Adaptive virtual reality stress training for spaceflight emergency procedures

by

Tor Teske Finseth

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Aerospace Engineering; Human Computer Interaction

Program of Study Committee:
Michael Dorneich, Co-major Professor
Nir Keren, Co-major Professor
Clayton C. Anderson
Warren Franke
Elizabeth Shirtcliff
Peng Wei

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

DEDICATION

To my parents, who read to me every night when I fell behind in the second grade.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I would like to thank my co-major professors, Michael Dorneich and Nir Keren for their unconditional guidance and support. They have been extraordinary mentors over the years. They have always expended enormous amounts of effort toward my research, treated me like a partner, and listened to my arguments about statistical methods. I am truly grateful.

I would also like to thank my committee members, Clayton C. Anderson, Warren Franke, Elizabeth (Birdie) Shirtcliff, and Peng Wei who throughout the course of this research were relentless advocates for my personal growth. They have spent endless hours teaching me new skills, discussing new and exciting ideas, telling me their astronaut stories, and going for goat walks.

In addition, I would also like to thank my friends and family. Being a graduate student can be tough at times, but having good company to work beside you in the cubicle, at a coffee shop, or at the kitchen table, makes the work more enjoyable.

ABSTRACT

Emergency training is an essential tool to mitigate safety risks to vehicles, operators, and for mission success. NASA astronauts go through extensive training to prepare for such situations. Astronauts can experience acute stress during hazardous, potentially life-threatening, situations that may erode any prior training and diminish remedial performance. Even high levels of skill training can succumb to the stress associated with the existential threat from an emergency. Incorporating stress training into the emergency training process may prepare astronauts to respond more favorably to stressful events. However, the implementation of stress training is difficult due to resource limitations, wide-variation between individual's stress responses, optimizing training to match user competency levels, and fidelity of the training environment.

The research objective is to develop and test an adaptive virtual reality (VR) stress training system as a countermeasure strategy against acute stress from spaceflight emergency operations. An adaptive VR training system may help astronauts develop resilience in preparation for high-stress operations. Four studies investigated the components and overall evaluation of the adaptive VR stress training system. The first study evaluated the effect of gradual exposure to stressors on building stress resilience. Participants were tasked with locating a fire on a virtual International Space Station (VR-ISS). Physiological and psychological measures were taken and results showed that prior exposure, as would be experienced during a gradual exposure to stress, enhanced relaxation behavior when confronted with a subsequent stressful condition. The second study developed and evaluated an emergency procedure, then manipulated a VR-ISS environment with three levels of stressors to induce psychological stress. The third study developed and tested a physiologically based stress detection system that uses

personalized interval methods to classify stress levels during tasks of ever-higher complexity, including an emergency fire procedure on the VR-ISS. A classifier was developed and tested against standard machine learning classifiers. Results from a human research study show high levels of accuracy in detecting multiple stress levels, even across tasks and when compared to other machine learning classifiers. The fourth study integrated the components from prior studies and evaluated a real-time adaptive stress training system. Using a VR simulation of a spaceflight emergency fire, predictions of the individual's stress levels were used to trigger adaptations of the environmental stressors (e.g., smoke, alarms, flashing lights), with the goal of maintaining an optimal level of stress during training. The adaptive training was compared to predetermined gradual increases in stressors (graduated), and trials with constant low-level stressors (skill-only). Results suggests that all training conditions lowered stress, but the adaptive condition was more successful decreasing multiple stress measures during the stress exposure. Lastly, the lessons learned from each of the studies was compiled into a list of recommendations to aid future researchers looking to improve training, stress detection, or adaptive systems.

CHAPTER 1. GENERAL INTRODUCTION

Objective

To contribute to the development of countermeasure support for space exploration class missions, this work describes the development and testing an adaptive virtual reality (VR) stress training system. This system aims to train healthy subjects to maintain performance under stress by methodically increasing stressor levels leading to increased resilience. Further, the system detects stress in real-time and autonomously changes the training environment to match the user's current capabilities. Supplementing current training practices with adaptive training that focuses on an individual's acute stress may prevent adverse behavior and performance degradation during actual emergencies.

Problem

Emergency training is an essential countermeasure tool to mitigate safety risks to vehicles and operators, while also increasing the probability of mission success. Astronauts perform tasks in environments with a multitude of hazards. When hazards develop into life-threatening emergency situations (e.g., fire, depressurization, toxic contaminate leaks; Marciacq & Bessone, 2009), astronauts can feel intense acute stress. This acute stress may erode any prior training, diminish performance, and jeopardize the lives of the crew. Therefore, how astronauts are trained for emergency situations plays an important role in survival. As Canadian astronaut Chris Hadfield explains, "If my focus ever wavers in the classroom or during an eight-hour simulation, I remind myself of one simple fact: space flight might kill me" (Hadfield, 2016, p. 320).

High levels of skill training can succumb to the stress associated with the existential threat of an emergency in the operational environment (Orasanu & Backer, 1996). *Skill training* is conducted under conditions that promote acquisition and retention of skills. When skills are

acquired, skill training often involves repeated drills to develop skill competency (Thompson & McCreary, 2006). This type of training is commonly used in astronaut and military training programs (Balmain & Fleming, 2009; Gancet, Chintamani, & Letier, 2012; Delahaij, Gaillard, & Soeters, 2006; Thompson & McCreary, 2006; Driskell, et al., 2008). Repeated drills have psychological benefits of automatizing the skill to instill a sense of control and minimize the demand on attentional capacity (Keinan & Friedland, 1996; Robson & Manacapilli, 2014). Astronauts train for emergency procedures by increasing the complexity of scenarios until they can reliably execute the procedures after launch (Balmain & Fleming, 2009). However, these training sessions require considerable physical and instructional resources. The training sessions may also use training practices that inadequately prepare individuals for coping with stressors or use training environments that lack the magnitude of stress felt in a real situation (Driskell et al., 2008). Inadequate training to handle specific stress in the operational environment can result in the diversion of cognitive resources for managing emotional states (e.g., fear, distress, and anxiety), leaving less resources available for task problem solving. Further, skill repetition does little to protect from novel stressors that can still degrade performance because of an inability to cope with the unfamiliar circumstance (Delahaij, Gaillard, & Soeters, 2006).

A NASA review of behavioral training methods for long duration spaceflight identified performance under stressful conditions and flexibility/adaptability as critical proficiencies needed by U.S. astronaut crews (Hysong, Galaza, & Holland, 2007). Nevertheless, despite the high level of skill training, NASA astronauts are only provided a short-duration classroom-based presentation on stress management without sufficient time to practice skills or receive feedback (Smith-Jentsch & Sierra, 2016).

Stress training could benefit astronauts preparing for a mission. Stress training is an effective method for preventing stress that supplements skill training by preparing individuals for performance in high stress operational environments (Driskell et al., 2008). Further, stress training can help individuals develop coping skills and flexibility to respond to unpredictable and uncontrollable stressors. Incorporating stress training into high-fidelity simulation or a NASA spaceflight analog may prepare astronauts to respond more favorably to emergent, but markedly stressful, events (Anglin et al., 2017).

Challenges

Training individuals to handle stress is difficult due to the wide variation on what is perceived as stressful. The perception of stress is unique to the environmental stressor and the individual's ability to cope with the stress. When an individual perceives a situation as stressful, stress will consistently influence physiology and human performance (Staal, 2004; Hockey, 1997). The general impact of stress on performance is the reduction of cognitive resources available for processing task relevant information as well as a reduction in cognitive regulation (Gillard, 2001; Eysenck et al., 2007). A deficiency in cognitive resources can affect many performance attributes including attention, decision time, and short-term memory (Staal, 2004). All these performance attributes may be critical for conducting an emergency procedure and avoiding harm. Developing strategies to train individuals for stress can reduce the impact on performance.

Several challenges exist with adopting stress training into emergency training environments. The first challenge is that stress training requires application and practice under conditions that approximate the operational environment (Driskell et al., 2008). However, scenarios that accurately portray the task and environment stress may not be suitable for the trainee's skill level. Subjecting unprepared trainees to extreme stress too early in training can

create a negative learning effect and learned helplessness (Keinan & Friedland, 1992; Driskell et al., 2008). Inversely, training that incorporates no stress can also be counterproductive because it may form mental models based on contextual factors in training that does not prepare the individual for the operational environment (Keinan & Friedland, 1992). Therefore, training with high-fidelity environments that allow the individual to become gradually more familiar with relevant stressors in a personal/unique/individualized manner will increase the trainee's capabilities in the field.

For decades, astronauts have used virtual reality environments (VRE) to practice for extra vehicular activity rehearsal, mass handling, and robotic arm manipulation (Garcia, Schlueter, Paddock, 2020; Homan & Gott, 1996; Cater & Huffman, 1995). The use of VREs has also been proposed to counteract stress from sensory deprivation and as psychological support systems for mental health during long-duration missions (Bachman, Otto, & Leveton, 2012; Salamon et al., 2018). However, the integration of VR with stress training has been limited by empirical research. As stated by Pallavicini et al. (2016), "Most of the studies, in fact, have VR-based stress management training programs only in theory, without providing data about trials conducted to test the effectiveness of the proposed approaches". Serino et al. (2014) found that challenges exist in the selection of effective technologies for delivering stress training, methodological rigor of the training pedagogy, and multi-dimensional assessment of stress. Therefore, more research is needed to verify the efficacy for VR stress training, especially before integration with NASA skill training.

The second challenge is that stress training methods have difficulty maintaining generalized training standards while simultaneously customizing the training to meet the trainee's specific needs and stress appraisal. The variation in stress response depends on the

experience, coping abilities, gender, age, and personality of the individual (Lu et. al., 2012; Sharma & Gedeon, 2012). Individual differences in physiological stress have also presented difficulties for building systems to detect and monitor stress (Giannakakis et al., 2019). Technologies to monitor stress need to be developed for NASA so that countermeasures can be introduced to support performance (Orasanu, Kraft, Tada, 2006).

The third challenge is that stress training can be vague and lack guidelines for choosing a context-specific pedagogy (Robson & Manacapilli, 2014; Crawford et al., 2013; Regehr et al., 2013). Several stress training strategies or techniques may be incorporated into high performance skills training (Meichenbaum, 2007; Driskell & Johnston, 1998). Stress training techniques may include cognitive control techniques, physiological control techniques, overlearning, mental practice, time-sharing skills, guided error training, decision-making training, flexibility training, and team training (Driskell et al., 2008). However, stress training is often conducted with the supervision of a trainer/psychologist with limited evidence on how much exposure to stressor is necessary (Robson & Manacapilli, 2014). Further, the training environment factors (e.g., trainee group size, number of training sessions, stressors, skills being trained) can influence the individual's competency and performance in future dynamic environments (Saunders et al., 1996). Therefore, stress training techniques should be selected to match the constraints of the trainee, task, and task stressors.

The last challenge is that astronaut training requires considerable resources and time. More research is needed to fully evaluate the direct benefits of stress training on task performance (Balmain & Fleming, 2009; Robson & Manacapilli, 2014). To emulate a stressful scenario, large requirements are placed on the facility and staffing capabilities. During the two-year training process for a six-month mission, NASA astronauts train approximately 32 hours for

emergency situations (Balmain & Fleming, 2009). Training requires rigorous travel scheduling and facility reservations with multiple instructors. Emergency exercises can include full-scale mock-ups of the International Space Station (ISS) modules and visiting vehicles (e.g., Soyuz capsule; Eichler, 2006; Marciacq & Bessone, 2009). While these training models are effective regarding fidelity, they are expensive and have difficulty simulating stressors (e.g., fire, depressurization). Further, over the course of a six-month increment on the ISS, it can be difficult for crewmembers to retain the skills learned on the ground, especially if emergency skills are not used very often (Balmain & Fleming, 2009). New training approaches are needed to mitigate safety risks, while also increasing the probability of mission success.

Approach

To contribute to the development of countermeasure support for exploratory class missions, this research describes the development and testing of an adaptive VR stress training system.

VR offers a safe and controlled environment for training resilience against traumatic or hazardous situations. Emergency operations can be trained for specific simulated tasks or stressors to strengthen transfer to the real world. Further, the advent of affordable VR technology has increased the mobile capability of head-mounted displays (HMDs). Stress training systems utilizing HMDs offer a unique opportunity for astronauts to train in immersive VR while at home, during space flights, or on the Martian surface.

Astronauts have the potential to encounter future stress; therefore, the aim of this research is preventative stress training with healthy people in task-oriented environments. Stress training can be accomplished by simultaneously practicing task skills while gradually increasing stress levels (i.e., graduated) over a series of sessions to promote control of the individual's threat appraisal. Graduated stress exposure is a component in two intervention therapies: Stress

Inoculation Training (SIT; Meichenbaum, 1985) and Stress Exposure Training (SET; Johnston & Cannon-Bowers, 1996). These interventions promote the training of coping skills and emotion regulation with graduated stress exposure until stress reaches the level expected in the real environment.

Training systems can be designed to be adaptive and manage crewmember acute stress levels during long duration missions. Sensing of the human state in real-time can enable the system to adapt a VR training environment based on the user's competency and momentary context. This can offer training advantages for crewmembers with different abilities to cope with stressful events. An adaptive system could ensure stress levels are maintained within a suitable physiological range assuring the crewmember is not overwhelmed. Adaptive VR stress training could promote competency and control, further enhancing the stress response and maintaining performance during life-threatening events.

The final product of the dissertation will be an adaptive system architecture as well as the evaluation of the system's ability to inoculate stress and enhance performance. Measures of physiological stress and graduated exposure can be joined with an adaptive system, which can assess the moment-to-moment stress of the user and adapt the VR spaceflight scenarios.

Research Questions

This research investigates how best to adapt user stress levels and improve task performance in the context of human spaceflight. Based upon the challenges of implementing stress training for emergency spaceflight hazards, investigation into the effects and adaptation of human stress will be explored through four linked studies. These will attempt to answer the following four research questions:

1. Can the combination of graduated stress exposure in an interactive 3D VR environment inoculate people against stress?

2. Can a spaceflight procedure simulated in VR be manipulated to evoke multiple stress levels?
3. Can multiple stress levels be detected and identified from physiological measures taken during simulated tasks with ever-higher levels of complexity?
4. Can a real-time physiology-driven VR adaptive system enhance resilience to stress without degrading performance?

Benefits and Contribution

Stress training has several practical applications for spaceflight training. First, graduated stress exposure offers unique advantages in comparison to traditional skill training; the latter can be effective under predictable conditions, but performance can degrade rapidly if novel stressors are introduced that reside outside the individual's coping ability (Driskell et. al., 2008; Driskell & Johnston, 1998; Keinan & Friedland, 1996). By introducing individuals to stressors through multiple sessions of graduated stress exposure, users should become familiar with the stress encountered in emergency situations while promoting competency and control of their stress response.

Second, there is only a medium correlation between subjective stress measures and the physiological stress responses, which may lower the sensitivity of subjective measures for rapidly changing emotional states (Campbell & Ehlert, 2012). A real-time physiological measurement may quickly assess the trainee's stress more reliably than observation or subjective measures (Smets, Raedt, & Van Hoof, 2019). Further, real-time physiological stress detection may have the capacity to account for individual differences in stress responses, making the detection more robust and accurate (Ćosić et al., 2010). Because of personalization, real-time stress measurement will recalibrate for crewmembers with physiological states that may change from day-to-day.

The third benefit is that an adaptive training system could adapt the training environment to optimize exposure to stressors and competency of the user. By considering stress-performance relationships and theories on learning, the adaptive training system can consistently adapt to keep the trainee in an optimal learning zone that provides a challenge without under- or over-whelming them (Parsons & Reinebold, 2012).

Finally, when ground support and training resources are limited, a mobile in-flight training system could be designed for long duration space flights. A mobile training system can minimize setup, implementation difficulty and require minimal trainer/psychologist support.

The contributions in this research are based on the integration of VR technology, adaptive systems, preventative stress training, and dealing with hazardous operations including spaceflight.

Document Organization

The remainder of this dissertation is organized as follows. Each research question corresponds to a study. The four studies mentioned above have been submitted as journal papers and reproduced here, but may contain work that is presented here for the first time. Chapter 2 introduces the literature on astronauts, stress, and training. Chapter 3 describes the research approach in greater depth. Chapter 4 contains Study 1, which found that graduated stress exposure in VR can reduce the stress response (Finseth et al., 2018). Chapter 5 contains Study 2, which describes the development of a VRE manipulated to evoke three levels of stress (Finseth et al., 2020). Chapter 6 contains Study 3, which tested the effectiveness of a physiologically-based stress detection system in identifying and classifying different stress levels using an Approximate Bayes algorithm (Finseth et al., 2021). Chapter 7 contains Study 4, which evaluated the effectiveness of the VR adaptive training system for enhancing resilience to stress

and maintaining performance (Finseth et al., 2021). Lessons learned and future work are discussed in Chapter 8. Conclusion and contributions are discussed in Chapter 9.

