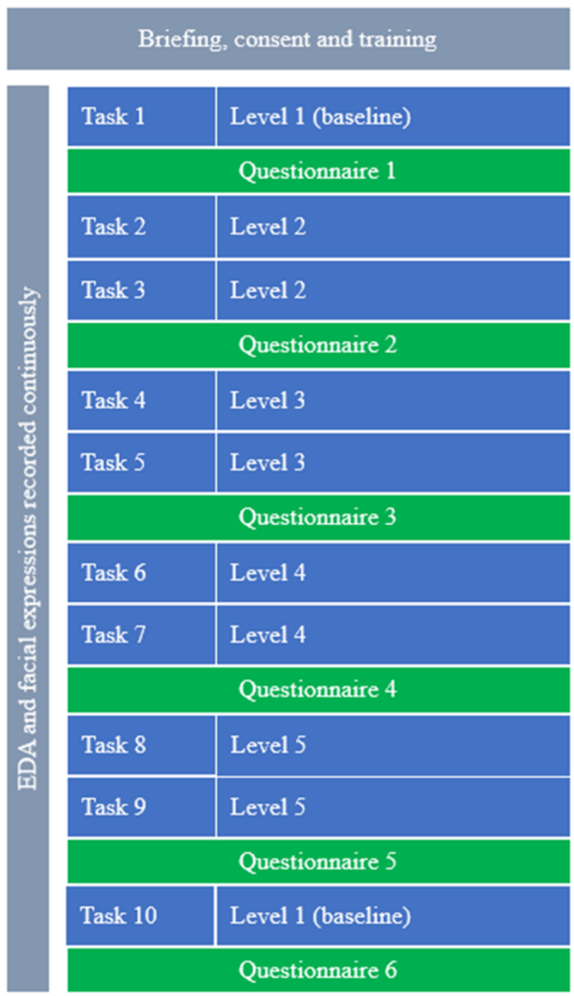
A total of 19 participants (11 females) were recruited from the undergraduate and graduate student population at McGill University with an age range from 19 to 35. The sample included several ethnic backgrounds: 5 Asians, 6 Caucausians, 2 Hispanics and 5 from other ethnicities. The current data collection was used to serve as a baseline for the larger pilot project and thus participants were selected who had limited to no aviation-related experience (private pilot license, aviation simulation gaming experience etc.), similar to beginner trainees. The simulation interface (X-Plane) and flying tasks are adapted to beginner level. Each participant was compensated ten dollars per hour for their participation.

Procedure

Figure 6 documents the experimental procedure. The duration of the experiment ranges from forty minutes to ninety minutes depending on the time required for the participants to complete training and tasks. Prior to the start of the experiment, participants were first given a briefing on the task and objectives of the experiment. Afterwards, participants provided consent and completed a demographic questionnaire. Next, they received instructions on how to use the X-Plane interface and the joystick for the flying tasks in this experiment. Subsequently, they went through a training session (approximately fifteen to forty minutes long) in which they practiced the flying maneuvers involved in the experiment (making turns, pitch up or down etc.). Once participants managed to return to the baseline parameters after making all the maneuvers, experimenters put BioNomadix EDA bracelets on participants, adjust the camera used for facial expression recording and perform calibration for both BioNomadix EDA bracelets and FaceReader software. Afterwards, the experimental tasks would start, while facial expression, EDA and X-Plane log file for performance analysis were recorded concurrently. There were ten tasks in total, including eight tasks of increasing difficulty and two tasks with minimal difficulty as baseline (see Figure 6). During each task, an experimenter read the instructions for the task to the participant. The instructions provided the specific requirements for the maneuver to be performed. For instance, task 2 is ‘maintain speed and altitude, make a right turn of 15-degree banking to the heading 0.’ Participants performed eight tasks grouped in pairs. Within each pair, both tasks had the same difficulty level with the second task being the reversal task. For instance, task 3 is the reversal task of task 2, hence task 3 requires the participant to make a left turn (instead of a right turn) with the same banking angle. A self-report questionnaire pertaining to affect appraisal and taskload constructs was administered 6 times (see figure 6) to capture perceived affect correlates throughout the study. Full questionnaires used in this experiment can be found in the appendix. After each task, participants took a break if they requested one.

Figure 6

Experimental Procedure

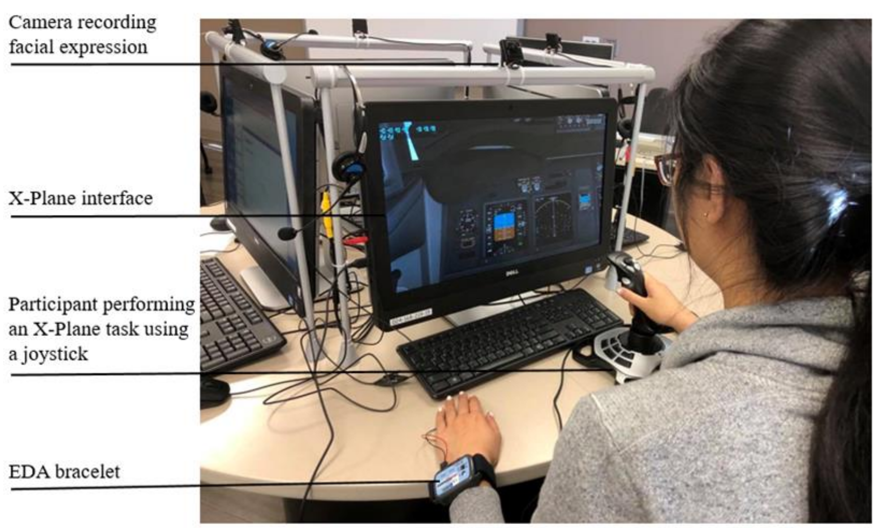


Experimental Setup

This experiment requires participants to perform aviation tasks on a computer-based flight simulation software by controlling a joystick with their dominant hand. A BioNomadix bracelet for EDA recording was placed on the non-dominant hand of the participants. Video recording of facial expression was capture by an external camera placed on top of the computer screen where X-Plane was displayed. A demonstration of this setup can be found in Figure 11. Participants performed the experimental tasks individually in a closed room with no external noise. One experimenter sat next to the participant to provide training and give verbal instruction at the start and end of tasks. Another experimenter monitored the recordings of video and EDA from another computer station, further away from the participant.

Figure 7

Experiment Set-up

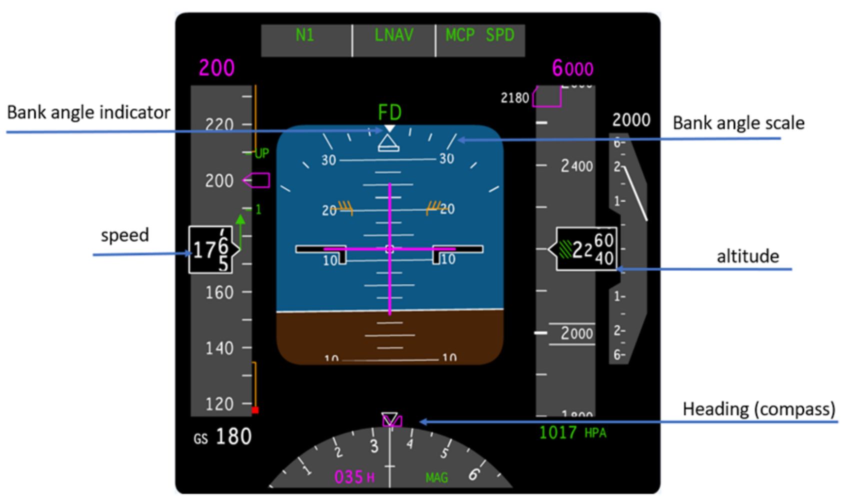


Experimental Tasks

The experimental tasks in this study were administrated on X-Plane, a computer-based aviation simulation with realistic simulation of aircrafts. It is a training environment often employed for trainees in the early phase of their training. The X-Plane tasks selected for this study challenges participants on maintaining and manipulating three basic variables of aircraft flying: altitude, heading (direction) and speed. The easiest task (level 1) requires no manipulation: participants were simply required to maintaine a course by holding the joystick steadily and observing the values as labeled in Figure 12. According to subject-matter experts in CAE Inc., the more variables to manipulate, the higher mental resources are required, as participants need to focus on all the variables simultaneously by cross-checking and making timely adjustments. A change of one variable might affect another: if any variable is overlooked for too long, the aircraft will ‘drift’ on that variable. For instance, to making a turn, the aircraft need to be tilted. A simultaneous drift in altitude will occur if it is not controlled as the aircraft turns. The task difficulty increases as the number of dimensions to manipulate increase. For instance, a task on changing altitude and making a turn is more difficult than a task with just a turn. Changes of speed is commonly deemed as more difficult than any other single variable change, as there are multiple ways to achieve it (by adjusting the engine power or by manipulating altitude in various pitch angles), and it destabilizes the aircraft. A detailed list of the ten tasks created by pilot instructors from CAE Inc. can be found in appendix v.

Figure 8

X-Plane Interface



Measurements

Demographic Questionnaire

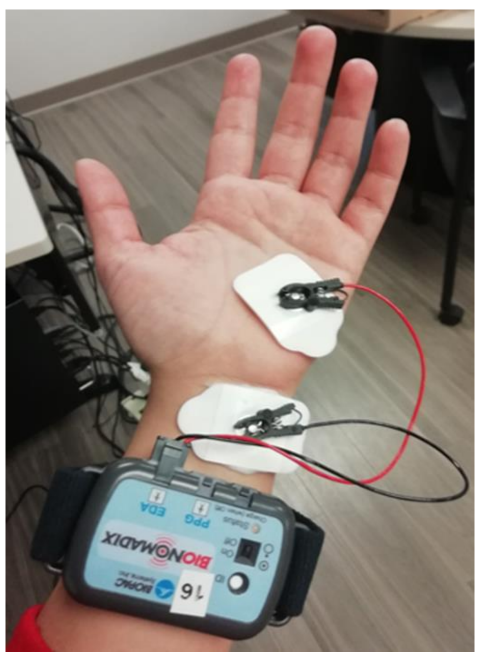
The demographic questionnaire was administed after the briefing and includes items on gender, ethnicity, age and past related experience. The items of this questionnaire are presented in the appendix i.