CHAPTER 2. LITERATURE REVIEW

This chapter presents a literature review on concepts and related research. Spaceflight is a dangerous occupation with many potential hazards. Some of the hazards are discussed as well as current training practices for spaceflight emergencies. Then the chapter reviews the human stress response, which can be elicited during encounters with potential hazards. To mitigate the stress from future hazardous tasks, the review then describes stress interventions and training, as well as stress training utilizing VR. Finally, the review provides background information on biofeedback, adaptive systems, and describes some adaptive VR systems that have been proposed for training individuals for stress.

Spaceflight Hazards and Training

Astronauts can experience a number of in-flight, life-threatening emergencies aboard the International Space Station (ISS). Although these emergencies are rare, several incidents have occurred in space operations (Evetts, 2009). In 1997, a *Vika* chemical oxygen generator malfunctioned aboard the *Mir* space station and caused a severe fire. Large amounts of toxic smoke filled the station for 45 minutes with near zero visibility (Linenger, 2000). In that same year, a Progress M-34 cargo vehicle collided with *Mir* causing decompression throughout the station and ultimately required the damaged *Spektr* module to be permanently sealed (Oberg, 1998). On the ISS, astronauts have also experienced several false fire alarms, including a false ammonium alarm in 2014 that resulted in the crew temporarily moving to the Russian side of the station (Hadfield, 2016; Kramer, 2015). Astronauts are responsible for reacting to remedy the situation or evacuate the station, both of which can be highly stressful in the presence of danger.

Threats to crew health have occurred during every phase of a mission (Evetts, 2009). Space operations entail long working days with a need for high levels of performance in critical

operations including extra-vehicular activity (EVA), docking crewed or cargo vehicles, or manipulation of a robotic arm close to the spacecraft (Orasanu, Kraft, & Tada, 2006). Further, advances in technology may do little to alter the perceived risk of living in a hostile environment characterized by microgravity, radiation, collisions with micrometeorites and supply vehicles, decompression, fires, and other environmental hazards (Palinkas, 2007). These hazard can be incredibly stressful and tiresome on crewmembers. Acute stress can accumulate in astronauts, resulting in psychological impairments including depression, anxiety disorders, and asthenia (Buckey, 2006). These psychological impairments increase the likelihood of errors that may have severe consequences, including damage to spacecraft and loss of life (Orasanu, Kraft, & Tada, 2006).

Astronauts and flight control teams train extensively for various emergency response strategies with classroom training and simulation (Uhlig et al., 2016). Spaceflight operations have mitigated risk by heavily relying on task-based training (Barshi & Dempsey, 2016). For this reason, task-based training has been the primary focus of research into spaceflight emergencies. For example, Olbrich et al. (2018) integrated VR with the European Space Agency's existent simulation framework and prototyped a lunar base emergency fire training simulation. Similarly, Uhlig et al. (2016) developed a virtual 3D simulator to train ground flight control personnel for ISS on-board fire emergencies and the corresponding response strategy. However, despite the heavy focus by researchers to support space agency operations, NASA's task-based crew training is not guided by empirically validated psychological principles and the historic spaceflight training for risk-mitigation does not support deep-space operations (Barshi & Dempsey, 2016). Training astronauts for future space missions must be qualitatively different than current practices.

Human Stress Response

Stress arises in transactional situations between an individual and the environment where the individual's perceived demands tax or exceed the perceived coping resources, which can result in physiological, psychological, behavioral, or social outcomes (Lazarus & Folkman, 1984). From this transactional perspective, stress is a coupled relationship between the person and the environment; stress is neither a characteristic of the individual or environment alone (Meichenbaum, 2007). External environmental stimuli are perceived as *stressors* and act upon the individual to evoke a *stress response*. The extent to which the situation is stressful is a function of the individual's perception. *Appraisal* is known as the process of simultaneously evaluating the demands and available resources to determine if the stressor is benign (i.e., positive appraisal), a challenge, or a threat (Folkman, 1984; Carpenter, 2016). The benign-positive appraisal results in no change to the physiological system. In contrast, a challenge appraisal refers to the stressor demands being taxing but within the individual's coping ability leading to the potential for growth. A threat appraisal refers to demands exceeding the individual's coping abilities, resulting in the undesirable state of being "stressed out" (McEwen, 2005).

The appraisal of a situation as a challenge or threat will have a large impact on the emotional and affective outcome. Challenge appraisal is usually associated with positive feelings, such as excitement and eagerness, whereas threat appraisal is associated with negative emotions, such as anger or fear (Folkman, 1984). In terms of adaptation, the challenge response is associated with increased energy mobilization for coping with the situational demands, whereas the threat response pattern serves to protect the individual from damage or harm (Olff, Langeland, & Gersons, 2005). Therefore, the stress response can be characterized as

physiological and psychological adaptation to meet the environmental demands for the purpose of achieving stability (McEwen, 2004).

A threat appraisal can lead to marked complications, both during an *acute* short-term duration or a *chronic* longer-term duration. Stress may result in:

- Physiological changes – increased heartbeat, labored breathing, trembling (Rachman, 1983), freezing behavior or tonic immobility (Abrams et al., 2009),
- Emotional reactions – fear, anxiety, frustration (Driskell and Salas, 1996),
- Cognitive effects – disruption in information processing (Gaillard, 2001), decreased search behavior (Streufert & Streufert, 1981), longer reaction time to peripheral cues (Wachtel, 1968), impaired or rigid decision making (Ellis, 2006; Starcke & Brand, 2012),
- Social effects – loss of team perspective (Driskell, Salas, & Johnson, 1999).

A stressful appraisal results in *coping*, which is the emotional, cognitive, and behavioral attempt to attenuate the stress response (Lazarus & Folkman, 1984). The effectiveness of the coping is not inherent in the selected coping strategy, but instead on the success to adapt to environmental demands. However, coping strategies are highly situational and some strategies may be more inflexible and consequential than other strategies. For example, while rumination, emotional numbing, escape, and intrusive thoughts may be beneficial for surviving hostile and inhospitable environments, these same coping strategies are associated with high levels of psychological distress and later can be ineffective in environments that require high levels of sustained attention during stress (McEwen, 2013; Ellis, Guidice, & Shirtcliff, 2013; Eysenck, 2007; Thompson et al., 2010). Further, consistent reliance on a single strategy, such as avoidance, can create maladaptive feedback resulting in chronic failure to alleviate stress, along with physical and mental health problems (Wadsworth, 2015).

Due to the severe consequences of failing to cope with stressors, increasing attention has been paid to building psychological *resilience* (Russo et al., 2012). Resilience is defined as an active, dynamic adaptation from successful coping with stressors that provides partial immunization against negative effects of futures stressors (Russo et al., 2012; Fletcher & Sarkar, 2013; Kalisch, Müller, & Tüscher, 2015). In other words, self-regulation during and after a stress response can shorten recovery periods until eventually a positive appraisal is achieved and normal levels of functioning are maintained (i.e., resilience; Fletcher & Sarkar, 2013). Some factors may contribute to positive adaptation (e.g., predispositions, genotypes), but they do so by facilitating intra-individual coping skills following an adverse life event or period of difficult life circumstances (Kalisch et al., 2017). Repeated and frequent use of copings skills to manage thoughts, feeling, and actions with focus on attaining personal goals builds the capacity to self-regulate in demanding environments (Ozhiganova, 2018). Therefore, individuals forced by new challenges to develop emotional regulation strategies increase the chances they will show optimized stress response in the same future situation (Kalisch, Müller, & Tüscher, 2015). However, this is dependent on the stressor(s) and perceived coping resources (e.g., intra-, inter-, extra-personal circumstances) during and after the exposure, and most importantly, if the individual has conscious intent for self-improvement (Kalisch et al., 2017).

Stress Interventions and Training

Interventions exist to help individuals learn about stress and develop coping strategies for current or future stress appraisals. Stress interventions can be classified into three models: primary, secondary, and tertiary (Lamontagne et al., 2007). These models allow for the distinction between the categories of *stress management* and *stress prevention*. First, primary interventions are the simplest solution because they aim to reduce exposure to stress by physically modifying the work environment or removing the stressor. When it is not possible to

remove a stressor, secondary interventions focus on stress prevention by providing education and skill development to improve participants' knowledge, skills, and ability to deal with stressful situations. Therefore, secondary interventions can be referred to as stress prevention which targets individuals deemed to be at risk during acute stress. Some secondary intervention frameworks include Stress Inoculation Training (SIT; Meichenbaum, 1985) and Stress Exposure Training (SET; Johnston & Cannon-Bowers, 1996). Lastly, tertiary stress interventions are most commonly utilized in clinical therapy for the treatment of people already experiencing strenuous circumstances or aliments. This form of care is referred to as stress management where individuals are already symptomatic due to past exposure to a stressor; therefore, the treatment is focused on care and support (Staal, 2004; Lamontagne et al., 2007). Commonly referenced intervention frameworks include Cognitive Behavioral Therapy, Mindfulness, Stress Resilience Training, Exposure Therapy, Toughness, and Stress Management Training.

Secondary interventions have three established requirements: (a) trainees should be given the opportunity to become familiar with stressors of the actual task situation; (b) such stressors should be introduced into the training process in a manner that prevents the build-up of anxiety and, (c) training should minimize interference with acquisition of skills required to perform the task (Friedland & Keinan, 1992). These requirements have been built into the two interventions frameworks of SIT and SET.

Stress inoculation training (SIT) can help individuals build resilience to future acute, sequential, and chronic stressors (Meichenbaum, 1985). The SIT approach is a three-phased flexible form of cognitive behavioral therapy. The initial phase of training (Phase 1) is conceptual education on the nature of stress and the stress effect. The second phase of SIT (Phase 2) involves acquisition of coping skills and consolidation of skills already possessed.

These coping skills are rehearsed prior to stressful training scenarios. The third phase of SIT (Phase 3), called application and follow-through, includes application of coping skills across multiple inoculation trials with increasingly demanding levels of stressors (Meichenbaum, 2007; Saunders, Driskell, Johnston, & Salas, 1996). As inoculation implies, developing coping strategies and incremental exposure over time decreases the potential for future negative cognitive, psychological, and behavioral reactions (Meichenbaum & Cameron, 1989; Serino et al., 2014; Leipold & Greve, 2009). Resilience can be achieved when the individual's appraisal promotes protective coping without experiencing mental health disruptions despite being subject to stressors (Fletcher & Sarkar, 2013; Kalisch et al., 2015). SIT has been validated for people with multiple sclerosis (Foley et al., 1987), injured athletes (Perna et al., 2003), and veterans suffering from anxiety, depression, or PTSD (Jackson et al., 2019; Hourani et al., 2011). In addition, SIT has been shown to reduce neuroendocrine stress response and lower stress appraisal to acute psychosocial stress in healthy subjects (Gaab et al., 2003).

Stress Exposure Training is a modified version of SIT, but for helping healthy individuals prevent stress in an operational environment where task performance is crucial (Johnston & Cannon-Bowers, 1996). SET is often referred interchangeably with SIT throughout literature because of their similar frameworks (Robson & Manacapilli, 2014). However, SET is commonly administered by trainers rather than clinical psychologists and used to enhance performance during complex cognitive tasks outside the original clinical domain (Thompson & McCreary, 2006; Johnston & Cannon-Bowers, 1996). Driskell et al. (2001) found that SET improved performance and reduced subjective stress in laboratory studies following exposure to novel stressors and tasks. Both SET and SIT have similarities through their third phase which practices skills under graduated simulated stress conditions.

Stress Training in Virtual Reality

VR offers a safe and controlled environment for exposure to traumatic or hazardous situations. Simultaneously, VR has the potential to solve the problem of treatment consistency and reconcile differences between the training environment and the environment in which the task is performed (Meichenbaum, 2007). In clinical exposure therapy, VR has become an effective tool for treating anxiety disorders and phobias (Meyerbröker & Emmelkamp, 2010), treating fear of flying (Rothbaum, Hodges, & Smith, 2000), treating claustrophobia (Malbos, Mestre, & Note, 2008), providing stress prevention for combat training in Iraq (Stetz, et al., 2007), and stress prevention for tactical and medical training (Wiederhold & Wiederhold, 2006).

A comparison of characteristics for SIT/SET experiments that utilized electronic treatment (e.g., video, phones, computers, virtual reality) is provided in Table 1, modified from original work by Serino et al. (2014), to include the participant sample, intervention purpose, SIT/SET phases, and third phase details. For the participant sample, Rose et al. (2013) makes the distinction that stress programs may have different effects on unhealthy and healthy stress samples. Unhealthy samples could include persons suffering from conditions like high anxiety or PTSD. Intervention purpose is provided to distinguish between stress prevention and stress management; stress management aims to develop recovery-focused coping strategies, whereas prevention-coping focuses on positive appraisal techniques to build resilience. SIT/SET phases describe which phases were included in the experimental protocol. For these studies, Phase 3 was defined as practicing coping skills in at least one session without a trainer. This could be through mental imagery, video, computer game, or VR. Lastly, VR was defined as an interactive virtual system, such as a head mounted display (HMD) or CAVE system. The table also provides characteristics for Phase 3, including the number of sessions if it was administered in VR, if the

sessions were graduate, and if task performance was measured during Phase 3. Research published after 2015 was included if SIT/SET was administered with VR.

The SIT experiment characteristics show that while many experiments were conducted with healthy subjects for the purpose of prevention against acute stress, specific intervention details differ greatly. Of the included studies, 16 out of 17 include the SIT Phase 3 of coping skill practice during stressful situations, only five of them provided graduated sessions, only 9 used VR, and only three measured task performance.

There are several methodological challenges surrounding the implementation of SIT/SET with VR and other electronic mediums. First, many studies were not included in Table 1 due to the lack of adhereing to SIT/SET guidelines, despite the claim of conducting stress inoculation. While SIT/SET has a flexible framework for personalizing treatment, few studies include all the phases and some omit key aspects about phases they were implementing (Pallavicini, Argenton, & Toniazzi, 2016). For example, negative training effects can occur when phase 2 task acquisition is conducted simultaneously during graduated exposure in Phase 3 (Friedland & Keinan, 1992). Second, while most of the studies implemented phase 3, only five implemented a form of graduated exposure, and no study explicitly mentioned graduated exposure. Several of the studies that used graduated exposure followed a 10-session preventative SIT framework by giving military participants presentations on stress education and coping, then conducting VR exposure sessions with breath training (Ilnicki, Wiederhold, & Maciolek, 2011; Kosinska et al., 2013; Maciolek et al., 2013; Zbyszeswki et al., 2012). In the short term, soldiers were able to reduce their arousal at the conclusion of 10 sessions (Kosinska et al., 2013). However, upon returning after a 19-month military deployment, no soldier showed long-term inoculation

Table 1: SIT/SET experiment characteristics. Adapted from Serino et al. (2014, pg. 77).

	Intervention Purpose		**Sample**		**Experiment details**		**SIT/SET**			**Phase 3 characteristics**			
	Stress Management	Stress Prevention	Unhealthy	Healthy	Sample	# of sessions	Phase 1	Phase 2	Phase 3	# of sessions	VR	Graduated Phase 3	Task Perform
Villani et al. (2013)	X		X		30 oncology nurses	8	X	X	X	2	-	-	-
Grassi et al. (2011)	X			X	75 university students	6	X	X	X	2	-	-	-
Riva et. al. (2007)	X			X	30 university students	6	X	X	X	2	-	-	-
Ilnicki et al. (2012)		X		X	118 soldiers	10	X	X	X	6	X	X	-
Maciolek et al. (2012)		X		X	118 soldiers	10	X	X	X	6	X	X	-
Zbyszeswki (2012)		X		X	120 soldiers	10	X	X	X	6	X	X	-
Kosinska et al. (2012)		X		X	4 soldiers	10	X	X	X	6	X	X	-
Hourani et al. (2011)		X	X	X	77 soldiers	2	X	X	X	1	-	-	X
Timmons et al. (1997)	X		X		68 veterans	12	X	X	X	2	-	-	-
Rose et al. (2013)		X		X	59 graduate students	6	X	X	-	N/A	N/A	N/A	-
Stetz et al. (2011)	X			X	60 soldier medics	5	X	X	X	1	-	-	-
Stetz et al. (2007)		X		X	25 soldier medics	2 (or 4)	X	X	X	2 (or 4)	X	-	-
Stetz et al. (2008)		X		X	63 student solider medics	2 (or 4)	X	X	X	2 (or 4)	X	-	-
Winslow et al. (2015)		X		X	40	1	-	-	X	5	-	X	X
Kluge et al. (2021)	X			X	30 university students	3	-	X	X	3	X	-	-
Prachyabrued et al. (2019)		X		X	60	1	-	-	X	1	X	-	-
Lugrin et al. (2016)		X		X	22 teachers	1	-	X	X	1	X	-	X

(Maciolek, 2013). While these results have potential, the small sample sizes (N=4 for autonomic system measurements; Kosinska, 2013) leave room for future work with larger samples

Third, SIT/SET techniques are stated to help maintain task performance during stress, but only three of the studies measured performance outcomes. Hourani et al. (2011) measured performance based on reaction time, although group statistical comparisons were not reported. Lugrin et al. (2016) measured teacher performance, including time to resolve, time to react, coping strategy, appropriate tone, and self-control in the classroom. No changes were found in the teachers' performance. Winslow et al. (2015) increased levels of stressors with five military scenarios that differed in task, novelty, predictability, and controllability. Task performance results did not differ between the experimental groups. These studies demonstrate that while SIT and SET are promoted as frameworks to improve performance under stress, more research is needed to validate the actual performance benefits.

Biofeedback

Biofeedback systems are an effective way to develop self-regulating mechanisms for improving health and performance by having instruments "feed back" information to the user (Schwartz & Andrasik, 2017). The biofeedback concept is synonymous with a closed-loop control of biological variables, with the direct aim at treatment of pathologies (i.e., train the patient so they can recover impaired functions to normal levels; Gaume et al., 2016). In a typical system, the user observes a feedback signal, which provides a form of explicit information to help the user self-regulate their biosignal (Gaume et al., 2016). The feedback information provided by the systems is usually directly proportional to the physiological signal, remaining constantly activated while the system is online.

While the practice of self-relaxation techniques with biofeedback may help manage current stress and eventually develop coping strategies, it assumes the body is in a current resting

state of arousal and must self-regulate to alleviate stress. Therefore, biofeedback is practiced with minimum stimuli rather than under conditions that approach stress in real-life situations. This may limit the transfer of self-regulation skills to acutely stressful real-life environments (Parnandi & Guiterrez-Osuna, 2015). Further, if self-regulation has been achieved, there is no further adversity for building resilience to situations beyond the users' past experience. For these reasons, biofeedback is recommended for phase 2 of SIT/SET to develop coping skills, rather than a method to apply coping within event-based scenarios for phase 3 of SIT/SET (Driskell et al., 2008).

To use biofeedback within event-based scenarios, several researchers have attempted to use biosignals to implicitly change some detail(s)/feature(s) of the simulated environment (Nacke et al., 2011; Parnandi, Son, & Gutierrez-Osuna, 2013). Implicit feedback is more subtle than biofeedback; the user is not directly aware of the biosignal change, but still experiences the indirect effects. For example, physiological arousal levels can be used to manipulate road visibility during a driving game or changing the field of view for police engaging hostiles (Parnandi & Guiterrez-Osuna, 2015; Brammer et al., 2021). The development of self-regulation through implicit feedback is theorized to rely more on autonomic regulation rather than volitional control (i.e., executive function and conscious cognitive strategies), consequently using less cognitive resources and attentional processes (Gaume et al., 2016). Therefore, the integration of implicit feedback with the event-based scenarios may subconsciously develop self-regulating strategies while simultaneously introducing adversity through simulations, and over time, build resilience.

A challenge that remains for biofeedback systems is that when a user has mastered self-regulation, higher levels of stress cannot be introduced because the displayed feedback signal (or

simulated feature) is directly proportional to the user's biosignal. In other words, the biofeedback system will always maintain a degree of authority even when users are in full control of their self-regulation. This may be problematic for users who want more authority over feedback information than the static system will allow and may lead to negative training or skill development (Byrne & Parasuraman, 1996). Further, the biofeedback system will continue to be constrained at a single level of operation that was fixed at the system design stage. This may limit the degree to which a simulated event-based scenario can be adjusted to reflect a real-world stressful situation. As explained in the SIT/SET phases, phase 3 is intended to introduce stress gradually and avoid overwhelming or sensitizing the user to stress (Driskell et al., 2008). Due to these limitations, rather than biofeedback, an autonomous system is better suited to vary the environment depending on the situational demands or reallocate control between the system and human.

Adaptive Physiological Systems

Adaptive automation is the process of dynamically allocating control of system functions between a human operator and/or computer over time (Kaber et al., 2005). The adaptive system is the technology component of that human-machine adaptive automation. Adaptive systems have primarily been characterized as adaptations of function allocation where taskwork is distributed between the human and system to maximize task performance (Parasuraman, Sheridan, & Wickens, 2000). In addition to dynamic function allocation, adaptive systems can modify information content, human-machine interaction style, and task scheduling (Feigh, Dorneich, & Hayes, 2012). The human-automation pair considers the agents' capability to assume authority for a function and individual responsibility to perform the collective set of functions (Pritchett, Kim, Feigh, 2014). However, adaptations can be used for other purposes

than taskwork function allocation including the ability help humans learn and control affective, physiological, or cognitive skills.

Rather than use physiological signals directly like biofeedback, some adaptive systems can monitor the human state, including metrics such as workload, fatigue, and stress. Specifically, much research has been devoted to the monitoring of stress detection in real-time (Smets et al., 2019). At a basic level, stress detection uses machine learning on data collected from physiological sensors to classify states of stress (Giannakakis et al., 2019). By classifying stress into different states, the system can adapt functionality given specific conditions; moreover, the system can decide when to engage with users and vary the amount of interaction and authority for a task. Therefore, the integration of stress detection with an adaptive system may be an effective way to build resilience and prevent stress (Zahabi & Razak, 2020).

Adaptive Virtual Reality Stress Training

A number of researchers (Ilnicki et al., 2011; Serino et al., 2014; Pallavicini, Argenton, & Toniazzi, 2016) have recognized the benefits of using adaptive systems to modify VR environments in order to conduct stress interventions. Further, some researchers have proposed system designs or tested components for a VR adaptive system that integrates stress prevention or stress management (Zahabi & Razak, 2020).

The Virtual Reality Adaptive Simulation (VRAS) was developed to prevent serious mental health problems prior to military deployment (Popović et al., 2009; Ćosić et al., 2010a; Ćosić et al., 2010b; Ćosić et al., 2011). The VRAS system uses concepts from SIT to gradually expose trainees to stressful stimuli with simultaneous practice of stress coping skills. Stimuli are generated through static pictures, sounds, and real-life video clips. The subjective emotional response of the trainees informs the adaptations and is measured through voice, speech, facial

expressions, heart rate, skin conductance, and respiration. This research was proof of concept and only tested the component that estimates emotional states of the user.

Another adaptive system for military application is the Virtual Reality for Cognitive Performance and Adaptive Treatment (VRCPAT 2.0; Parsons & Reinebold, 2012). The purpose of the system is to test a soldier's readiness to return-to-duty following a traumatic experience or head injury. This system uses the current cognitive state derived from psychophysiological signals and task performance, then adapts the VR environment (Rizzo et al., 2011). First, the system profiles the user's baseline stress level into two classifications (high, low) based on psychophysiological indices (heart rate, skin response, and pupil diameter). The user then engages in an HMD VR simulation driving a military Humvee. Performance is assessed by the ability to follow and maintain separation to a lead convoy Humvee. Several stimuli were used as adaptations, including the lead Humvee's speed, explosive blasts, insurgent AI characters, in-game weather conditions, and a scent machine that produces gunpowder odor. A classifier was developed for three stress levels using a VR psychological incongruence task to establish ground truth (Wu & Parsons, 2011). However, this research is proof of concept and classification accuracy and experimental results have yet to be published.

A system proposed by Jones and Dechmerowski (2016) leverages mobile technology and SIT framework to objectively measure stress levels during augmented reality (AR) and VR training. The system proposes to use unobtrusive physiological stress monitoring, including cardiovascular and respiratory measures and electrodermal activity. The system uses a linear stochastic gradient descent binomial classifier that differentiates stress during a public speaking task at 95% accuracy (Winslow et al., 2016). The researchers proposed that physiological measures with real-time performance could drive adaptive training, during which the repeated

exposure to a stressor would help build effective coping mechanisms. They propose that a closed-loop system can create an ever-changing environment, thus pushing users to create coping mechanisms for use in a live transfer environment. However, similar to the adaptive stress training system mentioned previously, this research is proof of concept and experimental results have yet to be published.

CHAPTER 3. RESEARCH APPROACH

To develop the VR adaptive system and evaluate its effectiveness, four inter-related studies investigated the four research questions, as illustrated in Figure 1. These studies iteratively develop and evaluate different components for manipulating, classifying, and adapting stress levels as well as the final closed loop adaptive system.

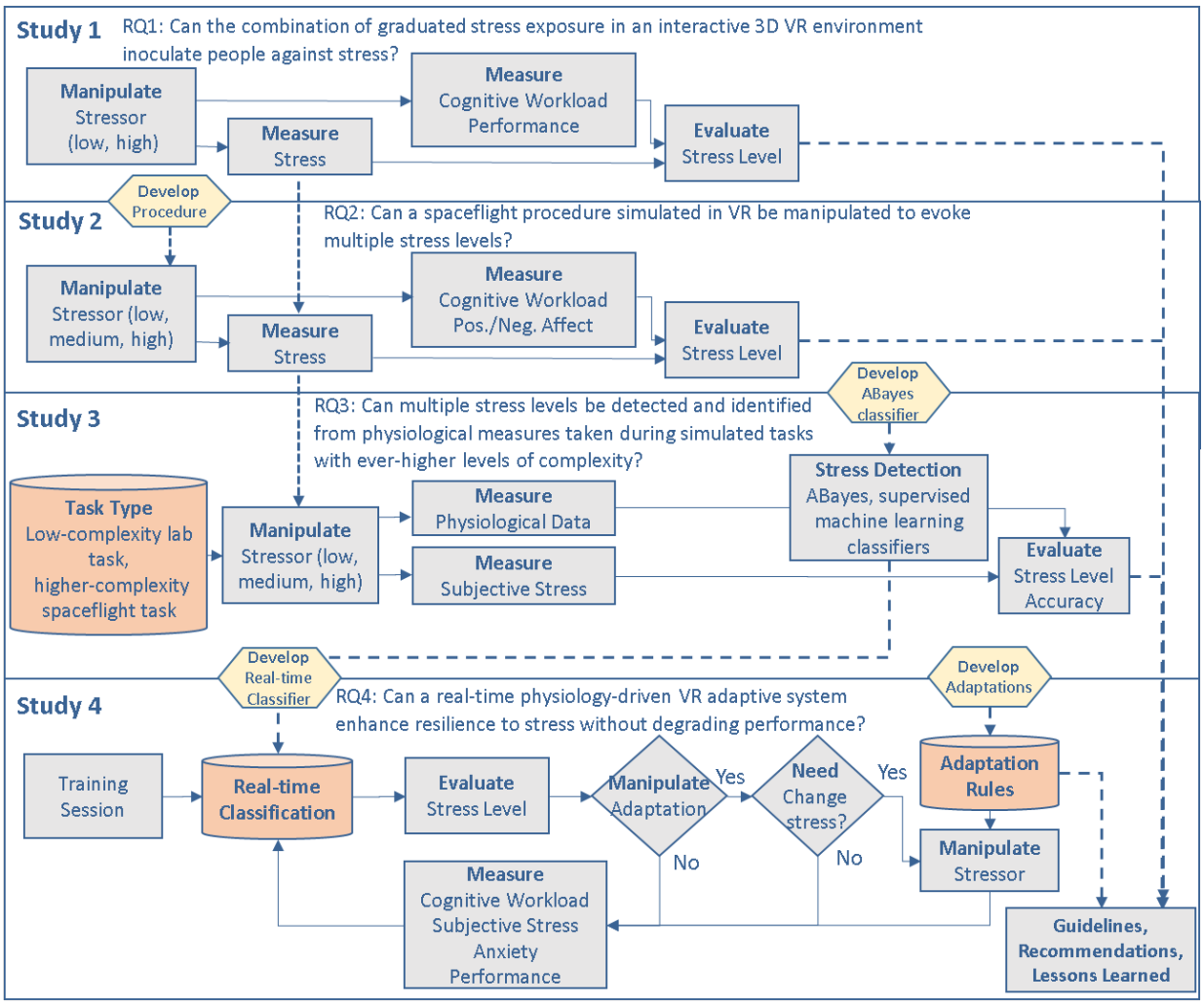

Figure 1: The overall vision of research.

Study 1 (Chapter 4) addressed the first research question of: *Can the combination of graduated stress exposure in an interactive 3D VR environment inoculate people against stress?* The experiment investigated how prior exposure to stress effected the physiological stress

response, subjective stress, time-to-completion, and cognitive workload during a subsequent increasingly stressful situation. The training condition was compared to a group that was not given prior stress exposure. This study's experimental design was based on past research with SET/SIT and graduated stress exposure (Meichenbaum, 2017; Saunders et al., 1996; Keinan & Friedland, 1996; Johnston & Cannon-Bowers, 1996). The training pedagogy of using graduated stressors with a simulated spaceflight procedure was considered when designing the subsequent studies.

Study 2 (Chapter 5) addressed the second research question of: *Can a spaceflight procedure simulated in VR be manipulated to evoke multiple stress levels?* This study addressed the second research question by designing a VR spaceflight environment with three-stressor levels. Further, an emergency fire task procedure was created with the help of subject matter experts in spaceflight procedures and human stress response. Conducting this study helped verify that VR training task accurately replicates the existing task procedure and has ability to induce different, distinguishable, levels of stress in trainees. Both these outcomes were necessary to ensure the relevant task skills were trained and that graduated stress exposure could have multiple levels.

Study 3 (Chapter 6) addressed the third research question of: *Can multiple stress levels be detected and identified from physiological measures taken during simulated tasks with ever-higher levels of complexity?* This study is based on past research on physiological stress classification (Plarre et al., 2011; Singh, Conjeti, & Banerjee, 2014). This experiment addressed the challenges of classifying individual stress responses, using machine learning to classifying time-series data that has temporal correlations, and classifiers that use indirect methods to approximate class conditional probabilities. The objective was to assess the extent to which an

individualized stress detection system can be effective in classifying three levels of stress during a low-complexity lab task and a higher-complexity VR emergency fire procedure. Physiological data was collected and input in a Bayes classifier, known as Approximate Bayes (ABayes), which was developed and evaluated for its accuracy in predicting stress. ABayes was compared to traditional machine learning classifiers and then implemented in a real-time stress detection framework.

Study 4 (Chapter 7) addressed the fourth research question of: *Can a real-time physiology-driven VR adaptive system enhance resilience to stress without degrading performance?* Studies 1, 2, and 3 evaluated critical components of the system in isolation, whereas study 4 evaluated the real-time adaptive stress training system with all components integrated. Using a VR simulation of a spaceflight emergency fire, predictions of the individual's stress levels were used to trigger adaptations of the environmental stressors (e.g., smoke, alarms, flashing lights), with the goal of maintaining an optimal level of stress during training. The adaptive training was compared to predetermined gradual increases in stressors (graduated), and trials with constant low-level stressors (skill-only).

The lessons learned from each of the studies was compiled into a list of recommendations to aid future researchers looking to improve training, stress detection, or adaptive systems (Chapter 8).

CHAPTER 4. EVALUATING THE EFFECTIVENESS OF GRADUATED STRESS EXPOSURE IN VIRTUAL SPACEFLIGHT HAZARD TRAINING

Tor Finseth[1], Nir Keren[2], Warren D. Franke[3], Michael C. Dorneich[4], Clayton C. Anderson[1] & Mack C. Shelley[5]

Modified from a manuscript published in *Journal of Cognitive Engineering and Decision Making* (Finseth et al., 2018).

Statement of Authorship

As the lead author on this paper, I developed the theoretical background, designed and conducted the experiment, and performed the analysis. Dr. Keren conceived the original idea. Dr. Keren, Dr. Franke, and Dr. Dorneich helped supervise the project and interpret the results. Clay Anderson contributed to the development of the tasks. Dr. Shelly provided statistical feedback and guidance. I wrote the manuscript with review and input from all authors.