Physiological Arousal: Electro-dermal Activity and BioPac®

Electrodermal activity was measured by BioNomadix® EDA module (Braithwaite, Watson et al. 2013) with sampling rate set as 1000 Hz. As demonstrated in Figure 7, a bracelet was placed on the wrist of the participant’s non-dominant hand. Two electrodes are placed on the inner palm and the wrist of the hand. Electrical current flowing between these two points were recorded.

Figure 9

BioNomadix EDA Sensor and Bracelet Placement

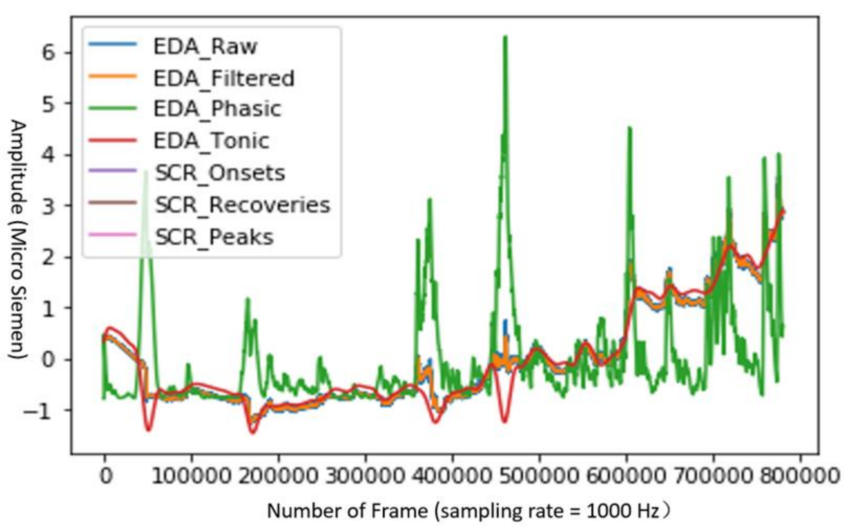


Normalized skin conductance response features (SCR) and skin conductance level were extracted through Makowski's algorithm for EDA processing, implemented in a python-based toolkit, Neurokit (Makowski 2016). Figure 8 demonstrates a plot of the extracted features by time. SCR, otherwise referred to as phasic EDA, is the rapid changes of the electrical conductivity of the skin. It is a product of the fluctuation of the sympathetic arousal level (Boucsein, 2012; Braithwaite et al., 2013). The SCR features we extracted are mean level of SCR per task (phasic), total number of SCR peaks (significant increases of skin conductance) per task, and mean amplitude of SCR peaks per task. The frequency of SCR peaks per minute was computed by the formula below.

Frequency of SCR peaks per minute in task 1= $\frac{\text{SCR peak count in task 1}}{\text{duration of task 1 (s)}} \times 60$

Figure 10

Diagram of Raw and Processed EDA Signal by Time, Generated by Neurokit



Behavioral Cues: Facial Expression and FaceReader 6.0®

Facial expression was recorded by a Microsoft LifeCam hd5000® camera during the experiment with frame rate at 30 Hz. The recording was processed using commercial software

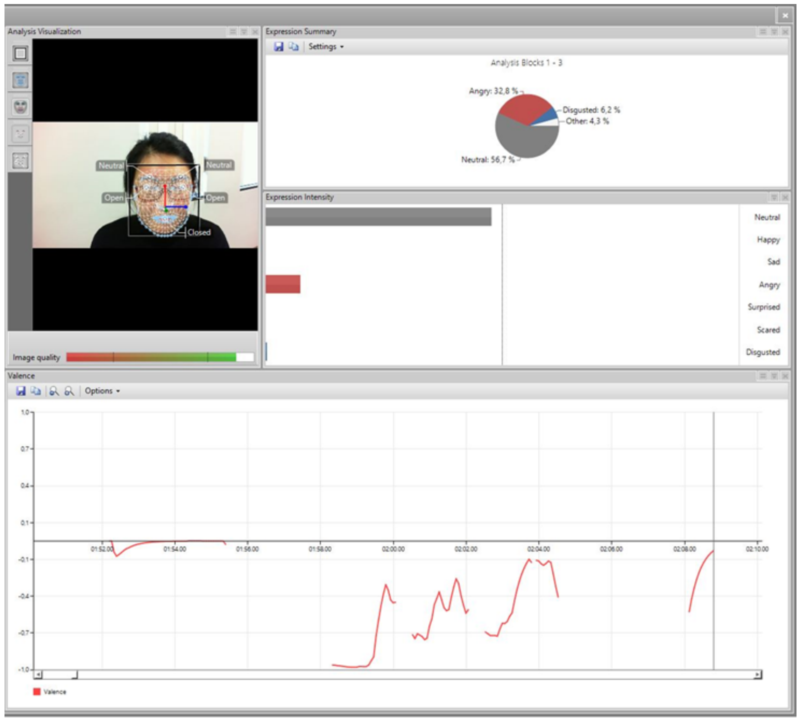
FaceReader 6.0® (Den Uyl & Van Kuilenburg, 2005) every second frame (sampling rate 15 Hz). For each participant, at each task the proportion of each basic emotion (happy, surprised, neutral, sad, angry, scared and disgusted) was analyzed along with the overall intensity of valence and that of arousal (described below). Figure 9 demonstrates the FaceReader interface as a video is processed and analyzed. A significant emotion state is recorded when the intensity of one emotion is consistently higher than all other emotions for more than 5 seconds. The proportions of significant emotions in each video analysis are automatically generated by FaceReader. To obtain the proportion of each emotion in each task, we processed a detailed FaceReader logfile segmented by task. The procedure of by-task segmentation is explained in section of data alignment. The proportion of each emotion was calculated by the count of this emotion divided by the total count of all emotions in this task, as illustrated in the formula below.

$$\text{proportion of emotion A in task 1} = \frac{count\ of\ emotion\ A\ in\ task\ 1}{count\ of\ all\ emotions\ in\ task\ 1}$$

A valence intensity is calculated in FaceReader at each frame by the sum of intensities of all positive emotions subtracting that of all negative emotions. An arousal intensity is calculated by FaceReader at each frame by the sum of intensities of all activating emotions subtracting that of all deactivating emotions. A mean valence and mean arousal per frame are calculated for each participant at each task. For instance, in the case where the intensity of ‘Happy’, ‘Sad’, ‘Angry’, ‘Scared’ and ‘Disgusted’ are 0.8, 0.1, 0.0, 0.05 and 0.05 respectively, the valence. is 0.7 (Noldus FaceReader 6.0, 2015).

Figure 11

FaceReader 6.0 Video Analysis Interface



Experiential Measurement: Subjective Self-reports as 'Ground Truth'

Subjective self-reports assess affect from its experiential dimension. The subjective self-reports in this study include two parts: 1) an appraisal questionnaire on perceived control (Perry et al., 2005) and value (Wigfield & Eccles, 2000a) of the task 2) NASA Task Load Index (NASA TLX). Both the appraisal questionnaire and the NASA TLX are completed between tasks.

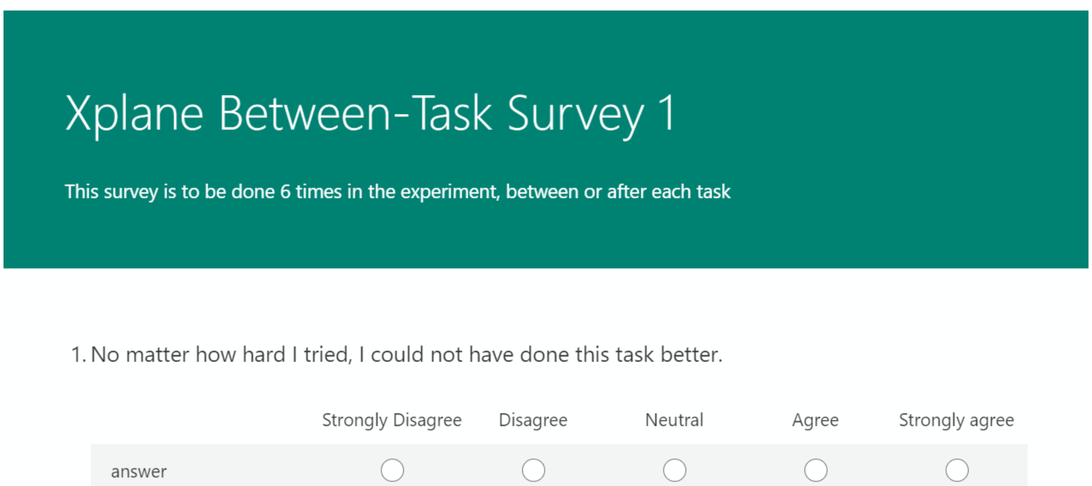
To assess appraisal, questionnaire items on perceived value were adapted from Wigfield and Eccles' expectancy-value theory of achievement motivation (2000b). Two items were used for each of the three dimensions of perceived value: utility value, interest and importance. Items on perceived control are adapted from the Perceived Control Scale (Perry et al., 2005). Two items were used to assess perceived control. The questionnaire items used in this study are listed in the appendix ii.

The NASA TLX is a questionnaire established to assess the subjective experience of workload related elements such as mental workload and fatigue. It is especially adapted to the aviation context where both physical and mental tasks are required (Hart, 2006). For our experiment, we included mental workload, physical workload, fatigue and effort. The item on temporal workload is not used for this study as there was no time limit for the experimental tasks. NASA TLX includes one item per construct with explanations, definitions and rephrases of each item attached. The items included in this study could be found in the appendix iii.

Self-reports are used as the 'ground truth' measurement in this study to converge and validate output from EDA and facial expression analysis. All self-report questionnaires are administrated online on a laptop during the experiment as demonstrated in Figure 10. The participants rate their experience using 5-point Likert scales on relevant activities such as aviation simulation, driving and video games involving joysticks. All questionnaires can be found in the appendices i - iii.

Figure 12

Example of a Question Administrated in Online Questionnaire



Performance Measurement

Performance is scored based on a rubric developed by subject-matter experts from CAE Inc. to evaluate beginner trainees, as attached in appendix iv. In aviation training, two dimensions of performance are often evaluated: 1. Management of the aircraft environment throughout the task. 2. Accuracy and timeliness of task completion. Currently, there is no automated evaluation system for aviation training. The management dimension is often scored based on professional instructor's overall impression throughout out the practice. Instructors score the accuracy dimension by assessing the final altitude, heading and speed and how much they differ from the targets. Hence, we developed a data mining protocol that only scores the performance accuracy of aviation task performance: First, the altitude, heading and speed at the end of each task was extracted from X-Plane log files. No time limit was set for each task to isolate task difficulty as the stressor or stimuli of affective changes in the experiment. A task was considered complete when the participants were satisfied with their metrics or when they gave up on the task. Since it is very difficult to achieve the exact value of altitude, heading and speed as instructed, performance accuracy was evaluated as the only measure of performance in this experiment. Participants' metrics (altitude, heading, speed) at the end of each task were compared with the targeted metrics to obtain performance accuracy on each metric. A test rubric provided by subject matter experts from CAE Inc. was used to grade performance accuracy. The overall performance accuracy score was calculated by averaging the accuracy grade on all three metrics.

Data Alignment and Processing

As the analyses in this thesis requires by-task comparisons, data output from each channel was aligned by time and segmented by task. As participants were instructed to press a ‘timer’ button on the joystick each time they start and finish a task, the start and end time for each task was logged in X-Plane. Using the timer logs and the timestamps generated automatically during video and EDA recording, I aligned and segmented the FaceReader output and raw EDA output by task. It is particularly important to segment the EDA recordings before the analysis, because there is a high level of irrelevant noise (movements, thought wanders…) between tasks and these noises may skew the feature extraction result. All questionnaires were pre-labeled by sequence as demonstrated in the procedure in Figure 6.

CHAPTER 4: RESULTS

This chapter presents in sequence the experimental results regarding the three research questions listed in Chapter Three.

Data Convergence

Research Question 1: Can affective inferences obtained from physiological indicators, behavioral cues and subjective self-report align with each other in an aviation training context? We aim to investigate whether affective inferences obtained from physiological indicators, behavioral cues and subjective self-report align with each other in an aviation training context. This question is answered by assessing data convergence of experimental data collected with a multimodal affect assessment designed based on the three-dimensional characteristic of affect manifestation (behavioral, physiological, experiential). Data convergence across measurements are examined by analyses regarding three sub-questions: a) relationship between physiologically inferred arousal and behaviorally inferred arousal b) relationship between physiologically inferred arousal and self-reported affect correlates c) relationship between behaviorally inferred discrete emotions and self-reported affect correlates. Data from all channels (EDA, facial expression and self-reports) are analyzed across 19 participants and 6 tasks (N = 114). Only tasks with corresponding questionnaires are analyzed (6 out of 10 tasks are analyzed) at this stage, as the other tasks were performed for future analyses which are out of the scope of this thesis. Outliers from all variables among all data channels with standardized value greater than 3.29 or smaller than -3.29 were excluded in this analysis. Descriptive statistics are demonstrated in Table 1.

Table 1

Descriptive Statistics of Facial Expression Variables, EDA features and Questionnaire Responses

Channel	Variable	N	Missing	Mean	Std. Deviation	Skewness	Kurtosis	Min	Max
Facial expression	Happy	113	1	.08	.11	1.80	3.22	0	.50
	Angry	114	0	.27	.21	.11	-1.13	0	.80
	Sad	111	3	.02	.06	4.55	22.77	0	.39
	Surprised	113	1	.07	.12	1.65	1.60	0	.44
	Scared	109	5	.01	.02	4.62	22.60	0	.13
	Disgusted	113	1	.07	.11	1.80	2.63	0	.47
	Valence	114	0	-.34	.30	-.18	-.38	-.97	.34
	Arousal	112	2	.31	.06	-.34	.97	.13	.49
EDA	SCR peak amplitude (μS)	112	2	3.87	10.03	4.34	19.87	.02	59.20
	Phasic EDA (μS)	112	2	2.79	4.77	2.64	7.85	-5.33	26.47
	SCR peak frequency (normalized)	112	2	2.63	.84	.04	-.35	.69	4.42
	SCR peak amplitude (normalized)	113	1	-.00	1.56	.83	1.25	-3.93	5.52
	Phasic EDA (normalized)	113	1	.05	1.62	.14	-.18	-3.97	4.32
	SCR peak frequency (per min)	112	2	24.31	166.48	10.56	111.66	0	1769.13
Self-report	Mental workload	114	0	3.44	1.16	-.40	-.81	1	5
	Physical workload	114	0	2.80	1.22	.49	-.94	1	5
	Effort	114	0	3.46	1.12	-.38	-.82	1	5
	Fatigue	114	0	3.23	1.14	-.06	-1.17	1	5
	Perceived value	114	0	3.56	1.01	-.64	-.26	1	5
	Perceived control	114	0	3.77	.85	-.42	-.84	2	5

Convergence between Behavioral and Physiological Dimensions of Affect

In this study, arousal can be inferred both from physiological changes and behavioral cues. EDA is a commonly used measure for physiological arousal. Discrete emotions inferred from behavioral cues such as facial expressions also contain information on arousal. Therefore, arousal as the overlapping construct across data channels is examined. Regarding to Research Question 1A, bivariate correlational analyses were conducted between means of arousal for seven facial expressions variables (overall arousal, proportion of happy, angry, sad, surprise, scared, disgusted faces) and the mean arousal inferred from EDA features (SCR peak frequency and phasic EDA). Consistent with the methodology of past research, we only examined valenced, activating or deactivating emotions (Lajoie et al., 2019; Russell, 1980). Neutral is a non-valenced emotion with medium arousal and therefore was not of research interest. As demonstrated in Table 2, significant positive correlations are found between arousal inferred from facial expression and two EDA features, namely SCR peak amplitude (r = .188, p < .05) and phasic EDA (r = .279, p < .01). This suggests that arousal expressed behaviorally (facial expression) increases as arousal inferred from physiology (EDA) increases. Consistent with Hypothesis 1A, this finding supports data convergence between facial expression analysis and EDA.   Sad is positively correlated with SCR peak amplitude (r = .280, p < .01) and phasic EDA (r = .297, p < .01).

Table 2

Bivariate Correlation between EDA Features and Facial Expression Variables

		SCR peak amplitude	Phasic EDA	SCR peak frequency
Happy	Pearson Correlation	-.001	-.097	-.072
	Sig(2-tailed)	.995	.311	.454
	N	111	111	111
Angry	Pearson Correlation	-.062	-.022	-.115
	Sig(2-tailed)	.517	.819	.229
	N	112	112	112
Sad	Pearson Correlation	.280**	.297**	-.030
	Sig(2-tailed)	.003	.002	.754

	N	109	109	109
Surprised	Pearson Correlation	-.031	.028	-.060
	Sig(2-tailed)	.747	.769	.534
	N	111	111	111
Scared	Pearson Correlation	-.017	-.032	-.030
	Sig(2-tailed)	.861	.746	.759
	N	107	107	107
Disgusted	Pearson Correlation	-.018	-.070	-.056
	Sig(2-tailed)	.849	.463	.560
	N	111	111	111
Arousal	Pearson Correlation	.188*	.279**	.030
	Sig(2-tailed)	.048	.003	.758
	N	111	110	110

** Correlation is significant at 0.01 level (2-tailed)

*Correlation is significant at 0.05 level (2-tailed)

Convergence between Physiological Arousal and Experiential Self-reports

To answer Research Question 1B, bivariate correlations between EDA features and self-reported affect correlates were performed to assess the data convergence of physiological arousal indicated from EDA and the ‘ground truth’ of affective experience. As demonstrated in Table 3, marginally significant negative correlation is found between perceived value and SCR peak frequency (r = -.185, p = .05). This result suggests that as perceived value increases, physiological arousal decreases. However, as the result is only marginally significant, it should be interpreted with caution.

Table 3

Bivariate Correlation between Physiological Arousal and Self-reported Affect Correlates

		SCR Peak Frequency	SCR Peak Amplitude	Phasic EDA
Mental Workload	Pearson Correlation	-.130	.117	.129
	Sig(2-tailed)	.171	.218	.175
	N	112	112	112
Physical Workload	Pearson Correlation	-.072	.090	.045

	Sig(2-tailed)	.452	.345	.636
	N	112	112	112
Effort	Pearson Correlation	-.138	.160	.150
	Sig(2-tailed)	.147	.092	.115
	N	112	112	112
Fatigue	Pearson Correlation	-.107	.115	-.027
	Sig(2-tailed)	.260	.229	.775
	N	112	112	112
Perceived Value	Pearson Correlation	-.185*	-.015	-.085
	Sig(2-tailed)	.050	.879	.372
	N	112	112	112
Perceived Control	Pearson Correlation	.069	-.127	-.152
	Sig(2-tailed)	.472	.181	.109
	N	112	112	112

*Correlation is significant at 0.05 level (2-tailed)

Convergence between Facial Expression and Experiential Self-report

To address Research Question 1C, bivariate correlational analysis between the emotions inferred from facial expression and the self-reported affect correlates. As demonstrated in Table 4, happy positively correlates with mental workload (r = .292, p <.01), physical workload (r = .185, p = .05) and effort (r = .238, p < .05). This suggest that participants invest more physical and mental effort in accomplishing the task as they feel more joy through the task. Furthermore, scared is negatively correlated with perceived control (r = -.221, p < .05). Arousal intensity from facial expression is positively correlated with fatigue (r = .218, p < .05) and negatively with value (r = -.224, p < .017).

Table 4

Bivariate Correlation between Behavioral Cues and Self-reported Affect Correlates

		Happy	Angry	Sad	Surprised	Scared	Disgusted	Valence	Arousal
Mental Workload	r	.292**	-.116	-.065	.011	.055	.128	-.058	.008
	Sig	.002	.220	.496	.910	.572	.177	.543	.935
	N	113	114	111	113	109	113	114	112
Physical Workload	r	.185*	.008	.100	-.017	.136	.096	-.040	-.040
	Sig	.050	.933	.298	.858	.160	.311	.675	.676
	N	113	114	111	113	109	113	114	112

Effort	r	.236*	-.113	.055	-.011	.079	.172	-.113	.109
	Sig	.012	.232	.569	.904	.416	.068	.230	.255
	N	113	114	111	113	109	113	114	112
Fatigue	r	.125	.036	.050	-.172	.026	-.071	-.156	.218*
	Sig	.187	.702	.604	.069	.792	.456	.097	.021
	N	113	114	111	113	109	113	114	112
Perceived value	r	.114	-.139	-.029	.108	-.167	.015	.044	-.224*
	Sig	.229	.141	.761	.257	.082	.871	.641	.017
	N	113	114	111	113	109	113	114	112
Perceived Control	r	.034	-.020	-.026	-.053	-.221*	.136	.127	-.073
	Sig	.722	.832	.788	.577	.021	.151	.179	.446
	N	113	114	111	113	109	113	114	112

** Correlation is significant at 0.01 level (2-tailed)

*Correlation is significant at 0.05 level (2-tailed)

The results from the three correlation analyses indicate alignment among the physiological, behavioral and subjective channels of affect measurement. The correlations are consistent with our hypotheses, which is grounded on previous research and theories. Therefore, these results support the validity of EDA and facial expression analysis as measurements for affect in aviation context.