Abstract

Psychological and physiological stress experienced by astronauts can pose risks to mission success. In clinical settings, gradually increasing stressors help patients develop resilience. It is unclear whether graduated stress exposure can affect responses to acute stressors during spaceflight. This study evaluated psychophysiological responses to potentially catastrophic spaceflight operation, with and without graduated stress exposure, using a virtual reality environment.

[1] Dept. of Aerospace Engineering, Iowa State University, Ames, IA, 50011, USA
[2] Dept. of Agricultural and Biosystems Engineering , Iowa State University, Ames, IA, 50011, USA
[3] Dept. of Kinesiology, Iowa State University, Ames, IA, 50011, USA
[4] Dept. of Industrial Manufacturing and Systems Engineering, Iowa State University, Ames, IA, 50011, USA
[5] Dept. of Political Science and Dept. of Statistics, Iowa State University, Ames, IA, 50011, USA

Twenty healthy participants were tasked with locating a fire on a virtual International Space Station (VR-ISS). After orientation, the Treatment group (n=10) practiced searching for a fire while exposed to a low-level stressor (light-smoke), while the Control group (n=10) practiced without smoke. In the testing session, both groups responded to a fire while the VR-ISS unexpectedly filled with heavy smoke. Heart rate variability and blood pressure were measured continuously. Subjective workload was evaluated with the NASA Task Load Index, stress with the Short Stress State Questionnaire, and stress exposure with time-to-complete.

During the heavy smoke condition, the Control group showed parasympathetic withdrawal, indicating a mild stress response. The Treatment group retained parasympathetic control. Thus, graduated stress exposure may enhance allostasis and relaxation behavior when confronted with a subsequent stressful condition.

Introduction

Several uncontrollable and unpredictable emergencies have occurred aboard the International Space Station (ISS) including fire, depressurization, and toxic environments (Summers et al., 2005). Astronauts are responsible for either remedying the situation or being prepared to evacuate the station. These situations can be highly stressful, since the consequences of not responding appropriately to the situation can be catastrophic. To prepare for emergency situations, NASA astronauts train approximately 40 hours using a full-scale ISS mock-up at NASA Johnson Space Center. Emergency training is a small, but crucial, piece of the overall training process necessitating rigorous scheduling and international travel over the course of two years. Given the sheer volume of operational spaceflight training that astronauts undergo, it is challenging to incorporate training interventions focused on remaining calm during potentially life-threatening situations.

Stress arises in transactional situations where the individual's perceived demands tax or exceed the perceived coping resources, which can result in negative physiological, psychological, behavioral, or social outcomes (Lazarus & Folkman, 1984). The individual's appraisal of a situation determines the extent to which the situation is stressful. When stress occurs, the process of allostasis is the body's attempt to adapt, maintain, or regain stable levels of functioning (McEwen, 2001; McEwen & Wingfield, 2003). With appropriate training, a healthy allostatic state can be maintained during exposure to intense acute stress. However, if allostasis is not maintained, outcomes can include a disruption in information processing (Gaillard, 2001), impaired or rigid decision making (Ellis, 2006; Starcke & Brand, 2012), and declines in cognitive performance (Lieberman et al., 2005), potentially leading to freezing behavior or tonic immobility (Abrams et al., 2009) and an impairment in performance during the crisis (Delahaij, Gaillard, & Soeters, 2006).

Most spaceflight training is performed intensively and frequently, but it is primarily focused on mastering tasks. Over time, repetitively practiced skills become automated, thereby requiring less attention and being more resistant to disruption (Driskell et al., 2008). However, stress can negatively affect performance even with high levels of task training (Orasanu & Backer, 1996). Overlearned skills may not lead to effective coping, except in situations where the stressors are well-known (Delahaij, Gaillard, & Soeters, 2006). When individuals are exposed to unpredictable situations, new stressors or radical environmental changes, a maladaptive response may occur. Learning how to respond to unpredictable acute stress could be helpful in spaceflight. While task training focuses on the automaticity of the task itself, stress training focuses on reduction of the stress response through coping. Developing coping strategies decreases the potential for negative cognitive, psychological, and behavioral reactions

(Meichenbaum & Cameron, 1989; Serino et al., 2014, Leipold & Greve, 2009). Moreover, during acute stress, coping mechanisms increase the likelihood of staying calm and relatively relaxed.

Stress inoculation training (SIT) could potentially help astronauts stay calm by building resilience to acute, sequential, and chronic stressors (Meichenbaum, 1985). The SIT approach is a three-phased flexible form of cognitive behavioral therapy. The initial phase of training is conceptual education focusing on the nature of stress and stress effects. The second phase involves acquisition of coping skills and consolidation of skills already possessed, where a variety of coping skills are rehearsed in preparation for stressful situations. The final phase of SIT can include application of these coping skills across multiple inoculation sessions with increasingly demanding levels of stressors (Meichenbaum, 2017; Saunders et al., 1996).

Resilience can be achieved when the individual's appraisal promotes protective coping without experiencing mental health disruptions, despite being subject to stressors (Fletcher & Sarkar, 2013; Kalisch, Müller, & Tüscher, 2015). With the goal of improving resilience, SIT is commonly used for the prevention and management of stress. Stress management differs from stress prevention in that the former focuses on reactive care and support after a stressful incident while the latter focuses on proactive measures to reduce the stress response (Staal, 2004). SIT has been validated as a stress management intervention for individuals in chronically stressful work settings (Foley et al., 1987; Perna et al., 2003), anxiety, depression, or post-traumatic stress disorder (Hourani et al., 2011). While SIT is often used in clinical therapy, more research is needed to explore its utility with healthy individuals experiencing acute stress (Rose et al., 2013).

One component of clinical SIT, graduated stress exposure, provides a mechanism for becoming comfortable with stress by providing trainees with a heightened sense of control and

competency (Keinan & Friedland, 1996). Graduated stress exposure could be readily integrated into astronaut training. Here, stressors could be introduced during task training with the stress levels gradually increasing during a training session, or over a series of sessions, to promote control of the individual's threat appraisal (Fornette et al., 2012; Johnston & Cannon-Bowers, 1996). A single session may be sufficient for stress response improvement (Baumann, Gohm, & Bonner, 2011), although multiple sessions may be better at fostering confidence in preparation for realistic stress levels.

In clinical therapy, the flexible framework of SIT allows the training to be modified to fit a patient's needs. However, additional research is needed on stress training for healthy people working in challenging environments, such as astronauts (Rose et al., 2013). Limited evidence exists to guide trainers in the selection of effective SIT intervention techniques (Crawford, Wallerstedt, & Khorsan, 2013; Regehr, Glancy, & Pitts, 2013); thus, research is needed to determine how the separate components of SIT contribute differentially to appraisal, biological arousal, and training effectiveness during occupational tasks (Saunders et al., 1996; Robson & Manacapilli, 2014).

Therefore, the purpose of this study is to assess the extent to which the individual contribution of graduated exposure to a task-specific stressor affects the physiological response to a critical spaceflight hazard. Specifically, we hypothesize that exposure to a light level of virtual smoke, during training focused on responding to a fire threat on the International Space Station, would improve the psychophysiological responses to a subsequent simulated emergency with heavy-smoke exposure, in comparison to a group that was trained without smoke exposure during the prior training. Collectively, this study serves as a proof-of-concept that this aspect of SIT may be worthy of consideration for including into future astronaut training. The novel

contribution of this paper is to provide insight into the effectiveness of graduated stress exposure as a component of SIT on modifying the response to stress, and whether the brief exposure to an acute spaceflight hazard, in healthy people, improves psychological and physiological responses to subsequent exposure.

Methods

General Experimental Design

Study participants came to the laboratory three times. The first visit was an orientation tour of the International Space Station in a virtual reality environment (VR-ISS). Here, participants learned the emergency response procedure to a fire, practiced navigating the VR-ISS, and completed a written quiz indicating their wayfinding and emergency response abilities. The second visit, the Training session, focused on completing a fire response protocol in the VR-ISS; half the participants were exposed to light smoke during this session while the other half were not. The third visit, the Testing session, was another fire response protocol; however, both groups were exposed to an unexpected and rapid accumulation of heavy smoke during the session. Participants were randomly assigned using a 1:1 ratio to either the Treatment (exposed to low level of smoke in prior session) or Control (no exposure to smoke in prior session) group for the two fire drill sessions. The location of the source of the smoke was the same for both participant groups (i.e., Treatment and Control), but varied between the Training and Testing sessions. The purpose of changing the source location was to prevent a learning effect between the Training and Testing session. The level of task difficulty in terms of required procedures and source location was kept constant across Sessions and Groups. Physiological responses and psychological states were assessed in both sessions. All study procedures were approved by the Iowa State University Institutional Review Board (see Appendix).

Participants

Potential participants were excluded if they reported having severe anxiety, claustrophobia, pregnancy, simulation sickness, seizures, heart abnormalities, circulatory problems, or implanted electromagnetic devices. Twenty-two subjects consented, but two withdrew prior to finishing the experiment. The final sample was 20 adult males, mean age was 22.5 years (*SD* = 2.2), from the Iowa State University community. None of the participants had prior experience with VR. The demographic included 60% Caucasian, 15% African American, 15% Hispanic or Latino, and 10% Asian or Asian American.

Task / Scenarios

All participants were asked to follow a simplified ISS emergency fire response procedure in the VR-ISS with the goal of locating the source of the smoke. The simulation followed the NASA ISS emergency fire procedures which contained instructions for crew responsibilities, air contaminants location sampling, and ISS system configuration (NASA, 2013). During the simulation, smoke was generated from a source in one of the modules aboard the ISS US Orbital Segment (see Figure 1). Participants began the simulation in the Node 1 module, since this is the "safe haven", closest to the Russian operations segment and the Soyuz escape capsule on the ISS (see U.S. Orbital Segment interior of the ISS in Figure 1).

The simulation took place in the C6, a virtual reality room at Iowa State University; Figure 2a illustrates participants in the VR-ISS in the C6. Figure 2b is an example of a view that participants saw in the simulation, including the location of a fireport.

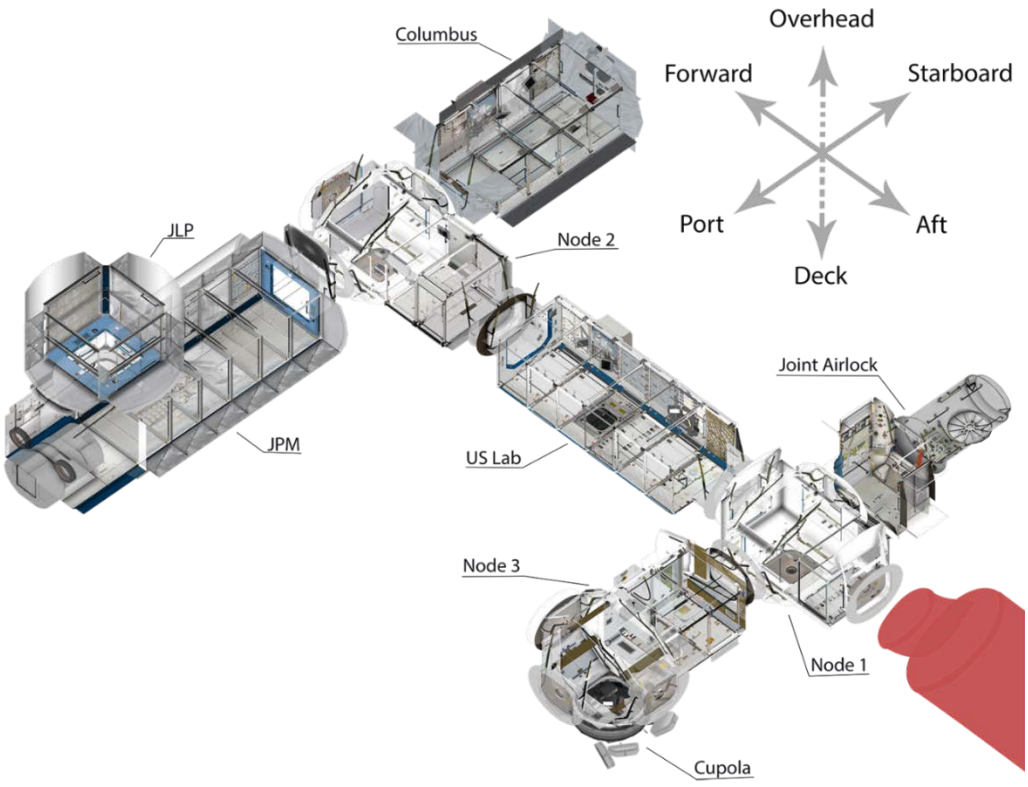

Figure 1: Simulated ISS configuration. The Russian segment of the ISS was not included in the simulation.

(a) (b)

Figure 2: The (a) VR-ISS and (b) a fireport.

To aid detection and location of the source of smoke, participants experienced different virtual smoke and corresponding atmospheric contaminant levels based upon the Treatment or Control group to which they were assigned (i.e., Training with light-smoke or no-smoke).

Atmospheric contaminant levels rose as a function of time and distance from the fire source. However, the smoke density changed as a function of time; therefore, participants could not rely solely on visual smoke cues to detect the source. Participants evaluated contaminant levels using a hand-held joystick programmed to emulate the NASA-used Compound Specific Analyzer–Combustion Products (CSA-CP) device. The purpose of the CSA-CP on board the ISS is to determine the level of atmospheric contaminants that are expected to be released due to potential fire, specifically dictating the length of time before Protective Breathing Apparatus is required. Participants were instructed to assess the atmospheric state by using the CSA-CP to display the levels of carbon monoxide (CO), hydrogen chloride (HCl), and hydrogen cyanide (HCN) in parts per million (Figure 3). Upon participant command, a floating window appeared in front of the participant with the contaminants concentration values. The window disappeared after five seconds. Based upon the participant's recall of the previously assessed contaminant levels in each VR-ISS module, the highest reading would indicate the approximate location of the source of smoke.

Figure 3: The CSA-CP in the VRE displays environmental noxious gas readings (left) and fireport with identifier code (right).

Once the participants identified the ISS module where the fire source was located, they began sampling fireports within the module to locate the "rack" that caused the fire. The VR-ISS

includes fireport labels accurately placed on the racks throughout the ISS (Figure 2). The labels have a unique code identifier which includes the module name, module surface, and rack number. Participants were trained on the fireport identifier codes during the first laboratory visit. The simulation ended when participants identified and reported the fireport label on the individual module rack which had the highest CSA-CP reading or when the simulation smoke became condensed to a level where visibility was almost zero (which occurred ten minutes from the beginning of the simulation), at which point the experiment controller stopped the experiment.

Procedure

The experiment was divided into three laboratory visits, each lasting approximately 60 minutes (Figure 4). The second and third laboratory visits, consisting of the experimental sessions, occurred at least 24 hours apart (*M*=39 h, *SD*=35). The first laboratory visit served to (1) educate and orient the participant to the ISS through VRE practice, and (2) develop task skills necessary to perform a simple fire response procedure. Participants were trained on the ISS layout and modules, how to navigate using module labels (e.g., PORT=left side, STBD=right side), and how to identify key landmarks within the modules (e.g., locations of the treadmill and cupola). The participants were then trained on the ISS fire procedures which included equipment, fireport rack labeling (e.g., “JPM1F3”), and proper procedure responses. Participants were given a guided acclimation walkthrough in the VR-ISS, which included reiteration of the ISS layout, navigation, landmarks and operation of the VRE. At the end of this visit, a written test was used to confirm participants’ ability to navigate the VR-ISS and perform the emergency fire procedure.

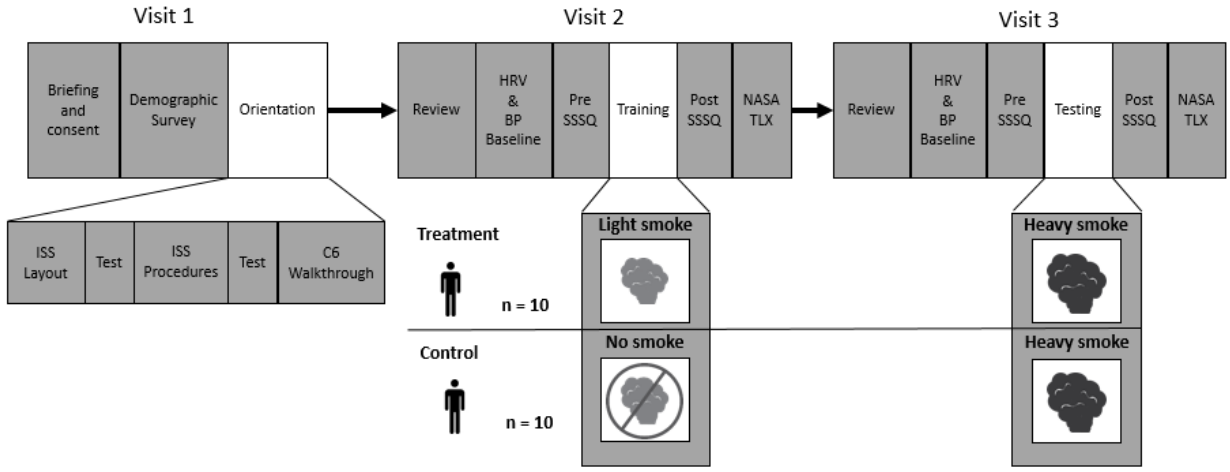

Figure 4: Design and procedure of the study.