Relationships between Affect, Workload and Performance

Research Question 2 is: How do affective valence, arousal and workload during aviation training together and separately influence performance? Multiple regression analyses were conducted to examine how affective valence, arousal and mental workload predict aviation task performance at different difficulty levels. One hierarchical multiple regression was performed for each difficulty level (easy to hard: level 1 to level 5). Valence as inferred from facial expression, arousal as inferred from phasic EDA and mental workload as reported in questionnaire were each entered as a separate step in the regression model to investigate the incremental effect each dimension on predicting performance. Conforming to the multicollinearity assumption of multiple regression, only one feature was chosen from each data channel enter the regression model: Phasic component of EDA was chosen to enter the regression models as it is an overall indicator of physiological arousal. A natural log transformation was performed on phasic EDA to adjust positive skewness of the data; Valence inferred from facial expression was chosen as the behavioral indicator of affect. The selection of predictors followed the two-dimensional (valence, arousal) definition of emotion. Mental workload was included in the predictive model as it represents the experiential dimension of cognitive and affective experience. Past research also demonstrated significant impact of it on performance (Hart, 2006). Assumption checks on outliers, multicollinearity, normality and homogeneity of variance were performed and these assumptions of multiple regression are fulfilled by the current dataset. It is expected that the physiological arousal, emotional valence and mental workload each have a significant incremental contribution to predicting performance.

No significant regression model was found for the baseline level tasks. For level 2 task, phasic EDA was a significant predictor of performance, F (1, 17) = 7.408, $p < .05$, *std* β = .551, accounting for 30.4% of the variance in performance. Therefore, for the level 2 task, every unit of change of phasic EDA led to .55 unit of change in performance. However, addition of valence and mental workload to this regression model did not bring any significant incremental improvement of the model.

For level 3 task, phasic EDA and emotional valence were not significant predictors. However, mental workload brought a significant incremental contribution to the model: F (1, 15) = 4.598, $p <.05$, *std* β = -.539, accounting for 22.1% of variance in performance, controlling for valence and phasic EDA. In a further analysis valence and phasic EDA were removed from the model and the regression model was improved: F (1, 18) = 6.517, $p < .05$, *std* β = -.526, with mental workload accounting for 22.7% of variance in performance. Therefore, for each unit of increase in mental workload, performance was compromised by .526 unit.

For the level 4 task (medium difficulty), valence and phasic EDA were not significant predictors either. Mental workload contributed to a significant incremental improvement of the model, F (1, 15) = 12.85, $p < .01$, *std* β = -.728, accounting for 43.3% of the variance in performance, controlling for valence and phasic EDA. In follow-up analysis, valence and phasic EDA were removed from the model. The model of mental workload regressing on performance yielded F (1,18) = 16.232, $p < .01$, *std* β = -.699. This time mental workload accounts for 48.8% of the variance in performance. Each unit of increase of mental workload led to .699 unit of compromise on performance. No significant predictive models were found for level 5 task.

Differences in Emotions Expressed during High versus Low Performance

Research Question 3 asks: Do high and low performers differ in emotional profiles? Chi-square tests of independence were conducted to identify if there were differences between the emotions experienced by high performers versus low performers.

Low and high performance are determined by the overall performance (mean score of altitude, speed and heading) in each task. Tasks with below median scores were grouped as low performance, while tasks with above median scores were grouped as high performance. To identify potential differences in emotion profiles across performance levels, a chi-square analysis of independence was conducted on the count of dominance for discrete emotions in high (performance score above median) and low (performance score below median) performance across 19 participants and 6 tasks (N = 114). Based on facial expression analysis output, the most frequently occurring emotion in each task was denoted as the dominant emotion of the task. Sad, disgusted and scared occurred seldomly across tasks and participants. To fulfill the assumption of chi-square that each case has at least five observations, these three rare emotions were excluded in the analysis. The counts of dominance for each discrete emotion during two performance levels are shown in Table 5. Pearson chi-square analysis yielded non-significant differences on emotion occurrence during high (pass) and low performance (fail): $\chi 2$ (3,114) = 6.848, p = .077. However, as this analysis showed medium effect (*Cramer's V* = .245, p = .077) and the p value is close to marginal significance, the differences of discrete emotions across performance levels should be examined with caution as shown in Figure 13.

Figure 13

Bar Chart Comparison of the Dominant Occurrence of Discrete Emotions (Angry, Happy, Surprised) during High and Low Performance

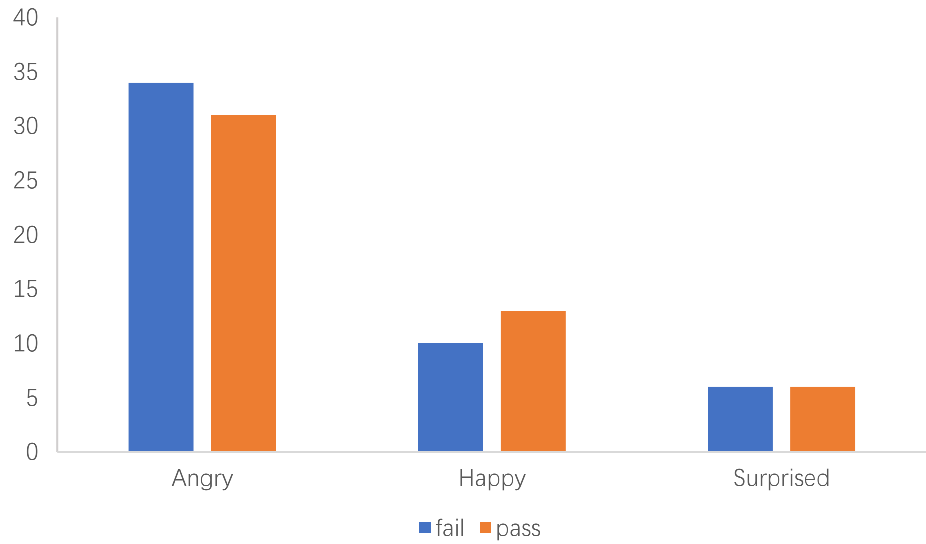


Table 5

Table of Observed and Expected Counts for Discrete Emotions Exhibited during High and Low Performance

		Performance		Total
		fail	pass	
Angry	Count	34_a	31_a	65
	Expected Count	35.4	29.6	65.0

Happy	Count	10_a	13_a	23
	Expected Count	12.5	10.5	23.0
Surprised	Count	6_a	6_a	12
	Expected Count	6.5	5.5	12.0

Each subscript letter denotes a subset of performance categories whose column proportions do not differ significantly from each other at the .05 level.

Happy more frequently emerged as the dominant emotion during high performance than in low performance. More angry dominance could be observed in low-performance tasks than in high-performance tasks. Equal dominance of surprised was observed across two performance levels.

CHAPTER 5: DISCUSSION AND CONCLUSION

The previous chapter presented the empirical results regarding the three research questions on the data convergence in multimodal protocol, the association between performance and affect (valence and arousal dimensions) and the difference in discrete emotions expressed between high and low performance. This chapter discuss the results by comparing them to the corresponding hypotheses and the past literature. Implications and limitations of this research are discussed in this chapter.

Data Convergence in a Multimodal Affect Assessment

This research addresses the gap in the literature about the role of affect in aviation. In particular, it first proposes a multimodal protocol for assessing affect during aviation training. More specifically, an experiment was conducted to test this multimodal affect assessment while participants performed simulated aviation tasks. By using a combination of behavioral, physiological, and self-report measures, data convergence is demonstrated in three folds. First, agreement was found between EDA and facial expression data on their overlapping construct, arousal. Positive correlations are discovered overall arousal from facial expression and EDA (SCR peak amplitude and phasic EDA). This is expected as an affective experience is coupled with expressions in physiology, behavior, and subjective feeling (Gross, 2015). For instance, the same feeling of anger or frustration experienced when participants repeatedly fail to achieve the altitude, heading and speed goals of the task could be expressed in the facial expression of frowning and jaw tensing (Ekman, 1992) and in the surge of SCR (Braithwaite et al., 2013). Alignment between these two data channels supports the construct validity of using EDA and facial expression as affect indicators while trainees interact with flight simulation. In other words, our result suggests that EDA and facial expression analysis are valid measurements for assessing affective arousal in aviation training context, as the results found in the physiological component were supported by the inferences from the behavioral data channel. Conflicting correlations are found for the sad facial expression and EDA: sad is positively correlated with SCR peak amplitude and phasic EDA. The directionality of this relationship is unexpected as past research categorize sad as a deactivating emotion (Pekrun & Linnenbrink-Garcia, 2014a; Scherer, 2005). Such conflicts between data channels are one of the greatest challenges in multimodal research (Harley et al., 2015). A potential explanation to these conflicts is that there is some variability of sensitivity across modalities to different dimensions of emotions. For instance, facial expression analysis is more sensitive to valence, whereas physiological channels such as EDA are more sensitive to arousal (Harley et al., 2015). Furthermore, alignment between data modalities is higher for some emotions (e.g. happy) than others (e.g. sad) (Mauss et al., 2005). Therefore, inferences from different data channels should be compared with the ‘ground truth’ of the experiment.

A ‘ground truth’ measurement is instrumental in multimodal studies as it can be used to examine alignment of other data channels and resolve conflicts among them. In this study, self-reports of workload (workload, effort fatigue) and appraisal (control and value) are chosen as the ‘ground truth’. These self-reports are validated measurements of correlates of affect. Therefore, by converging the correlation between other data channels and the ‘ground truth’, consistency of the current results with existing literature can be established. In this study, Research Question 1B and Research Question 1C aimed to examine the convergence of the biometrics data channels (EDA and facial expression) with the ‘ground truth’. SCR peak frequency correlateed negatively with perceived value, indicating that a physiological arousal increase was associated with a drop in perceived value. The same correlation was found between arousal as inferred from facial expression and perceived value. These converging findings suggest that the more activation is experienced, the less the task is valued. This correlation seems to contradict the control-value theory, which states that perceived value intensifies all achievement-related affects except boredom (Pekrun & Linnenbrink-Garcia, 2014a). However, the design of the protocol and the cyclic relationship between appraisal and affect could explain this correlation. As the control-value questionnaires were administrated after each task and the facial expressions were captured during each task, the measured perceived value of the aviation task could only be a consequence of the affects experienced during the tasks. Negative activating emotions such as anger and frustration could undermine intrinsic interest of the task, lead to disengagement and lower perceived value (D'Mello & Graesser, 2012). In the future, we would examine the relationship between appraisal of a task and the affect experienced in the following task to investigate the role of appraisal as antecedents of affect. We also hope to examine the type of values involved to further uncover the rationale behind the directionality of this relationship, further demonstrating the convergence between EDA and facial expression analysis.

Behaviorally inferred arousal was also positively correlated with fatigue. This is consistent with previous research demonstrating that prolonged stress could lead to fatigue (Shiomi et al., 2012). Furthermore, consistent with previous research (Pekrun & Perry, 2014), a scared state has a significant negative correlation with perceived control. However, happy is positively correlated with mental workload, physical workload and effort. As mentioned, the self-reports of workload, effort and fatigue were completed after each task. Hence, the task-related emotion could influence how participants engage in the task through motivational and cognitive changes. Positive activating emotions are associated with positive motivational and cognitive changes in learning. The feeling of happiness fosters the motivation to engage in the task of focus, supports attention allocation on the task performance (Pekrun & Linnenbrink-Garcia, 2014a). Therefore, participants could invest more effort and engage in higher workload as they are motivated and supported by the feeling of happiness. Although it is beyond the scope of this thesis, future research could analyze if perceived value mediates this relationship between happy and workload or effort. Our results failed to confirm our hypothesis 1C on the relationship between anger and the 'ground truth'. It was expected that anger would correlate negatively with workload, perceived value and perceived control (Pekrun & Linnenbrink-Garcia, 2014a). The conflicting results were not a result of low occurrence of anger in the current experimental context as anger was the most frequently experienced emotion (M = .26). A possible explanation is that negative emotions coupled with cognitive disequilibrium could sometimes motivate students to resolve an impasse (D'Mello & Graesser, 2012). To investigate if this is the case for anger, we have sought advice from FaceReader developers regarding this intriguing result. Through our communication with the FaceReader developers, we understood that a face of intense focus (e.g. frowning) can be coded as anger in FaceReader due to the similarity of the facial expression. As our experimental tasks are attention-drawing, a proportion of the 'anger' output could be interpreted as focus. Angry is a negative emotion, which could have the same effect as frustration, while focus simply gives information on attention and effort. The mixing of these two behaviors could lead to a non-significant result in this analysis. Future research should test other facial expression analysis algorithm which could identify the difference between anger and focus.

This study developed a multimodal protocol to measure affect in aviation. This mutltimodal methodology resulted in a validated, concurrent objective measurements of affect (facial expression, EDA) that addresses the shortcomings of traditional affect measurement in aviation (questionnaire, instructor evaluation). Facial expression and EDA analysis captures affective cues in aviation without disrupting training and overtasking instructors. Our findings could inform aviation instructors on how to infer trainees' workload and level of fatigue from biometrics measurement (e.g. high arousal inferred from EDA features suggests that trainees are having high workload and will experience higher fatigue). Such information could help instructors to design more effective training activities, reducing burnout and boredom. Furthermore, the current analysis attempts to generalize broad educational psychology theories such as the control-value theory to aviation training context. Despite some unexpected results, the current findings and the methodology could encourage aviation instructors to understand trainees' affective responses from the appraisal standpoint. Integration of appraisal theories of affect in aviation context could increase instructors' awareness of the motivational implication of trainees' affect and performance. Instructors could help trainees to reduce fatigue and boredom during training by balancing perceived control and challenge.

Predicting Aviation Performance with Affect Inferences

The previous section discussed our findings regarding data convergence across modalities of a multimodal affect assessment. Equally important in the field of aviation is how performance is influenced by affect. This section discusses the results obtained regarding the second research question: How do affective valence, arousal and workload during aviation training together and separately influence performance? As discussed in the aviation literature, affect could lead to detrimental results, such as fatal accidents. However, there is a lack of research on the role of affective states (except stress) in aviation training performance. In this study, this gap in affect research in aviation is addressed by examining the predictive value of affect in aviation training performance. Both physiological and behavioral data variables were included in the multiple regression analysis to account for the multiple expressions of affect. Phasic EDA is chosen as the overall indicator of physiological arousal. Overall valence from facial expressions is chosen as the indicator of affective valence. The choice of predictors also complies with two assumptions of multiple regression: (a) Multicollinearity: the independent variables cannot be highly associated with each other, and (b) power of the analysis is reduced if too many variables are included in the predictive model with a small degree of freedom.

No significant models were found for the baseline tasks with minimal difficulty. This might indicate that limited amount of mental workload and affective responses were experienced during easier tasks. Phasic EDA is a significant predictor of performance for the low difficulty task, positively predicting performance. This is consistent with past research as deactivating emotions such as boredom reduces attention, motivation and strategic efforts in problem solving (Noble, 2002; Pekrun et al., 2010; Pekrun et al., 2014). Furthermore, insufficient arousal could reduce the efficiency of psychomotor functions, resulting in poorer performance (Duffy, 1962). From the instructional standpoint, this result suggests that EDA could be a measure to identify disruptive emotional responses in training. Information on physiological arousal could help instructors to identify the potential causes for poor performance. For instance, as our results suggested, low performance in a low difficulty task could be due to lack of focus and boredom, instead of insufficient technical skills. Instructors could integrate arousal inferences from EDA to their decisions on instructional methods and training task design.

For medium to hard tasks, valence and phasic EDA do not significantly predict performance. The limitation of FaceReader in differentiating the expression of anger and deep focus could account for the non-significance of valence. Anger as a negative, activating emotion could hamper performance in difficult tasks as it reduces deep information processing and distracts attention from the task. However, in cases where FaceReader identified the expression of deep focus as anger, participants could achieve high performance as a result of focused attention and deep thinking (Pekrun & Perry, 2014). Furthermore, participants might have demonstrated a variety of cognitive and affective trajectories which are not accounted for by the current measures. For participants who are advanced in learning the experimental tasks, they could be fully engaged in strategical thinking and enter a state of flow, which is a positive, high arousal state (Hussain et al., 2011). In this case, arousal positively predicted performance. However, for those who were confused about how to interpret information from the simulation environment and did not know how to anticipate changes in parameters as they execute actions, high arousal might be an indicator of fear, confusion and stress (D'Mello & Graesser, 2012a). Coupling with negative valence, high arousal might negatively predict their performance. Although it is beyond the scope of the current thesis, future research could incorporate qualitative interviews in between experimental tasks to verify these interpretations obtained from quantitative biometrics measures. Furthermore, in the future we could explore the fit of dynamic predictive models such as dynamic Bayesian models to account for the differences in learning paths. Mental workload negatively predicted performance in medium to hard level tasks. This is consistent with previous theory and empirical research on mental workload. Cognitive overload is coupled with poor performance as it indicates insufficient cognitive resources for the current task (Itoh et al., 1990; Mansikka et al., 2018).

These findings support existing theories on arousal, valence and affective correlates and specifically demonstrate how affect influences aviation training. The design of the current analysis opens the discussion of using concurrent biometric measures to infer affect and workload and eventually predict performance. Our findings could inform aviation instructors on the causes of poor performance: in easy to medium tasks, boredom and low workload could lead to lack of attention and poor performance; whereas in harder tasks, poor performance could be a result of cognitive overload. In the next step of the bigger project on pilot training, we hope to replicate the current experiment design with pilot trainees as participants.

Differences in Discrete Emotions between High and Low Performance

As demonstrated in the previous section, affective states can be studied based on their positions on the valence and arousal dimensions. However, for arousal only has significant effects on performance in easy tasks and on workload in easy to medium level tasks. One way to explain the nonsignificant results in harder tasks is that some nuances of affect may not be fully expressed in terms of valence and arousal. This section further explores the relationship between affect and performance by discussing the results regarding the differences in discrete emotion expressed between high and low performance.

Discrete emotions carry additional complexities on appraisal of the situation (Barrett, 1998). Past research demonstrates that discrete emotions that arise in achievement activities and are closely associated with perfromance outcomes. Prospective emotions such as hope and anxiety are instigated by perceived control of the achievement activities where as retrospective emotions such as joy, anger and sadness are induced by perceived performance feedback (D’Mello & Graesser, 2012a; Dettmers et al., 2011; Pekrun & Linnenbrink-Garcia, 2014a). Furthermore, these emotions influence various cognitive processes that lead to performance consequences. For instance, enjoyment facilitates working memory and helps to focus attention on an achievement task, while anxiety has the opposite effect. In addition, discrete emotions are reactions or responses to feedback, indicating perception or appraisal of performance (Pekrun & Linnenbrink-Garcia, 2014a)

In our study chi-square test of independence is conducted to uncover potential associations between basic discrete emotions (happy, surprise, angry, sad, disgusted, scared) and performance. Close to significant differences are found: Angry occurred more in tasks with low performance than tasks with high performance. Happy was more frequently observed in tasks with high performance than in low performance. Same level of surprise was observed in both performance levels. These results are consistent with previous research on problem solving (D'Mello & Graesser, 2012): A state of cognitive disequilibrium and frustration is more likely to occur when an impasse is not efficiently resolved, which could be the case in low performance. In addition, this result could be interpreted together with the previous discussion on the correlation between anger and workload and appraisal. In the previous section (Data Convergence in a Multimodal Affect Assessment), one of the explanations for the correlation between the expression of anger and workload, perceived value and control is the similarity between the face of anger and the face of deep focus. Both deep focus and anger couple with cognitive disequilibrium, high workload and unresolved impasse, which could explain the association with low performance in the current task. However, future research on affective computing is needed to distinguish these two similar facial expressions, as they could lead to different learning outcomes in the long term: Deep focus, increased control and value could lead to adaptive achievement emotions and motivations and favorable learning outcomes in the long term (Pekrun et al., 2014), whereas anger and frustration could lead to disengagement from the task (D'Mello & Graesser, 2012). These findings add a perspective to the interpretation of low performance in a single task.