During the second visit, termed the Training session, participants completed the emergency procedure in either a light-smoke or no-smoke condition (Figure 4) based on their assignment to the Treatment or Control group, respectively. Participants were not informed of their group assignment, only that they would be performing the emergency fire response in the VR-ISS. Prior to entering the VR-ISS, the participants were given a brief review of the ISS layout and navigation. Before the session began, participants completed the Short State Stress Questionnaire (SSSQ) to assess subjective stress levels. Participants then sat quietly for 10 minutes while baseline physiological data were collected. They then entered the VR-ISS and completed the fire response procedure. Post-session SSSQs were completed after finishing the simulation.

In the third visit, termed the Testing session, participants completed the same emergency response procedure as in Training. However, both the Treatment and Control groups were exposed to higher ("heavy-smoke") levels of virtual smoke and atmospheric contaminants. Participants were unaware that the smoke levels would increase significantly. The review, questionnaires, and physiological measurements were administered as described for the Training

visit. At the end of the experiment, participants were debriefed on the implementation of the heavy smoke conditions.

Independent Variables

There are two independent variables in the experiment: Session (Training, Testing) and Group (Control, Treatment). Participants were placed into one of two Groups: (1) a Treatment group with prior light-smoke exposure (Training) followed by heavy-smoke (Testing); and, (2) a Control group with no-smoke exposure (Training) followed by heavy-smoke (Testing).

Dependent Variable Measures

The study used both psychological and physiological indices of stress: perceived subjective stress state and psychophysiological biomarkers of the stress response. It has been recommended that training in VREs should measure stress through multiple standardized avenues such as cognitive perceived state, mediating factors, or psychophysiological biomarkers (Serino et al., 2014). Thus, data collected in the experiment included autonomic nervous system responses and cognitive workload. The dependent variables are summarized in Table 1.

Table 1: Description of dependent variable metrics, units, and frequencies

Dependent Variable	Metric	Components	Association	Unit	Measurement Frequency
Autonomic stress response	Heart rate variability (HRV)	HR	Cardiac activity	BPM	Baseline before session, throughout each session
		HF n.u.	Parasympathetic (i.e., vagal activity)	dimensionless	
		LF n.u.	Sympathetic & Parasympathetic	dimensionless	
		LF/HF ratio	Sympathovagal balance	dimensionless	
		RMSSD	Parasympathetic	ms	
		pNN50	Parasympathetic	%	

Table 1 Continued

Dependent Variable	Metric	Components	Association	Unit	Measurement Frequency
Autonomic stress response	Blood pressure	Systolic (SBP)		mmHg	Before session, throughout sessions
		Diastolic (DBP)		mmHg	
Psychological stress response	Short Stress State Questionnaire (SSSQ)	Engagement, Distress, Worry		Likert scale	Pre- and post-session
Workload	NASA Task Load Index (TLX)			Likert scale	post-session
Time-to-completion				minutes	Throughout each session

Aut**onomic Nervous System (ANS) responses**

Two overlapping branches of the ANS, the sympathetic nervous system branch and parasympathetic nervous system branch, determine the arousal or restorative functions targeting organs in response to a stressor. The overall "stress response" of the heart and vascular system is a result of the interplay between these two branches. It includes the effects of locally secreted neurotransmitters (e.g., norepinephrine) as well as systemic modulators (e.g., epinephrine). In essence, the sympathetic nervous system primes the body for action while the parasympathetic nervous system regulates organ and gland functions during rest.

The ANS responses to stress were assessed with two measures: Heart rate variability (HRV) and blood pressure (BP). HRV is comprised of time domain and frequency indices that reflect the balance between ANS-mediated relaxation or arousal (Hourani et al., 2011; Malik, 1996). Heart rate data were collected via electrocardiogram (ECG, modified CS_5 lead configuration). The ECG was sampled at 2048 Hz using Biopac MP150 hardware and recorded using AcqKnowledge software (Version 3.8.2, Biopac Systems Inc). Spectral analysis of ECG

was performed using the Matlab-based toolbox Kubios HRV software (Niskanen, Tarvainen, & Ranta-Aho, 2004). The raw data were first inspected visually for artifacts and corrected using the Kubios artifact correction option; the default low artifact correction level of Kubios was used for detecting RR intervals differing “abnormally” from the local mean RR interval (Tarvainen & Niskanen, 2012). First order trend correction was applied. Spectral density analysis of the HRV was used to parse the data into a low-frequency (LF) (0.04–0.15 Hz) band reflecting sympathetic activity with vagal modulation, and a high-frequency (HF) (0.15–0.4 Hz) band reflecting parasympathetic activity. The very-low-frequency (VLF, <0.04 Hz) band was not included in this study because it is unreliable for short term recordings (<5 min) (Malik, 1996). The LF and HF components were normalized to their total power in order to remove influences of VLF and the influence of changes in total power that may occur with autonomic arousal (e.g., HF/ (HF+LF) × 100). The LF/HF ratio was calculated to assess sympathovagal balance, which is an index of the relative amount of sympathetic activity (the extent at which the individual is hyper-aroused for action; sometimes referred to as ‘fight or flight response’) relative to parasympathetic activity (the extent to which an individual feels at ease) of the ANS. Therefore, this ratio concisely represents the individual’s physiological stress response. Time domain analysis of the ECG was performed to quantify the amount of variance in the inter-beat-interval through the root-mean-square difference of successive normal R-wave intervals (RMSSD) and the proportion of the number of pairs of successive normal R-waves that differ by more than 50 ms (pNN50). Both RMSSD and pNN50 represent vagal control within the time domain and are correlated to HF power (Shaffer, McCraty, & Zerr, 2014). The HRV time domain and frequency bands for each participant were calculated in 60-second intervals over the duration of each

session. The first minute of the data were omitted to prevent anticipatory stress responses from skewing the assessments.

Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were collected as another measure of cardiovascular reactivity. DBP and SBP can reflect changes in the total peripheral resistance of blood vessels. Increases in local sympathetic activity causes constriction of blood vessels, while reductions in sympathetic activity (or more parasympathetic activity) lead to dilation. In the absence of changes in cardiac output, decreases in blood vessel constriction are usually reflected by decreases in DBP. In the present study, beat-to-beat blood pressure data were collected. A finger cuff was placed on the participants' non-dominant hand over the middle phalanx of either the long or ring finger (Finapres 2300; Ohmeda). The non-dominant arm was placed in an arm-sling to standardize the position of the hand relative to the heart between all participants. Data were recorded at 1,024 Hz. After instrumentation and before each session, participants sat quietly for 10 minutes while baseline physiological data were collected. To calibrate the finger cuff, an oscillometric non-invasive blood pressure cuff (CNAP Monitor 500, CNSystems Medizintechnik AG) was placed on the participant's dominant upper arm and BP was measured twice during this 10-minute baseline period. The two systems showed similar BP measurements suggesting the arm-sling sufficiently accounted for potential hydrostatic pressure differences between the fingers and heart level. The raw data were inspected visually for artifacts and corrected using AcqKnowledge software. BP values were saved in 15-second interval samples. To give ample time for a resting state to occur and to prevent anticipatory stress interference, baseline BP data were calculated as the mean of the data from minutes five to eight of the 10-minute baseline. The mean baseline value was subtracted from the session data to

determine change scores (Zhang & Han, 2009). The first minute of the data were omitted to prevent any anticipatory stress interference.

Psychological Stress Response

The Short State Stress Questionnaire (SSSQ) was administered before and after each laboratory visit to assess multiple dimensions of the subjective response to stressful environments. The SSSQ assesses three state factors: task engagement, distress, and worry (Helton, 2004). Engagement refers to qualities of energetic arousal, motivation, and concentration. Distress is characterized by feelings of tense arousal, hedonic tone, and confidence-control. Worry relates to self-focus, self-esteem, and cognitive interference (Matthews et al., 1999). The three factor SSSQ scale scores for pre- and post-session were calculated for each participant. The factor scores from both pre- and post-session were standardized against normative means and standard deviation values from a large sample of British participants obtained by Matthews et al. (2002) and using the method of Helton and colleagues (Helton, Matthews, & Warm, 2009; W. Helton, 2004). Average difference scores for each state measure were then calculated for Treatment and Control groups. These were used to calculate the absolute difference between sessions, resulting in a z-score.

Workload

The NASA Taskload Index (TLX) was used to assess the subjective workload of the task during exposure (Hart & Staveland, 1988). The NASA TLX measures six dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration level. The NASA TLX was administered after the completion of a session. Participant scores on the six numerical rating scales were computed in the 0 to 100 range and as an unweighted participant mean for each of the six-dimensional subscales (Nygren, 1991).

Time-to-complete

The time-to-complete the fire response procedure was used to assess the stressor duration between participants. Since the stressor intensity (i.e., smoke density) increased over time, the length of time spent completing the procedure could influence the stress response.

Materials

The research was conducted in the C6, the high resolution virtual reality room at the Virtual Reality Applications Center (VRAC) at Iowa State University. The room is a 10 ft. x 10 ft. x 10 ft. cube in which all six screens have projected 4K stereoscopic images that provide total immersion in a virtual world. VirtuTrace, a full-featured "experiment engine," allows researchers to develop immersive experiment protocols for display in a fully immersive system (see Keren, Franke, Bayouth, Harvey, & Godby, 2013). Users moved in the virtual environment by taking a step in the desired direction, whereupon the system would track standing body position to facilitate motion through the environment. The VR-ISS was created using NASA-provided models of the U.S. Orbital Segment interior of the ISS (Figure 1). The Russian segment of the ISS was not included in the simulation since a model of this segment was not available.

Data Analysis Plan

Data analysis was performed using SPSS software (Version 23.0; IBM Corp.). For comparison of HRV components and BP, a linear mixed model (LMM) was used to calculate the fixed effect interaction of Group and Session. Random effect from participant sampling was used in the covariance matrix. All HRV metrics and HR were Winsorized to three standard deviations to reduce the impact of outliers. The LF/HF variables had a moderate positive skew and was Log(x+1) transformed to normalize the data for parametric analysis. Subjective stress questionnaires (SSSQ) and workload questionnaires (TLX) were checked for normality and subsequently assessed using the non-parametric Mann-Whitney U test. Results were considered

significantly different at the $p \leq 0.05$ level. A statistical trend is defined as results with a $0.05 < p < 0.10$. All results shown are means (*M*) and standard error (*SE*).

Effect sizes were calculated for the fixed effects and interaction effects. Cohen's *d* (Cohen, 1988) was used for the standardized effect size in units of standard deviation. Cohen's *d* effect size guidelines are reported as small for $0.2 < |d| < 0.5$, medium for $0.5 < |d| < 0.8$, and large for $|d| > 0.8$ (Cohen, 1988). Effect size for the Mann-Whitney U tests were calculated with normal approximation *z* to *r*. Cohen's guidelines for Pearson correlation *r*-score effect size are adopted as small for $0.1 < r < 0.3$, medium for $0.3 < r < 0.5$, and large for $r > 0.5$ (Fritz et al., 2012).

To assess whether the standard deviations of the interaction differed between the Group and Session, a 95% confidence interval (*CI*) was generated by the parametric bootstrap resampling technique (Efron & Tibshirani, 1993). Parametric bootstrapping uses a fitted model based on the experiment sample data to generate synthetic data through replication, which yields a sampling distribution for a larger population. The bootstrap population distribution can then provide a robust empirical *CI* estimation. Some studies have recommended at least 2,000 replications for *CI* (DiCiccio & Efron, 1996; Efron, 1987), the present study used 10,000 replications for greater *CI* accuracy. Bootstrapping was performed using R software (Version 3.3.2; R Foundation for Statistical Computing). The mixed model in R was verified by comparison of maximum likelihood estimation in SAS software (Version 9.2; SAS Institute).

Results

ANS Stress Response by HRV and Heart Rate

Comparison of baseline physiological measures between groups revealed no significant differences. Physiological data during the VR-ISS procedures is presented in Table 2 and Table

3. The main effects of Group and Session were not significant for HR but differences were seen with the HRV variables.

Table 2: Descriptive statistics for the LMM measures of HRV and HR, mean (SE).

	Control		Treatment	
Metric	**Training**	**Testing**	**Training**	**Testing**
Log(LF/HF+1)	0.51 (0.059)	0.66 (0.062)	0.51 (0.056)	0.54 (0.057)
LF n.u.	63.3 (3.9)	74.3 (4.12)	62.6 (3.7)	65.3 (3.76)
HF n.u.	36.5 (3.88)	25.5 (4.1)	37.1 (3.68)	34.5 (3.75)
HR (BPM)	83.2 (4.05)	86.4 (4.08)	84.4 (4.01)	82.1 (4.02)
RMSSD (ms)	47.9 (4.34)	35.8 (4.13)	49.4 (4.13)	48.0 (4.20)
pNN50 (%)	18.6 (3.40)	10.7 (3.58)	18.2 (3.33)	17.9 (3.36)

Table 3: Significance level and effect size for the LMM measures of HRV and HR.

Metric	$P_{\text{Fixed effect}}$ **(Group)**	$P_{\text{Fixed effect}}$ **(Session)**	$P_{\text{Interaction}}$ **(Group*Session)**	**Cohens *d* (Interaction effect only)**
Log(LF/HF+1)	0.455	0.006 *	0.051	-0.49
LF n.u.	0.343	0.003 *	0.066	-0.48
HF n.u.	0.345	0.003 *	0.062	0.49
HR (BPM)	0.786	0.765	0.092	-0.30
RMSSD (ms)	0.232	0.01 *	0.042 *	0.57
pNN50 (%)	0.465	0.006 *	0.011 *	0.50

Note: * indicates Significance

A main effect was seen for Session, where LF/HF was significantly higher for Testing (M = 0.6, SE = 0.042) compared to Training (M = 0.50, SE = 0.041), $F(1, 58) = 8.13$, $p = .006$, d = -0.70. A significant increase was found for normalized LF for Testing (M = 69.8, SE = 2.8) compared to Training (M = 63.0, SE = 2.7), $F(1, 60) = 9.41$, $p = .003$, d = -0.79. In contrast, a significant decrease was seen for normalized HF for Testing (M = 30.0, SE = 2.8) compared to Training (M = 36.8, SE = 2.7), $F(1, 59) = 9.47$, $p = .003$, d = 0.79. The HRV time domain indices also changed. RMSSD was significantly decreased for Testing (M = 41.9, SE = 3.1) compared to Training (M = 48.7, SE = 3.0), $F(1, 46) = 7.14$, $p = .01$, d = 0.72, and pNN50 was significantly decreased for Testing (M = 14.3, SE = 2.4) compared to Training (M = 18.4, SE = 2.4), $F(1, 35) = 8.56$, $p = .006$, d = 0.54. No significant main effects for Group were found for

any HRV metrics but as described below, several Multiple Group × Session interaction effects were seen.

The interaction effect for normalized HF of HRV during Testing showed a statistical trend (i.e., $0.05 < p < 0.10$) for increasing in the Treatment group compared to the Control group, $F(1, 59) = 3.61$, $p = .062$, with a small-moderate effect size, $d = 0.49$ (Figure 5). The normalized LF had a decreasing trend for the Treatment group compared to the Control group, $F(1, 60) = 3.51$, $p = .066$, $d = -0.48$ (Figure 6). Likewise, the interaction effect for the LF/HF responses due to differences in Session approached statistical significance, $F(1, 58) = 3.97$, $p = .051$, $d = -0.49$ (Figure 7). The interaction effect for HR also showed a trend, $F(1, 19) = 3.15$, $p = .092$, $d = -0.30$ (Figure 8). The interaction effect for the time domain indices of RMSSD was significantly higher for the Treatment group compared to the Control group, $F(1, 46) = 4.39$, $p = .042$, $d = 0.56$ (Figure 9) and the pNN50 was significantly higher for the Treatment group compared to the Control group, $F(1, 35) = 7.25$, $p = .011$, $d = 0.50$ (Figure 10).