The effect of happy is consistent with past research as well: Once the impasse is resolved, cognitive equilibrium and a positive emotional state is restored (D'Mello & Graesser, 2012).

Therefore, it is within expectation that happiness is observed in high performance. In addition to co-occurring with cognitive states and performance outcomes, these discrete emotions could have an impact on performance by influencing motivational and cognitive processes. Positive activating emotions such as happiness could facilitate task performance by fostering deep cognitive processes, adaptive attention allocation and motivation towards the task (Pekrun & Linnenbrink-Garcia, 2014a). Hence, in this study, the feeling of happiness could have promoted task performance.

Limitations

This study is a pilot study for a large-scale project, which aims to examine a larger variety of potential biometrics assessments of affect and its correlates, such as workload, using experienced pilots and trainees as participants in a simulator with higher fidelity. The large scale project is still in progress and can not be reported on at this time. Although several of the hypotheses for this thesis were supported, it is acknowledged that there are limitations which could explain some unexpected results.

Firstly, one of the limitations is the facial expression recognition algorithm. It was noticed that angry is the most frequently experienced emotion for many of the participants. However, through retrospective interviews with the participants, qualitative information demonstrated that a lot of the participants did not experience anger related emotions. Some participants reported frustration and confusion at certain tasks. However, the majority of the participants reported being only slightly stressed and highly focused throughout. FaceReader developers have responded to my inquiry that there is high resemblance between a face of anger and a face of intense focus. The context of the experimental tasks in this study is highly complex and focus-eliciting. For some participants who struggle with certain maneuvers, the tasks might be anger-eliciting as well. In the current sample, there might be both cases. It would be meaningful to identify the difference between these two facial expressions in facial expression analysis. Although it is a limitation of the FaceReader algorithm, it is also a justification for using multimodal measurement for the current context. Combining and converging information gathered through different affect measurement channels not only captures various manifestations of affect, but also allows various data channels to complement each other on the information that current affective computing technologies cannot access individually.

Secondly, the analysis regarding the second research question used valence inferred from FaceReader, arousal inferred from EDA, and mental workload inferred from NASA TLX to predict performance. The number of variables included in the predictive model is limited by the multicollinearity assumption of multiple regression and the sample size. To avoid having highly correlated variables in the same model and to maximize the power of the analysis, the relationship between affect and performance is examined in both the two-dimensional perspective of affect and in terms of discrete emotions. In future studies, a larger sample size and less restricted predictive models could be explored. Machine learning algorithms could be implemented to further examine rich multimodal data as well. Future research should also explore the mediation effect of other constructs, such as goal structure. The frequently occurred emotions should be studied more systematically in aviation training context. For instance, specialized facial detection algorithm could be trained to identify differences between anger and focus in terms of facial expression.

Conclusion

In sum, this thesis presents a non-invasive, multimodal affect assessment protocol which could be potentially implemented in aviation training. This research contributes to the empirical methodology of aligning and converging multimodal data of affect. Using validated questionnaires, the validity of EDA and facial expression analysis as measurements of affect in aviation training context is supported. Our investigation on the association between affect (valence, arousal and discrete emotions) and performance addresses the gap in affect research in the aviation context. Through triangulation across data channels, we extend the generalizability of psychology theories on achievement emotions and appraisal to the aviation context. We provide empirical evidence on how these frameworks could be used to enhance our understanding of affect in the aviation context.

Our findings are consistent with past research in broader contexts: poor performance is associated with negative emotions (e.g. anger) and low activation. By establishing the connection between affect and performance in aviation training, this research confirms the importance of affect assessment and regulation to ameliorate training efficiency. Our results could inform instructors to allocate training resources efficiently by balancing trainees' perceived challenge and sense of control. Therefore, this research contributes to enhancing aviation training efficiency and ultimately improving flight safety

REFERENCES

Arnold, M. B. (1960). *Emotion and personality.* Columbia University Press.

Bakhtiar, A., Webster, E. A., & Hadwin, A. F. (2018). Regulation and socio-emotional interactions in a positive and a negative group climate. *Metacognition and Learning, 13*(1), 57-90. https://doi.org/10.1007/s11409-017-9178-x

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, *12*(4), 579-599. https://doi.org/10.1080/026999398379574

Bäuml, K. H., & Kuhbandner, C. (2007). Remembering can cause forgetting—but not in negative moods. *Psychological Science*, *18*(2), 111-115. https://doi.org/10.1111/j.1467-9280.2007.01857.x

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

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*(1), 1017-1034.

Burnham, J. F. (2006). Scopus database: a review. *Biomedical Digital Libraries*, *3*(1), 1-8. https://doi.org/10.1186/1742-5581-3-1

Causse, M., Dehais, F., Péran, P., Sabatini, U., & Pastor, J. (2013). The effects of emotion on pilot decision-making: A neuroergonomic approach to aviation safety. *Transportation Research Part C: Emerging Technologies, 33*, 272-281. https://doi.org/10.1016/j.trc.2012.04.005

Cohn, J. F. (2006). Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (pp. 233-238). https://doi.org/10.1145/1180995.1181043

Conati, C., Chabbal, R., & Maclaren, H. (2003). A study on using biometric sensors for monitoring user emotions in educational games. In *Workshop on Assessing and Adapting to User Attitudes and Affect: Why, When and How*.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life* (1st ed., Ser. Masterminds). BasicBooks.

Davis, H. A., DiStefano, C., & Schutz, P. A. (2008). Identifying patterns of appraising tests in first-year college students: Implications for anxiety and emotion regulation during test taking. *Journal of Educational Psychology, 100*(4), 942. https://doi.org/10.1037/a0013096

D'Arcey, J. T. (2013). *Assessing the validity of FaceReader using facial EMG.* [Master's thesis, California State University]. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.916.6560&rep=rep1&type=pdf

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

Den Uyl, M. J., & Van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. In *Proceedings of Measuring Behavior* (Vol. 30, No. 2, pp. 589-590). Wageningen.

Dettmers, S., Trautwein, U., Lüdtke, O., Goetz, T., Frenzel, A. C., & Pekrun, R. (2011). Students' emotions during homework in mathematics: Testing a theoretical model of antecedents and achievement outcomes. *Contemporary Educational Psychology*, *36*(1), 25-35. https://doi.org/10.1016/j.cedpsych.2010.10.001

Dewaele, J.-M. (2011). Reflections on the emotional and psychological aspects of foreign language learning and use. *Anglistik: International Journal of English Studies, 22*(1), 23-42.

Diaz-Piedra, C., Rieiro, H., Cherino, A., Fuentes, L. J., Catena, A., & Di Stasi, L. L. (2019). The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied Ergonomics, 77*, 92-99. https://doi.org/10.1016/j.apergo.2019.01.012

Doorey, A., Denenberg, B., Sagar, V., Hanna, T., Newman, J., & Stone, P. H. (2011). Comparison of myocardial ischemia during intense mental stress using flight simulation in airline pilots with coronary artery disease to that produced with conventional mental and treadmill exercise stress testing. *The American Journal of Cardiology*, *108*(5), 651-657. https://doi.org/10.1016/j.amjcard.2011.04.010

Duffy, E. (1957). The psychological significance of the concept of" arousal" or" activation.". *Psychological Review, 64*(5), 265. https://doi.org/10.1037/h0048837

Duffy, E. (1962). *Activation and behavior.* Wiley. https://psycnet.apa.org/record/1963-04290-000

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

Ekman, R. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64. https://doi.org/10.1518/001872095779049543

Flin, R., Martin, L., Goeters, K.-M., Hormann, H., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety, 3,* 97-120.

Freina, L., & Ott, M. (2015). A literature review on immersive virtual reality in education: State of the art and perspectives. In *The International Scientific Conference Elearning and Software for Education: Vol. 1*. (pp. 10-17).

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39*(3), 281-291. https://doi.org/10.1017/S0048577201393198

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

Hannula, M., Huttunen, K., Koskelo, J., Laitinen, T., & Leino, T. (2008). Comparison between artificial neural network and multilinear regression models in an evaluation of cognitive workload in a flight simulator. *Computers in Biology and Medicine, 38*(11-12), 1163-1170. https://doi.org/10.1016/j.compbiomed.2008.09.007

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.). (2015). *Emotions, technology, design, and learning*. Elsevier. https://doi.org/10.1016/B978-0-12-801856-9.00005-0

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

Hart, S. G. (2006). NASA-task load index (NASA-TLX): 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*(9), 904-908. Sage. https://doi.org/10.1177/154193120605000909

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47-77. https://doi.org/10.3102/00346543058001047

Hormeño-Holgado, A. J., & Clemente-Suárez, V. J. (2019). Effect of different combat jet manoeuvres in the psychophysiological response of professional pilots. *Physiology & Behavior, 208*. https://doi.org/10.1016/j.physbeh.2019.112559

Hussain, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011). Affect detection from multichannel physiology during learning sessions with AutoTutor. In G. Biswas, S. Bull, J. Kay, A. Mitrovic (Eds) *International Conference on Artificial Intelligence in Education* (pp. 131-138). Springer. https://doi.org/10.1007/978-3-642-21869-9_19

Itoh, Y., Hayashi, Y., Tsukui, I., & Saito, S. (1990). The ergonomic evaluation of eye movement and mental workload in aircraft pilots. *Ergonomics*, *33*(6), 719-732. https://doi.org/10.1080/00140139008927181

Jarrell, A. (2015). *The emotional twists and turns of problem solving :An examination of learners' behavioral, phsyiological and experiential emotion responses to unexpected events during problem solving* (unpublished dissertation). McGill University Libraries.

Jensen, A. R. (1965). Scoring the Stroop test. *Acta Psychologica, 24*(5), 398-408. https://arthurjensen.net/wp-content/uploads/2014/06/Scoring-the-Stroop-Test-1965-by-Arthur-Robert-Jensen.pdf

Jensen, R. S. (1997). The boundaries of aviation psychology, human factors, aeronautical decision making, situation awareness, and crew resource management. *The International Journal of Aviation Psychology*, *7*(4), 259-267.