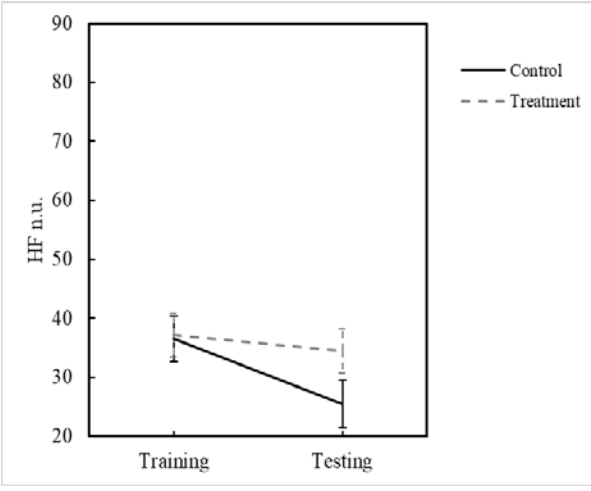

Figure 5: Mean and standard error of HF (n.u.)

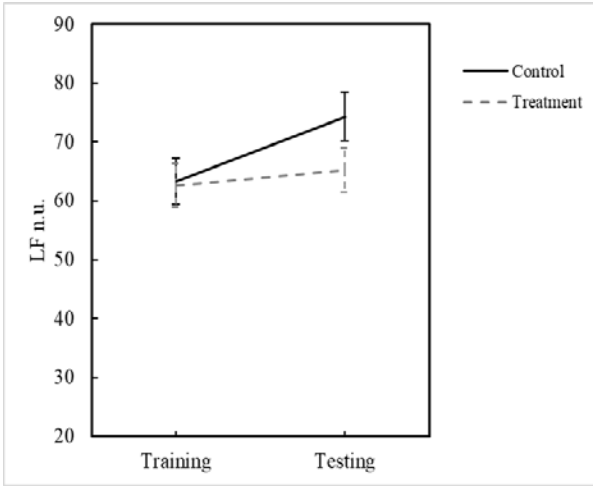

Figure 6: Mean and standard error of LF (n.u.)

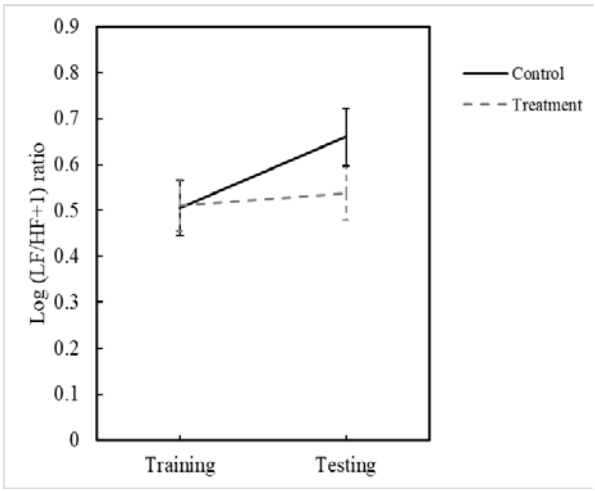

Figure 7: Mean and standard error of Log(LF/HF+1) ratio

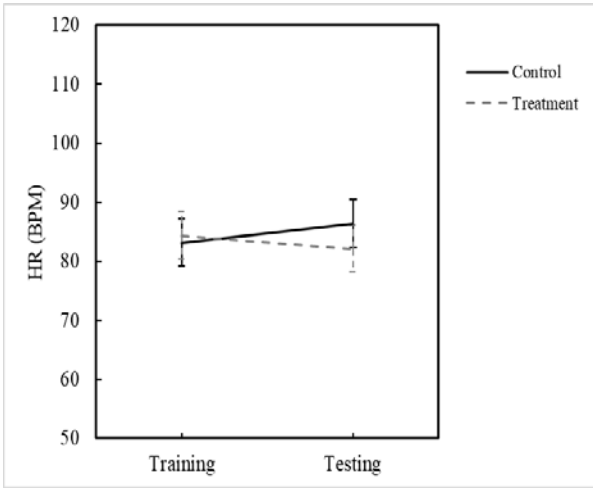

Figure 8: Mean and standard error of HR

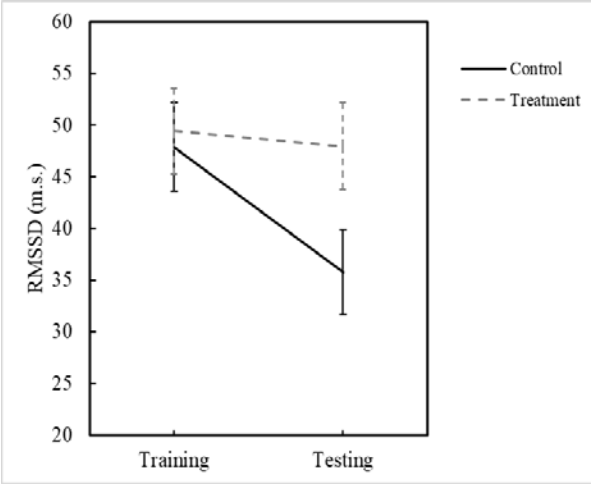

Figure 9: Mean and standard error of RMSSD

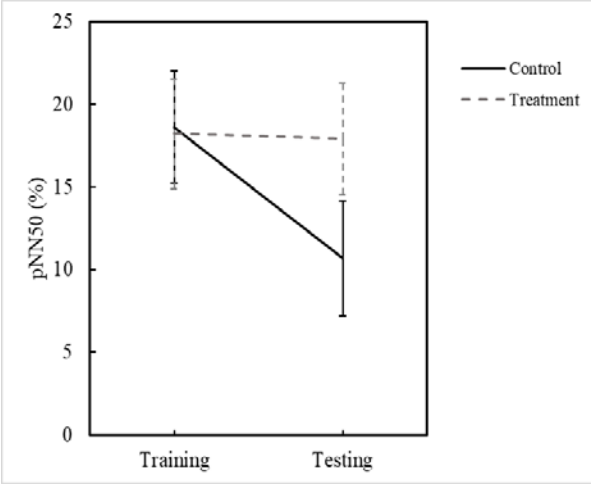

Figure 10: Mean and standard error of pNN50

The Group × Session interaction effect confidence intervals for a large population distribution, calculated via parametric bootstrap, showed that the original sample estimate was highly reliable when the original sample model was assessed for 10,000 replications. The interaction effects for LF/HF, RMSSD, and pNN50 were significant. The interaction effects for normalized LF, normalized HF, and HR showed a statistical trend (Table 4).

Table 4: Estimate of Interaction Effect with bootstrap 95% confidence intervals

	Estimate	Lower Bound	Upper Bound	Bootstrapped *p*-value (Group*session)
Log(LF/HF+1)	0.27	0.011	0.55	0.047 *
LF n.u.	-8.21	-16.5	1.08	0.067
HF n.u.	8.28	-0.76	16.8	0.065
HR (BPM)	-6.46	-13.8	1.14	0.091
RMSSD (ms)	12.8	1.57	23.6	0.023 *
pNN50 (%)	7.80	2.14	13.6	0.008 *

Note: * indicates Significance

ANS Stress Response by Blood Pressure

From the data presented in Table 5 and Table 6, no significant Group or Session effect was found. The DBP response for the Group × Session interaction effect illustrated in (Figure 11) was not significant, $F(1, 40) = 1.91$, $p = .174$, $d = .32$. Similarly, no significant difference was found with the SBP interaction effect, $F(1, 26) = .043$, $p = .836$, $d = -.06$ (Figure 12). Parametric bootstrap of the Group × Session interaction effect shown in Table 7 confirmed that DBP and SBP would remain unchanged even in a larger population sample.

Table 5: Descriptive statistics for the LMM measures of BP, mean (SE)

	Control		Treatment	
	Training	Testing	Training	Testing
SBP (mmHg)	25.2 (1.04)	21.8 (1.28)	26.4 (0.73)	27.0 (1.60)
DBP (mmHg)	11.0 (0.49)	9.05 (0.80)	11.1 (0.47)	14.1 (0.64)

Table 6: Inferential statistics for the LMM measures of BP, p-values and effect size of interaction

	$P_{\text{Fixed effect}}$ (Group)	$P_{\text{Fixed effect}}$ (Session)	$P_{\text{Interaction}}$ (Group*Session)	Cohens *d* (Interaction effect)
SBP (mmHg)	0.44	0.87	0.84	-0.06
DBP (mmHg)	0.31	0.40	0.17	0.32

Note: * indicates Significance

Table 7: Estimate of Interaction Effect with bootstrap 95% confidence intervals

	Estimate	**Lower Bound**	**Upper Bound**	**Bootstrapped *p*-value (Group*Session)**
SBP (mmHg)	-0.54	-18.3	18.2	0.955
DBP (mmHg)	3.84	-3.93	11.7	0.334

Note: * indicates Significance

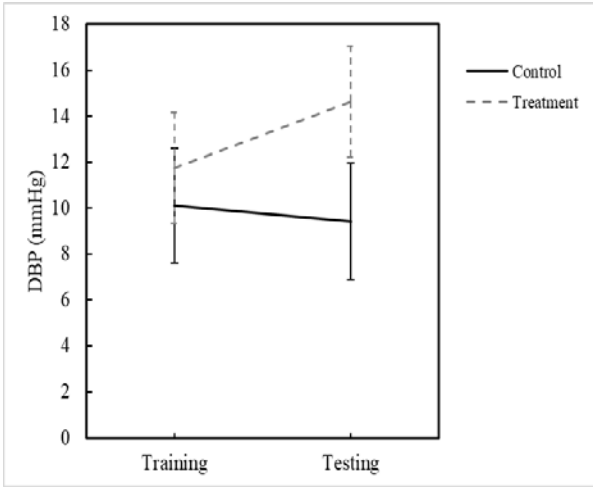

Figure 11: Mean and standard error of DBP from baseline.

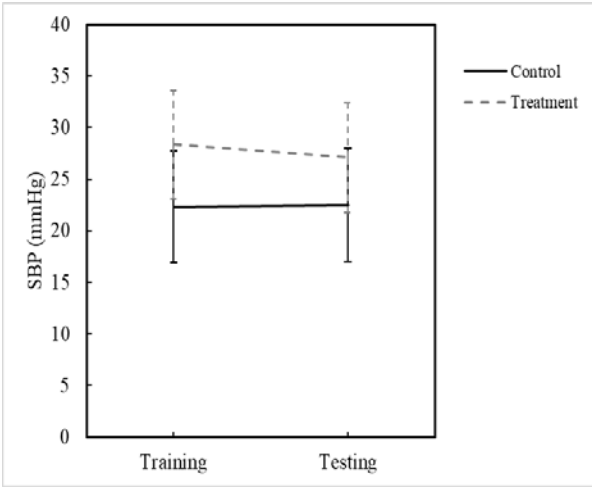

Figure 12: Mean and standard error of SBP from baseline.

Stress State

The main effect of Group and the main effect of Session were not significant for the SSSQ scale factors. However, the simple main effect for the task engagement factor during

Training session was significantly less for the Treatment (light-smoke) condition (M = 0.22, SE = 0.13) than for the Control (no-smoke) condition (M = 0.46, SE = 0.071), U = 23.5, p = .041, r = -.46. The simple main effect in the Testing session was not significantly different between the Treatment (M = 0.41, SE = 0.14) and Control (M = 0.34, SE = 0.11), U = 47.5, p = .85, r = -.043.

Figure 13 illustrates the change in z-score between the Training and Testing for both the control and Treatment sessions for task engagement, distress, and worry. The task engagement change score was not significantly different between the Treatment group (M = 0.19, SE = 0.2) compared to the Control group (M = -0.12, SE = 0.12), U = 31, p = .137, r = -.33. The distress change score was not significantly different between the Treatment group (M = 0.095, SE = 0.093) compared to the Control group (M = 0.095, SE = 0.067), U = 48, p = .876, r = -.035. The worry change score was not significantly different between the Treatment group (M = 0.32, SE = 0.23) compared to the Control group (M = 0.055, SE = 0.097), U = 43.5, p = .621, r = -.11. All other sources of variance in the analysis lacked statistical significance or trends, $p > 0.10$.

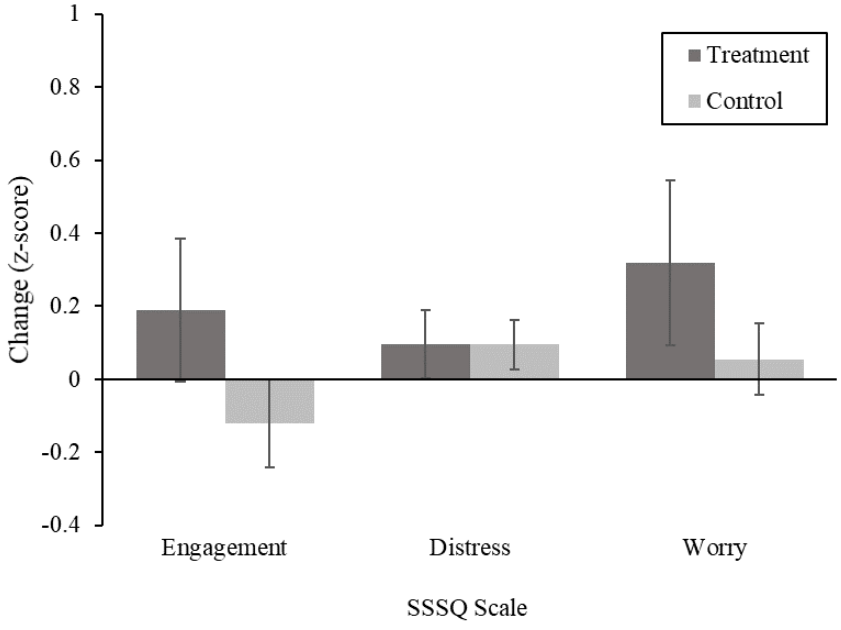

Figure 13: Mean and standard error of SSSQ change scores between sessions.

Workload (NASA-TLX)

Figure 14 illustrates the workload scores for both Training and Testing. No significant differences were found for the Group main effect. The main effect of Session on temporal demand indicated that Testing (*Mdn* = 76.2) had significantly greater temporal demand than Training (*Mdn* = 49.3), *U* = 114.5, *p* = .02, *r* = -.37. The main effect of Session on physical demand, but also that Testing (*Mdn* = 28.6) has greater physical demand than Training (*Mdn* = 23.8), *U* = 127.5, *p* = .048, *r* = -.31. When evaluating the TLX within a single session, simple main effects for Group were not significant.

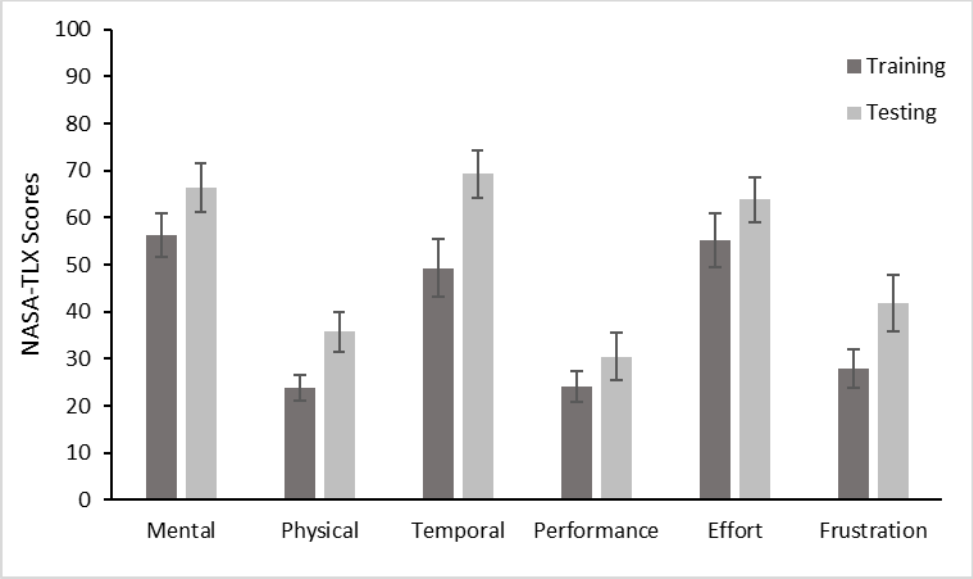

Figure 14: Workload profile for Session main effects (mean and standard error).

Time-to-complete

The main effect for Group on the time-to-complete was significantly higher for Treatment (*M* = 7.29, *SE* = 0.66) than for the Control group (*M* = 5.32, *SE* = 0.66), *F*(1, 18) = 4.43, *p* = .05, with a large effect size *d* = 1.19. A main effect trend was detected for Session, such that the time to complete the emergency procedure was higher for Training (*M* = 6.73, *SE* =

0.52) than for Testing (M = 5.89, SE = 0.52), $F(1, 18) = 3.30$, $p = .086$, $d = 0.51$. No significant interaction was found.

Discussion

The purpose of this study was to assess the extent to which graduated exposure to a task-specific stressor affected the physiological response to a critical spaceflight hazard. Specifically, we hypothesized that exposure to a light level of virtual smoke, during training that is focused on responding to a fire threat on the International Space Station, would attenuate the psychophysiological responses to a subsequent simulated emergency with an unexpected heavy-smoke condition. Compared to the Control group, the autonomic responses of the Treatment group suggested that they had relatively less activation of the sympathetic nervous system, enhanced allostasis, and adaptability to a stress response. These results suggest that prior exposure to a low-level stressor attenuates the sympathovagal response to a more stressful condition of the same task in virtual reality. Moreover, these psychophysiological measures suggest that using only one component of the SIT framework, graduated exposure to a stressor, can positively affect the responses to exposure to a more severe version of the stressor (i.e., Testing).

When evaluating the ability to stay calm as indicated by the ANS stress response, the unchanged state of the Treatment group's normalized HF component, RMSSD, and pNN50 suggests that participants retained parasympathetic modulation. In contrast, the Control group's parasympathetic withdrawal is characteristic of a mild stress response. This response suggests they were more stressed and relatively unable to maintain a calm behavioral state (Shaffer, McCraty, & Zerr, 2014). The Control group was relatively unaware of this change, as evident by a lack of between-group differences in their SSSQ scores.

Any changes seen in the HRV were primarily meditated by withdrawal of the parasympathetic branch of the ANS (Porges, 1995) leading to relatively greater sympathetic control as indicated by the decrease in normalized HF and increase in normalized LF (i.e., comparing Figure 5 to Figure 6. This interpretation is consistent with Hjortskov et al. (2004) who found that short lasting exposure to psychosocial stressors indicated parasympathetic withdrawal along with an unchanged sympathetic activity to be responsible for increase in LF/HF ratio. Further, associations between resting autonomic balance and psychological resilience are supported by the polyvagal theory (Porges, 1995) in which the primary response to a stressor is mediated by the parasympathetic nervous system component of the ANS (Lewis et al., 2015). Collectively, these findings suggest that without prior exposure to a mild threat, participants were unable to relax when exposed to the more severe threat, thus contributing to greater autonomic arousal and higher stress levels.

The results reinforce the use of HRV, rather than HR, as a tool for measuring psychophysiological stress response. Results showed a weakly elevated HR for the Control and decrease for Treatment between the sessions. However, the trend is small and only a subtle reflection of the ANS activity (Shaffer et al., 2014). Similarly, blood pressure remained unchanged suggesting no change in vasoconstriction during the stress response. While HR and blood pressure can be perceivable indications of stress at times, the non-reaction during a stressful situation verifies the usefulness of HRV to detect stress without relying on human subjectiveness.

The Control group's SSSQ task engagement during Training was higher than the Treatment group. Due to the Control group not experiencing smoke exposure in Training, the absence of physiological stress may have resulted in a sense of control and challenge-related

appraisal, and therefore higher levels of engagement for the Control group. Challenge appraisal and effective coping have been shown to be associated with higher parasympathetic activation (Geisler et al., 2013; Laborde, Lautenbach, & Allen, 2015). The Treatment group's unchanged engagement between sessions possibly reflects the parasympathetic activation, indicating that inoculation fostered a challenge perspective and coping. However, the Control groups parasympathetic withdrawal during the Testing session suggests the challenge became a threat appraisal and subsequently resulted in stress (Schwerdtfeger & Derakshan, 2010). While the absence of stress during Training may have been relatively beneficial in the short-term for the Control group, it did little to prepare them for the more severe Testing.

The subjective assessment of task workload showed no change between Groups, but changed between Sessions for the temporal and physical workload dimensions. These results are reinforced by the high SSSQ engagement results for both groups. Previous studies have suggested that task engagement has moderate correlations to the temporal and physical dimensions of the NASA TLX, but also has high correlations to TLX's mental demand and effort (Matthews et al., 2013). Possible explanations for the increased temporal and physical response to Testing are the increase in perceived time pressure of locating a fire and the greater physical fatigue associated with the stress response evoked by the heavy smoke scenario. The unchanged state of mental demand and effort between Sessions can be explained by the already relatively high mental demands of practicing the emergency fire procedure during the Training session. The procedure requires navigation and spatial knowledge as well as short-term memory task of remembering containment levels to support any movement decisions. Navigational processes include visual attention, spatial memory, and working memory operations which feed information to executive function decision making (Bajaj et al., 2008).

In stress training interventions which involve performing tasks, it may be difficult to distinguish whether the stressor or the task is primarily influencing HRV. One could question whether changes in HRV between sessions may be partly attributed to the change in workload rather than the exposure of the experimental stressors. Changes in task complexity and workload influence HRV (Jorna, 1992), which are more pronounced in executive level functions such as sustained attention (Luque-Casado et al., 2016). Similarly, HRV is correlated to the stress caused by evaluations of threat and safety and inhibition of unwanted memories and intrusive thoughts (Shaffer et al., 2014; Thayer et al., 2012). Based on this interconnectedness, both the perceived threat caused by the stressors and the executive control required by the task are theorized to be feedback and feedforward in nature, allowing both functions to simultaneously influence regulation of autonomic nervous system (Thayer et al., 2009). Therefore, without a clear understanding of the extent to which HRV measurements are influenced by task or stressor, stress response improvements (lower HRV) might be due to task learning effect from improved executive control rather than resilience or an increased ability to relax during the stressor. To deconfound this problem, we can compare changes in task engagement and changes in workload across sessions and groups.

Executive control has been shown to be associated with to SSSQ task engagement (Matthews & Zeidner, 2012). Executive functions require short-term memory, focused attention, or manipulation of new information (Thayer et al., 2009). In this task, the executive functions are remembering the NASA emergency procedures while maintaining sustained attention to fire-related cues and navigating the VR-ISS to detect and locate the fire. The progression from the Training to Testing session showed increase in temporal and physical workload, while the SSSQ task engagement from remained unchanged. However, the autonomic nervous system activities

reflected by HRV showed a large parasympathetic decrease for the Control group during the experiment, but no change for the Treatment group (see Figures 9 and 10). The workload demands increased from the Training session to the Testing session even though there was no change in the task engagement, which suggests that the executive function caused by the task was not responsible for the stress response. Collectively, the lack of SSSQ task engagement differences and the presence of HRV differences between Groups in the Testing session suggest that the perceived threat from the simulated environment (i.e., scenario change from light/no smoke to heavy smoke) was the primary source of stress for participants, without direct influences of stress elicited from either the changing workload or the procedures.

In the present study, high workload levels and a decreased engagement between trials for the Control group would have provided evidence that experiencing a more stressful situation without inoculation may result in an overload of attentional resources. As complexity increases for memory and search tasks, allocation of free resources can be impaired and decrease the task engagement (Matthews & Davies, 2001). The impairment is due to visual search tasks and spatial information using common mechanisms in working memory (Woodman & Luck, 2004). We would have expected that the simulated threat would have interfered with executive function more in the Control group that the Treatment group because of the larger difference in the manipulated smoke levels (from training to testing). Therefore, we would have expected to see the task engagement decrease from Training to Testing in the Control group, but remain unchanged from Training to Testing for the Treatment group. However, the task engagement from the Training to Testing session for the Control group did not rise to the level of significance (p = .137, see Figure 13).

The time to complete the procedure was one minute longer for Treatment compared to Control, but also two minutes longer for Training compared to Testing. Although the treatment group took two minutes longer to complete the task compared to the Control group, the (LF/HF) power spectra during the training session suggests no difference in the stress response. Both groups completed the task quicker in the Testing session than the Training session, most likely due to familiarization with the procedure. While the time to complete the procedure differed between Group and Session, the participants were only told they would be practicing the emergency procedure and not informed that a rapid emergency response would be evaluated or was more favorable. In other words, "performance" was not emphasized as a desired attribute. The longer completion time for the Treatment group during Training might have resulted from decrease in visibility during the task due to the increase in virtual smoke density.

It is important to address the reduced emphasis on performance. One could consider time-to-completion as a performance metric that we would expect would decrease with increased decision making. Keren et al. (2013) studied firefighters' decision making under stress in a virtual reality environment. The results indicated that 'time-to-decision' was significantly longer with veteran firefighters when compared to novice firefighters, potentially indicating novices out performed veterans. However, the investigation demonstrated that veterans used their experience to better assess the time window available for a response, which in turn they used to enhance their situational awareness and to better size up the situation at hand, rather than debating the decision task. Thus, an increase in speed did not reflect better performance. Post-experiment decision analysis revealed that veterans' decision quality was higher. Bayouth et al., (2011) analyzed firefighters' decision performance under two different stressors: (1) difficult tradeoff; and, (2) time pressure. When under time pressure, time-to-decision did not vary between

veterans and novices. The quality of veterans' decision quality was less prevalent when under time stress rather than under difficult tradeoffs. Thus, performance is a difficult, complex attribute to assess.

Several factors may limit the generalizability of our conclusions. The participant sample size was adequate to measure psychophysiological measures, but a larger sample size would likely increase the reliability of our subjective stress measurements. The participant sample included males only, thereby potential gender effects could not be detected. The goal of this study was to assess the effect of stress training for healthy people working in challenging environments; however, future research on the efficacy of using stress training for specific tasks would benefit from a sample more closely related to the relevant occupational demographic, such as astronauts. While stress appraisal and coping has substantial variability between individuals, task proficiency can lead to heightened sense of control and mitigate stress (Orasanu & Becker, 1996). Therefore, because an astronaut sample may be more proficient at emergency procedures than a college students sample, the effect of stress training may not sufficiently generalize to the representative population. An experiment with a sample with similar education, age, and gender that reflects astronaut population would strengthen the generalizability of the findings.

Although the graduated stress exposure elicited an improved physiological response after only two sessions, it remains uncertain whether more sessions will affect the benefits. In a review by Saunders et al. (1996), the beneficial effects of SIT increase as the number of training sessions increase. A single session can be sufficient for tempering the stress response, but five to seven sessions produce a robustly positive effect. Further, the present study only evaluated the application phase of SIT (3^{rd} phase) and intentionally omitted the two phases focused on stress education and acquisition of coping skills. Participants were not given any instruction in these

two areas and their coping abilities were not assessed prior to the experiment. Individual differences in coping abilities could potentially have affected our results. Several limitations also exist for the HRV analysis and interpretation of results. There is an influence of respiration rate on the cardiac cycle which limits clear interpretation of the HF spectrum as an index of cardiac vagal tone. RMSSD and pNN50 may be correlated to HF power, however the influence of respiration rate on the indices is uncertain (Shaffer, McCraty, & Zerr, 2014). Further, there are concerns that the LF/HF ratio cannot be considered as an index of sympathovagal balance due to assumptions of vagal and sympathetic nerve traffic to the heart (Eckberg, 2000).

Conclusion

This work addresses a gap in the literature with respect to enhancing the resilience of healthy individuals to acute stress, which may be more pronounced in the realm of spaceflight. The current findings suggest that even modest graduated stress exposure shows promise as a useful tool during spaceflight procedure training to prepare for emergencies. Spaceflight is dangerous and much of the current training paradigm for astronauts is focused on task exposure and mastery. However, the present study suggests that psychophysiological responses to life-threatening situations can be mitigated by graduated stress exposure training and this can be done concurrent with task-specific training. The results suggest that participants who received prior exposure to a stressor enhanced their ability to remain calm during an emergency procedure task in a virtual ISS. Moreover, this study shows that graduated exposure training is a beneficial training component in SIT.

While not the focus of this study, the results suggest that a VRE system for training astronauts to handle stressful situations would likely increase their ability to cope and perform under acute highly stressful situations. The use of a VRE to administer gradual stress levels proved to be effective at eliciting an appropriate stress response to varying conditions. VREs

offer a safe and controlled environment for training for traumatic or hazardous situations. Simultaneously, VREs have potential to solve two common stress training problems, namely treatment consistency as well as reconciling differences between the training environment and the environment in which the task is performed (Meichenbaum, 2017). To date, VRE simulations for intra-vehicular activity have been used far less during NASA training in lieu of full-scale mock-ups (Aoki et al., 2008; Gancet, Chintamani, & Letier, 2012). Since a high-fidelity VRE utilizes less resources than traditional fire training using mockups, emergency training in VREs may be a suitable alternative for NASA astronauts in preparation for some aspects of missions to the ISS.

Future work is needed to study further the inoculation effects of using a stress training pedagogy for spaceflight applications. The current study did not explore whether the effects of SIT enhance performance for spaceflight operations, as has been theorized for other occupations (e.g., McClernon et al., 2011). In addition, although preliminary findings indicate that graduate exposure can attenuate relaxation mechanisms, future research should investigate the use of VR for all phases of SIT used in a preventative approach. Robson & Manacapilli (2014) note that when SIT in implemented in the full three-phases, current use of VR during SIT primarily occurs in the application phase, and this use is in the context of the full three-phased implementation of SIT. To administer SIT in its entirety in VR, a phased training approach with the three phases separated minimizes the interference of stressors affecting new trainees trying to learn skills (Friedland & Keinan, 1992). Although our intent was to evaluate graduated stress exposure in isolation, the full SIT framework in VR may have more pronounced improvements in stress response.

Acknowledgements

The authors are grateful to the NASA Johnson Space Center ER7's Virtual Reality Laboratory and Integrated Graphics Operations and Analysis Laboratory for providing the 3D models of the ISS. The authors thank Kevin Godby for his assistance in integrating the ISS models into VirtuTrace and for running VirtuTrace during the sessions and also thank Dr. Elizabeth Shirtcliff for offering expertise on the human stress response.

References

Abrams, M. P., Carleton, R. N., Taylor, S., & Asmundson, G. J. (2009). Human tonic immobility: measurement and correlates. *Depression and anxiety*, *26*(6), 550–6. doi: 10.1002/da.20462

Aoki, H., Oman, C., Buckland, D., & Natapoff, A. (2008). Desktop-VR system for preflight 3D navigation training. *Acta Astronautica*, *63*(7-10), 841–847. doi: 10.1016/j.actaastro.2007.11.001

Bayouth. S. T. (2011). *Examining firefighter decision making process and choice in virtual reality* (Doctoral Dissertation). Retrived from Iowa State University Digital Repository. 10350.

Bajaj, J., Hafeezullah, M., Hoffmann, R., Varma, R., Franco, J., Binion, D., Hammeke, T., & Seian, K. (2008). Navigation skill impairment: Another dimension of the driving difficulties in minimal hepatic encephalopathy. *Hepatology*, *47*(2), 596–604. doi: 10.1002/hep.22032

Baumann, M. R., Gohm, C. L., & Bonner, B. L. (2011). Phased Training for High-Reliability Occupations: Live-Fire Exercises for Civilian Firefighters. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 548–557. doi: 10.1177/0018720811418224

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillside, NJ: Lawrence Earlbaum Associates.

Crawford, C., Wallerstedt, D., & Khorsan, R. (2013). A systematic review of biopsychosocial training programs for the self-management of emotional stress: potential applications for the military. *Evidence-Based Complementary and Alternative Medicine*. doi: 10.1155/2013/747694

Delahaij, R., Gaillard, A.W.K., & Soeters, J.M.L.M. (2006). *Stress training and the new military environment.* In Human Dimensions in Military Operations – Military Leaders' Strategies for Addressing Stress and Psychological Support (pp. 17A-1 – 17A-10). Meeting Proceedings RTO-MP-HFM-134, Paper 17A. Neuilly-sur-Seine, France: RTO.

DiCiccio, T., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, *11*(3), 189–212.

Driskell, J. E., Salas, E., Johnston, J. H., & Wollert, T. N. (2008). Stress exposure training: An event-based approach. *Performance under stress*, 271–286. London: Ashgate.