Jensen, R. S. (2017). *Pilot judgment and crew resource management.* Routledge.

Jones, C. M., & Troen, T. (2007). Biometric valence and arousal recognition. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces* (pp. 191-194). https://doi.org/10.1145/1324892.1324929

Kaempf, G. L., & Klein, G. (2017). 11 Aeronautical Decision Making: The next generation. In N. Johnston, N. McDonald, & R. Fuller (Eds.). *Aviation psychology in practice*. 223-252. Routledge.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615. https://doi.org/10.1037/0278-7393.33.3.615

Knecht, C. P., Muehlethaler, C. M., & Elfering, A. (2016). Nontechnical skills training in air traffic management including computer-based simulation methods: From scientific analyses to prototype training. *Aviation Psychology and Applied Human Factors, 6*(2), 91–100. https://doi.org/10.1027/2192-0923/a000103.

Kochan, J. A., United States. Office of Aviation, M., & Civil Aeromedical, I. (1997). *A new approach to aeronautical decision-making : the expertise method*. U.S. Dept. of Transportation, Federal Aviation Administration, Office of Aviation Medicine. http://purl.access.gpo.gov/GPO/LPS84883

Kuroda, I., Fujiwara, O., Okamura, N., & Utsuki, N. (1976). Method for determining pilot stress through analysis of voice communication. *Aviation, Space, and Environmental Medicine*. *47*(5), 528–533. https://psycnet.apa.org/record/1980-24447-001

Lajoie, S. P., Zheng, J., Li, S., Jarrell, A., & Gube, M. (2019). Examining the interplay of affect and self regulation in the context of clinical reasoning. *Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2019.101219

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, 46(4), 352-367. https://doi.org/10.1037/0003-066X.46.4.352

Lee, A. T. (2017). *Flight simulation: Virtual environments in aviation*. Routledge.

Leininger, K. (2000). Interindexer consistency in PsycINFO. *Journal of Librarianship and Information Science, 32*(1), 4-8. https://doi.org/10.1177/096100060003200102

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology, 66,* 799-823. https://www.annualreviews.org/doi/full/10.1146/annurev-psych-010213-115043

Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics, 7*(4), 227-236. https://psycnet.apa.org/doi/10.1037/npe0000033

Li, S. Z. (2009). *Encyclopedia of Biometrics: I-Z* (Vol. 2). Springer Science & Business Media.

Main, L. C., Wolkow, A., & Chambers, T. P. (2017). Quantifying the physiological stress response to simulated maritime pilotage tasks: the influence of task complexity and pilot experience. *Journal of Occupational and Environmental Medicine*, 59(11), 1078-1083. https://doi.org/10.1097/JOM.0000000000001161

Mandler, G. (1989). Affect and learning: Causes and consequences of emotional interactions. In: McLeod D.B., Adams V.M. (eds) *Affect and Mathematical Problem Solving (pp. 3-19).* Springer. https://doi.org/10.1007/978-1-4612-3614-6_1

Mansikka, H., Virtanen, K., & Harris, D. (2018). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems*, *49*(1), 1-9. https://doi.org/10.1109/THMS.2018.2874186.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175-190. https://doi.org/10.1037/1528-3542.5.2.175

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*(2), 209-237. https://doi.org/10.1080/02699930802204677

Muis, K. R., Psaradellis, C., Lajoie, S. P., Di Leo, I., & Chevrier, M. (2015). The role of epistemic emotions in mathematics problem solving. *Contemporary Educational Psychology*, *42*, 172-185. https://doi.org/10.1016/j.cedpsych.2015.06.003

Murray, P. S., & Martin, W. L. (2012, November). Beyond situational awareness: A skill set analysis for situational control. In *Australian Aviation Psychology Association Symposium,* Sydney, Australia.

Niedenthal, P. M., Halberstadt, J. B., Margolin, J., & Innes-Ker, Å. H. (2000). Emotional state and the detection of change in facial expression of emotion. *European Journal of Social Psychology*, *30*(2), 211-222. https://doi.org/10.1002/(SICI)1099-0992(200003/04)30:2<211::AID-EJSP988>3.0.CO;2-3

Nittala, S. K., Elkin, C. P., Kiker, J. M., Meyer, R., Curro, J., Reiter, A. K., Xu, K. S., & Devabhaktuni, V. K. (2018). Pilot skill level and workload prediction for sliding-scale autonomy. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/ICMLA.2018.00188.

Noble, C. (2002). The relationship between fidelity and learning in aviation training and assessment. *Journal of Air Transportation World Wide*, *7,* 33-54. https://ntrs.nasa.gov/search.jsp?R=20020074981

Noldus FaceReader 6.0. (2015). https://student.hva.nl/binaries/content/assets/serviceplein-a-z-lemmas/media-creatie-en-informatie/media--communicatie/observatorium/factsheet-facial-coding-reference-manual.pdf?2900513938585

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press.

Panganiban, A. R., Matthews, G., Funke, G., & Knott, B. A. (2011b). Effects of anxiety on performance and workload in an air defense task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*(1), 909-913. https://doi.org/10.1177/1071181311551189

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*(3), 531.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36-48.

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology*, *106*(3), 696.

Pekrun, R., & Linnenbrink-Garcia, L. (2014a). *International handbook of emotions in education*. Routledge. http://ebookcentral.proquest.com/lib/mcgill/detail.action?docID=1656765

Pekrun, R., & Linnenbrink-Garcia, L. (2014b). Introduction to emotions in education. In R. Pekrun, & L. Linnenbrink-Garcia (Eds.) *International handbook of emotions in education* (pp. 11-20). Routledge.

Pekrun, R., Muis, K. R., Frenzel, A. C., & Götz, T. (2017). *Emotions at school*. 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. 130-151). Routledge.

Perry, R. P., Hladkyj, S., Pekrun, R. H., Clifton, R. A., & Chipperfield, J. G. (2005). Perceived academic control and failure in college students: A three-year study of scholastic attainment. *Research in Higher Education, 46*(5), 535-569. https://doi.org/10.1007/s11162-005-3364-4

Pintrich, P. R. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* https://eric.ed.gov/?id=ED338122

Regula, M., Socha, V., Kutilek, P., Socha, L., Hana, K., Hanakova, L., & Szabo, S. (2014). Study of heart rate as the main stress indicator in aircraft pilots. In *Proceedings of the 16th International Conference on Mechatronics-Mechatronika 2014* (pp. 639-643). IEEE. https://doi.org/10.1109/MECHATRONIKA.2014.7018334

Roscoe, A. H. (1978). Stress and workload in pilots. *Aviation, Space, and Environmental Medicine, 49,* 630–636. https://psycnet.apa.org/record/1980-26622-001

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., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502. https://doi.org/10.1037/0022-3514.57.3.493

Saproo, S., Shih, V., Jangraw, D. C., & Sajda, P. (2016). Neural mechanisms underlying catastrophic failure in human–machine interaction during aerial navigation. *Journal of Neural Engineering, 13*(6). https://iopscience.iop.org/article/10.1088/1741-2560/13/6/066005/meta

Sawin, D. A., & Scerbo, M. W. (1995). Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. *Human Factors, 37*(4), 752-765. https://doi.org/10.1518/001872095778995616

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 695-729. https://doi.org/10.1177/0539018405058216

Shiomi, K., Itano, K., & Suzuki, A. (2012). Development and evaluation of the fatigue and drowsiness predictor. *Archives of 28th International Congress of the Aeronautical Sciences.* http://www.icas.org/ICAS_ARCHIVE/ICAS2012/PAPERS/920.PDF

Soll, H., Proske, S., Hofinger, G., & Steinhardt, G. (2016). Decision-making tools for aeronautical teams: FOR-DEC and beyond. *Aviation Psychology and Applied Human Factors, 6*(2), 101–112 https://doi.org/10.1027/2192-0923/a000099

Starcke, K., & Brand, M. (2012). Decision making under stress: A selective review. *Neuroscience & Biobehavioral Reviews, 36*(4), 1228-1248. https://doi.org/10.1016/j.neubiorev.2012.02.003

Stokes, A. F., & Kite, K. (2017). *Flight stress: Stress, fatigue and performance in aviation*. Routledge.

Storbeck, J., & Clore, G. L. (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass, 2*(5), 1824-1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x

Terzis, V., Moridis, C. N., & Economides, A. A. (2010). Measuring instant emotions during a self-assessment test: The use of FaceReader. *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research. 18,* 1-4. https://doi.org/10.1145/1931344.1931362

Tole, J., Stephens, A. T., Harris, R., & Ephrath, A. (1982). Visual scanning behavior and mental workload in aircraft pilots. *Aviation, Space, and Environmental Medicine*. *53*(1), 54-61. https://psycnet.apa.org/record/1982-10909-001

van Erp, J. B., Eriksson, L., Levin, B., Carlander, O., Veltman, J., & Vos, W. K. (2007). Tactile cueing effects on performance in simulated aerial combat with high acceleration. *Aviation, Space, and Environmental Medicine, 78*(12), 1128-1134. https://doi.org/10.3357/ASEM.2133.2007

Warm, J. S., Matthews, G., & Finomore Jr, V. S. (2008). Vigilance, workload, and stress. Hancock, P. A., & Szalma, J. L. *Performance under stress*, 115-143. Ashgate. https://www.researchgate.net/profile/Gerald_Matthews2/publication/286031710_Vigilance_workload_and_stress/links/59288802a6fdcc4443552118/Vigilance-workload-and-stress.pdf.

Wickens, C. D., Lee, J. D., Liu, Y., & Becker, S. G. (2009). *An introduction to human factors engineering*. Pearson Education.

Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight [Article]. *Human Factors*, *44*(2), 171-188. https://doi.org/10.1518/0018720024497943

Wigfield, A., & Eccles, J. S. (2000a). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81. https://doi.org/10.1006/ceps.1999.1015

Zeidner, M. (1998). *Test anxiety: The state of the art*. Springer Science & Business

APPENDICES

Appendix i: Demographic Questionnaire

1. How experienced are you at similar tasks? (rate 1 - 5)
   Flying aircrafts

   Driving

   Driving games

   Flying games

   Other video games with joysticks

   Other: (please specify)

2. Gender:
   Female, male, other

3. Age:
4. Ethnicity:
5. ___ year in ___ (Master, PhD, undergraduate or pilot training program) program OR profession

Appendix ii: Control-Value Questionnaire

Control

Adapted from Wigfield & Eccles (2000)

Rate from 1 – 5 (strongly disagree to strongly agree)

- I have a great deal of control over this flying task I just did.
- I feel like the more effort I put in, the better I did at this flying task
- No matter what how hard I tried, I could not have done this task better

Value

Rate from 1 – 5 (strongly disagree to strongly agree)

Adapted from MSLQ (Pintrich, 1991)

Utility value

- I think I can use what I learnt from doing this flying task to other things in other situations
- Comparing to other things you are learning, how useful is this task

Interest

- I had fun doing this task.
- I find the task interesting.

Importance

- Comparing to other things you learn, how important is it for you to learn this flying task?
- how important is being good at this task for you?

Appendix iii: NASA TLX

The following questions are adapted from NASA Task Load Index

Mental load

Description: the amount of thinking, deciding, remembering, looking one needs to do; how challenging or complex is the task.

- How mentally demanding was the previous flying task? Rate 1-5

Physical load

Description: how much physical activity is required for the task (controlling, moving etc.); how physically challenging or laborious the task is.

How physically demanding was the previous flying task? Rate 1-5

Effort

Description: how hard you have worked (mentally and physically) to accomplish your level of performance in the task

How hard did you have to work to complete this flying task? Rate 1- 5

Fatigue

How tired were you (physically and mentally) doing the previous flying task? Rate 1-5

Appendix iv: Test Rubrics (Beginner Level)

Prepared by Hugh Grenier and Alain Bourgon, CAE Inc.

Grading Scale:

The following assessment guidelines are based on AQP scoring on a scale of 1 to 4. The emphasis of this study is to measure the “flying” performance. Threat and Error Management of CRM will not be assessed in this case.

Assessment of the resulting performance parameters applicable to the fields of **Heading**, **Altitude**, **Airspeed** and **VSI**:

Score	Definition	Description	Comments
4	Excellent	Achieves and maintains target parameters accurately. Very few minor deviations.	Exemplary.
3	Standard	Small deviations from target parameters occur but are corrected rapidly. No large deviation.	Safe and effective.
2	Acceptable	Frequent and / or large deviations from target parameters. Attempts are still made to return to target parameters.	Margin of safety is reduced.
1	Unsatisfactory	Large errors occur and are not corrected in a timely fashion. Target parameters are not achieved.	Safety of flight is jeopardized.

Control	Description	Score			
		1	**2**	**3**	**4**
Heading	Variations in degrees of direction from the target heading.	>30	≤30	≤20	≤10
Altitude	Variations in feet from the target altitude.	>1000	≤1000	≤500	≤200

Airspeed	Auto-thrust (Speed mode) will be used during experiments with ab-initio candidates on the Marinvent simulator.	N/A			
VSI	Variations in feet per minute from the target VSI.	>1500	≤1500	≤1000	≤300

Appendix v: Experimental Tasks and Instruction

Sequence	Difficulty	Task details
1 Survey 1	1	Maintain baseline for 30 seconds: altitude 10.0, speed 247, heading 001.5
2	2	Maintain airspeed and altitude, turn L at 15 degree bank angle to the heading of 330
3 Survey2	2	Reverse back to 001.5 heading, maintain altitude (10.0), speed (247)
4	3	Ascend to 11.0 by doing a 3 degree pitch up, once stabilized, turn L at 30 degree bank angle to heading of 330
5 Survey3	3	Reverse back to 001.5 heading by 30 degree bank to the right, stabilize, then descend to 10.0 altitude
6	3/4	Maintain bank and altitude, increase airspeed to 270 and stabilize
7 Survey4	3/4	Change speed back to 247
8	5	Ascend to an altitude of 11.0, stabilize, adjust airspeed to 230, stabilize. Then turn L to heading 330 with 30 degree bank angle, stabilize.
9	5	Descend to 10.0 feet, stabilize, adjust speed to 247, stabilize, turn R to heading 001.5 with bank angle 30 and stabilize.
Survey5		
10	1	Maintain altitude, bank and speed, fly for 60 seconds
Survey6		

Appendix vi: Consent Form

CONSENT FORM FOR PARTICIPATION IN THE EXPERIMENT ON PILOT TRAINING

Study Title: Inferring pilot trainee's affective and cognitive states from biometric data during the training processes

Researcher: Tianshu Li, Susanne Lajoie

Researcher's Contact Information: tianshu.li@mail.mcgill.ca

Faculty Supervisor: Susanne Lajoie

Faculty Supervisor's Contact Information: susanne.lajoie@mcgill.ca

You are being invited to participate in the research study mentioned above. This form provides information about what participating would mean. Please read it carefully before deciding if you want to participate or not. If there is anything you do not understand, or if you want more information, please ask the researcher.

A. PURPOSE

The purpose of the research is to develop biometric methods to infer pilot trainee's affective and cognitive states during the training process.

B. PROCEDURES

Before and after the experiment starts, you may be asked to fill in a questionnaire related to the training task. We may interview you for more information. The interview will be video-taped.

The experiments will be divided into two kinds: one is in a controlled lab, where some basic and general (may not be necessarily flight related) training activities will be conducted while your physiological and neurological signals will be recorded; the other will be conducted in a flight simulator located at Concordia or at CAE, where a session of flight training will be conducted while your physiological and neurological signals will be recorded. The training will follow a protocol provided by CAE. You may be asked to participate one of the two or both experiments, depends on the availability of the devices and the instructors.

In the beginning of the experiment, you will be asked to rest with eyes closed. Then, you will be starting the training specified above. Before the experiment ends, you will be asked to rest again with eyes closed. The following data will be recorded during the experiment:

- Electroencephalogram (EEG): electrical activity of the brain.
- Electrocardiogram (ECG): electrical activity of the heart. From ECG, heart rate and heart rate variability can be inferred.
- Eye tracking data such as eye movement, blinking frequency and pupil diameter.
- Galvanic skin response: change in electrical conductance of the skin.
- Respiration rate: the number of breaths per minute.

- Video and audio data: including body movements, gestures and facial expressions.
- Your learning behavior will be recorded in the training devices that you will use (a computer or a simulator).

C. RISKS

There is no risk participating in this experiment.

D. CONFIDENTIALITY

By participating, you agree to let the researchers capture your EEG, ECG, eye tracking data, galvanic skin response, respiration rate, video and audio data. You also agree to fill in the questionnaire, and conduct the pre- and post-experiment interviews. We will not allow anyone to access the information, except the researchers directly involved in conducting the research, and except as described in this form. We will only use the information for the purposes of the research described in this form. The information gathered will be coded. That means that the information will be identified by a code. The researcher will have a list that links the code to your name.

We intend to publish the results of the research. However, it will not be possible to identify you in the published results.

E. CONDITIONS OF PARTICIPATION

In order to be eligible for this experiment, ***you must not have any pre-existing mental disorder or must not be under prescribed medication for mental or psychological problems***.

If you are eligible and do participate, you can stop at any time. You can also ask that the information you provided not be used, and your choice will be respected. If you decide that you don’t want us to use your information, you must tell the researcher within TWO WEEKS after the experiment.

There are no negative consequences for not participating, stopping in the middle, or asking us not to use your information.

F. PARTICIPANT’S DECLARATION

I have read and understood this form. I have had the chance to ask questions and any questions have been answered. I ☐ **agree/☐ do not agree** to participate in this research under the conditions described.

NAME (please print)____
SIGNATURE____
DATE____

If you have questions about the scientific or scholarly aspects of this research, please contact the researcher. Their contact information is on page 1. You may also contact their faculty supervisor. If you have concerns about ethical issues in this research, please contact the Manager, Research Ethics, Concordia University, 514.848.2424 ex. 7481 or oor.ethics@concordia.ca.

Appendix vii: Briefing Script

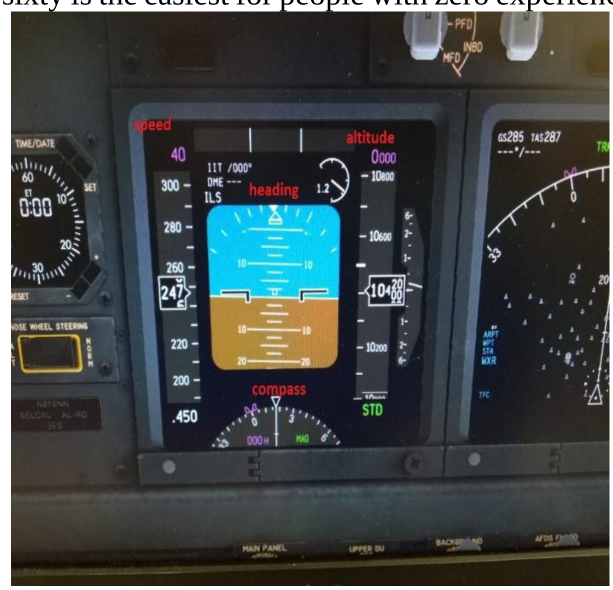
Overview

- 'x-plane is a flight simulation software/game. This research project aims to develop an emotion assessment method for pilot trainees during simulated flight trainings. This method will help us to understand the role of affect in aviation training, hence to develop more effective and multi-dimensional aviation training'
- 'you are going to use only the joystick to control three meters on the instrument screen: altitude, speed and direction
- Altitude is controled by pitch angle (ascend and descend)
- Speed is controlled by pitch angle and power
- Direction is controlled by heading

Joystick Introduction

- Thrust: the handle on the bottom of the joystick
- Trim: a button around the top of the joystick, you could gently massage it almost. We can play with trim to make different levels of difficulty as well, since a very well trimed plane would need little effort and control of the joystick to maintain baseline
- Baseline: put the two triangles together, one or two degrees above the horizon

The bank heading: 060, could be read at zero six zero, or sixty, or six. Whichever is understandable for the participant. Maybe sixty is the easiest for people with zero experience.



Terms and Instruction Language

- Thrust: the level of energy used by the plane
- Trim: establish baseline and achieve a like-autopilot condition (if let loose of the joystick, the plane will stay in baseline)
- Baseline: maintain altitude, speed and direction
- Bank: direction
- Initiate what degree bank turn, to which heading
- Roll in roll out: initiate a turn and stop a turn