Eckberg, D. L. (2000). Physiological basis for human autonomic rhythms. *Annals of medicine*, *32*(5), 341-349.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, *82*(397), 171–185.

Efron, B. & Tibshirani, R. J. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, *57*, 1-436.

Ellis, A. (2006). System breakdown: The role of mental models and transactive memory in the relationship between acute stress and team performance. *Academy of Management Journal*, *49*(3), 576–589. doi: 10.5465/AMJ.2006.21794674

Fletcher, D., & Sarkar, M. (2013). Psychological resilience: A Review and Critique of Definitions, Concepts, and Theory. *European Psychologist*, *18*(1), 12-23. doi: 10.1027/1016-9040/a000124

Foley, F. W., Bedell, J. R., LaRocca, N. G., Scheinberg, L. C., & Reznikoff, M. (1987). Efficacy of stress-inoculation training in coping with multiple sclerosis. *Journal of consulting and clinical psychology*, *55*(6), 919–22. doi: 10.1037/0022-006X.55.6.919

Fornette, M. -P., Bardel, M. -H., Lefrançois, C., Fradin, J., Massioui, F., & Amalberti, R. (2012). Cognitive-Adaptation Training for Improving Performance and Stress Management of Air Force Pilots. *The International Journal of Aviation Psychology*, *22*(3), 203–223. doi: 10.1080/10508414.2012.689208

Friedland, N., & Keinan, G. (1992). Training effective performance in stressful situations: Three approaches and implications for combat training. *Military Psychology*, *4*(3), 157. doi: 10.1207/s15327876mp0403_3

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology. General*, *141*(1), 2–18. doi: 10.1037/a0024338

Gaillard, A. W. K. (2001). Stress, workload, and fatigue as three biobehavioral states: A general overview. *Stress, workload, and fatigue*, 623-640. Mahwah, NJ: Lawrence Erlbaum.

Gancet, J., Chintamani, K., & Letier, P. (2012). Force feedback and immersive technologies suit (FITS): an advanced concept for facility-less astronaut training. In *International symposium on artificial intelligence, robotics and automation in space, Turin, Italy*.

Geisler, F., Kubiak, T., Siewert, K., & Weber, H. (2013). Cardiac vagal tone is associated with social engagement and self-regulation. *Biological Psychology*, *93*(2), 279–286. doi: 10.1016/j.biopsycho.2013.02.013

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Helton, W. (2004). Validation of a Short Stress State Questionnaire. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(11), 1238–1242. Los Angeles, CA: SAGE Publications. doi: 10.1177/154193120404801107

Helton, W. S., Matthews, G., & Warm, J. S. (2009). Stress state mediation between environmental variables and performance: The case of noise and vigilance. *Acta psychologica*, *130*(3), 204–213. doi: 10.1016/j.actpsy.2008.12.006

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, *92*(1-2), 84–89. doi: 10.1007/s00421-004-1055-z

Hourani, L. L., Kizakevich, P. N., Hubal, R., Spira, J., Strange, L. B., Holiday, D. B., Bryant, S., & McLean, A. N.(2011). Predeployment stress inoculation training for primary prevention of combat-related stress disorders. *Journal of CyberTherapy & Rehabilitation 4*(1), 101-116.

Johnston, J., & Cannon-Bowers, J. (1996). Training for stress exposure. *Stress and human performance*, 223–256. Mahwah, NJ: Lawrence Erlbaum.

Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology*, 237–257. doi: 10.1016/0301-0511(92)90017-O

Kalisch, R., Müller, M. B., & Tüscher, O. (2015). A conceptual framework for the neurobiological study of resilience. *The Behavioral and brain sciences*, *38*. doi: 10.1017/S0140525X1400082X

Keinan, G., & Friedland, N. (1996). Training effective performance under stress: Queries, dilemmas, and possible solutions. *Stress and human performance*, 257–277. Mahwah, NJ: Lawrence Erlbaum.

Keren, N., Franke, D. W., Bayouth, S. T., Harvey, E. M., & Godby, K. M. (2013). VirtuTrace: Training for Making Decisions under Stress in virtual environments. *Proceedings of the annual Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, FL.

Laborde, S., Lautenbach, F., & Allen, M. S. (2015). The contribution of coping-related variables and heart rate variability to visual search performance under pressure. *Physiology & behavior*, *139*, 532–40. doi: 10.1016/j.physbeh.2014.12.003

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping.* New York: Springer-Verlag.

Leipold, B., & Greve, W. (2009). Resilience: A conceptual bridge between coping and development. *European Psychologist, 14*(1), 40-50. doi: 10.1027/1016-9040.14.1.40

Lewis, G., Hourani, L., Tueller, S., Kizakevich, P., Bryant, S., Weimer, B., & Strange, L. (2015). Relaxation training assisted by heart rate variability biofeedback: Implication for a military predeployment stress inoculation protocol. *Psychophysiology*, *52*(9), 1167–74. doi: 10.1111/psyp.12455

Lieberman, H. R., Bathalon, G. P., Falco, C. M., Morgan, C. A., Niro, P. J., & Tharion, W. J. (2005). The fog of war: decrements in cognitive performance and mood associated with combat-like stress. *Aviation, Space, and Environmental Medicine*, *76*(7), C7–C14.

Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: the autonomic response to task demands. *Biological psychology*, 113, 83-90.

Malik, M. (1996). Heart Rate Variability. *Annals of Noninvasive Electrocardiology*, *1*(2), 151–181. doi: 10.1111/j.1542-474X.1996.tb00275.x

Matthews, G., Szalma, J., & Panganiban, A. R., Neubauer, C., & Warm, J. S. (2013). Profiling task stress with the dundee stress state questionnaire. *Psychology of stress: New research*, 49-90.

Matthews, G., Campbell, S., Falconer, S., Joyner, L., Huggins, J., Gilliland, K., Grier, R., & Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*(4), 315. doi: 10.1037/1528-3542.2.4.315

Matthews, G., & Davies, D. R. (2001). Individual differences in energetic arousal and sustained attention: A dual-task study. *Personality and individual Differences*, *31*(4), 575-589.

Matthews, G., Joyner, L., Gilliland, K., Campbell, S. E., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire: Towards a state big three. *Personality psychology in Europe*, *7*, 335-350.

Matthews, G., & Zeidner, M. (2012). Individual differences in attentional networks: Trait and state correlates of the ANT. *Personality and Individual Differences*, *53*(5), 574–579. doi: 10.1016/j.paid.2012.04.034

McClernon, C., McCauley, M., O'Connor, P., & Warm, J. (2011). Stress training improves performance during a stressful flight. *Human factors*, *53*(3), 207–18. doi: 10.1177/0018720811405317

McEwen, B. S. (2001). Plasticity of the hippocampus: adaptation to chronic stress and allostatic load. *Annals of the New York Academy of Sciences*, *933*, 265–77. doi: 10.1111/j.1749-6632.2001.tb05830.x

McEwen, B. S., & Wingfield, J. C. (2003). The concept of allostasis in biology and biomedicine. *Hormones and behavior*, *43*(1), 2–15. doi: 10.1016/S0018-506X(02)00024-7

Meichenbaum, D. (2017). Stress inoculation training. *The Evolution of Cognitive Behavior Therapy: A Personal and Professional Journey with Don Meichenbaum*, 101.

Meichenbaum, D. (1985). *Stress Inoculation Training*. New York: Pergamon Press.

Meichenbaum, D., & Cameron, R. (1989). Stress Inoculation Training. *Stress Reduction and Prevention*, 115–154. New York: Springer.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews of the Cambridge Philosophical Society*, *82*(4), 591–605. doi: 10.1111/j.1469-185X.2007.00027.x

Niskanen, J. P., Tarvainen, M. P., Ranta-Aho, P. O., & Karjalainen, P. A. (2004). Software for advanced HRV analysis. *Computer methods and programs in biomedicine*, 76(1), 73-81. doi: 10.1016/j.cmpb.2004.03.004

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, *33*(1), 17-33.
Orasanu, J., & Backer, P. (1996). Stress and military performance. *Stress and human performance*, 89–125. Mahwah, NJ: Lawrence Erlbaum.

Perna, F. M., Antoni, M. H., Baum, A., Gordon, P., & Schneiderman, N. (2003). Cognitive behavioral stress management effects on injury and illness among competitive athletes: a randomized clinical trial. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, *25*(1), 66–73. doi: 10.1207/S15324796ABM2501_09

Porges, S. (1995). Cardiac vagal tone: A physiological index of stress. *Neuroscience & Biobehavioral Reviews*, *19*(2), 225–233. doi: 10.1016/0149-7634(94)00066-A

Porges, S. (2011). The Polyvagal Theory: Neurophysiological Foundations of Emotions, Attachment, Communication, and Self-regulation (Norton Series on Interpersonal Neurobiology). New York: *WW Norton & Company*.

Regehr, C., Glancy, D., & Pitts, A. (2013). Interventions to reduce stress in university students: a review and meta-analysis. *Journal of affective disorders*, *148*(1), 1–11. doi: 10.1016/j.jad.2012.11.026

Robson, S., & Manacapilli, T. (2014). *Enhancing Performance Under Stress: Stress Inoculation Training for Battlefield Airmen*. Santa Monica, CA: The Rand Corporation.

Rose, R. D., Buckey, J. C., Zbozinek, T. D., Motivala, S. J., Glenn, D. E., Cartreine, J. A., & Craske, M. G. (2013). A randomized controlled trial of a self-guided, multimedia, stress management and resilience training program. *Behaviour research and therapy*, *51*(2), 106–12. doi: 10.1016/j.brat.2012.11.003

Saunders, T., Driskell, J. E., Johnston, J., & Salas, E. (1996). The effect of stress inoculation training on anxiety and performance. *Journal of occupational health psychology*, *1*(2), 170–186. doi: 10.1037/1076-8998.1.2.170

Schwerdtfeger, A., & Derakshan, N. (2010). The time line of threat processing and vagal withdrawal in response to a self-threatening stressor in cognitive avoidant copers: Evidence for vigilance-avoidance theory. *Psychophysiology*, *47*(4), 786–795. doi: 10.1111/j.1469-8986.2010.00965.x

Serino, S., Triberti, S., Villani, D., Cipresso, P., Gaggioli, A., & Riva, G. (2014). Toward a validation of cyber-interventions for stress disorders based on stress inoculation training: a systematic review. *Virtual Reality*, *18*(1), 73–87. doi: 10.1007/s10055-013-0237-6

Shaffer, F., McCraty, R., & Zerr, C. L. (2014). A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in psychology*, 5. doi: 10.3389/fpsyg.2014.01040

Staal, M. A. (2004). *Stress, cognition, and human performance: A literature review and conceptual framework* (NASA Tech. Memorandum 212824). Moffett Field, CA: NASA Ames Research Center.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. *Neuroscience and biobehavioral reviews*, *36*(4), 1228–48. doi: 10.1016/j.neubiorev.2012.02.003

Summers, R. L., Johnston, S. L., Marshburn, T. H., & Williams, D. R. (2005). Emergencies in space. *Annals of emergency medicine*, *46*(2), 177-184. doi: 10.1016/j.annemergmed.2005.02.010

Tarvainen, M. P., & Niskanen, J. P. (2012). Kubios HRV. *Finland: Biosignal Analysis and Medical Imaging Group (BSAMIG), Department of Applied Physics, University of Eastern Finland.*

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, *36*(2), 747-756.

Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009). Heart rate variability, prefrontal neural function, and cognitive performance: the neurovisceral integration perspective on self-regulation, adaptation, and health. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine, 37*(2), 141–53. doi: 10.1007/s12160-009-9101-z

United States. National Aeronautics and Space Administration (NASA). (2013). *International Space Station, Emergency Procedures 1a: Depress, Fire, Equipment Retrieval* (No. JSC-48566). Houston, TX: NASA Johnson Space Center.

Woodman, G. F., & Luck, S. J. (2004). Visual search is slowed when visuospatial working memory is occupied. *Psychonomic bulletin & review, 11*(2), 269-274.

Zhang, L., & Han, K. (2009). How to Analyze Change from Baseline: Absolute or Percentage Change. *D-level Essay in Statistics.* Dalarna University.

Appendix. Approval for Research (IRB)

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
1138 Pearson Hall
Ames, Iowa 50011-2207
515 294-4566
FAX 515 294-4267

Date: 5/8/2015

To: Tor Finseth
1200 Howe Hall

CC: Dr. Nir Keren
102 I Ed II

From: Office for Responsible Research

Title: Stress Inoculation Technique: The effect of exposure to presence of low-level hazardous conditions during full-scale simulated operations on performance and stress during simulated high-level hazardous conditions

IRB ID: 15-144

Approval Date:	5/7/2015	**Date for Continuing Review:**	5/6/2017
Submission Type:	New	**Review Type:**	Expedited

CHAPTER 5. DESINGING TRAINING SCENARIOS FOR STRESSFUL SPACEFLIGHT EMERGENCY PROCEDURES

Tor Finseth[1], Michael C. Dorneich[2], Nir Keren[3], Warren D. Franke[4], & Stephen Vardeman[2]

Modified from a manuscript published in *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)* (Finseth et al., 2020).

Statement of Authorship

As the lead author, I developed the theoretical background, designed and conducted the experiment, and performed the analysis. Dr. Dorneich was involved in detailed supervision. Dr. Dorneich, Dr. Keren, Dr. Franke, and Dr. Vardeman helped supervise the project and interpret the results. I wrote the manuscript with review and input from all authors.

Abstract

Graduated stress exposure aims to alleviate the negative effects of stress on task performance during high-stress conditions. Skills are practiced in increasing stress conditions that approximate the operational environment. Practice continues until stress resilience and task proficiency are achieved. The use of virtual reality (VR) for inducing a stress response has increased in popularity in recent years. The ability to simulate operational tasks could create training based on graduated stress exposure. However, more research is needed to verify that stress levels can be effectively manipulated in the virtual environment during training, and that the VR training task accurately replicates the existing task procedure. The objective of this study was to investigate the creation of different VR stressor levels from existing emergency spaceflight procedures and validate three distinguishable stressor levels (i.e., low, medium,

[1] Dept. of Aerospace Engineering, Iowa State University, Ames, IA, 50011, USA
[2] Dept. of Industrial Manufacturing and Systems Engineering, Iowa State University, Ames, IA, 50011, USA
[3] Dept. of Agricultural and Biosystems Engineering , Iowa State University, Ames, IA, 50011, USA
[4] Dept. of Kinesiology, Iowa State University, Ames, IA, 50011, USA

high). Experts in spaceflight procedures and the human stress response helped design a VR spaceflight environment and emergency fire task procedure. A within-subject experiment was conducted using the three stressor levels. Sixty-one healthy participants completed three trials in VR to locate and extinguish a fire on the International Space Station (VR-ISS). Self-assessment was implemented for each stressor level; NASA Task load index, Post Task Stress Reaction scale, Free stress scale, Positive and Negative Affect Scale, and Short Stress State Questionnaire were used for assessment. The results suggest that the stressors can induce different, distinguishable, levels of stress in trainees for use in graduated stress exposure training.

Introduction

The spaceflight environment has a multitude of hazards that can evolve into life-threatening emergencies (e.g., fire, depressurization, or toxic contaminate leaks). Astronauts practice emergency procedures by increasing the complexity of scenarios over time until they can reliably execute the task. However, intense acute stress can erode skills developed in prior training and diminish performance in-flight, jeopardizing the lives of the crew.

Stress is defined as a transactional process in which environmental demands tax or exceed the resources available by an individual resulting in psychological, physiological, behavioral, or social outcomes (Lazarus & Folkman, 1984). When environmental demands exceed the mental capacity of an individual, particularly under hazardous spaceflight emergencies, highly aroused stress states can lead to impaired decision-making capabilities, decreased situational awareness, and distress, which could reduce performance in the operational environment (Healey & Picard, 2005).

Graduated stress exposure aims to alleviate the negative effects of stress on task performance in high-stress conditions (Finseth et al., 2018). Skills are practiced in gradually increasing stress conditions that approximate the operational environment until stress resilience

and task proficiency are achieved. Psychotherapies such a Stress Inoculation Training (SIT) and Exposure Therapy have relied on gradually increasing stress exposure as a treatment intervention for individuals suffering from aliments like post-traumatic stress disorder or anxiety.

Some researchers have aimed to enhance task performance outside of a clinical setting using Stress Exposure Training (SET; Johnston & Cannon-Bowers, 1996). SET is a preventative stress training for healthy individuals encountering operational environment stressors when task performance is crucial. SET is commonly administered by trainers rather than clinical psychologists and used for complex cognitive tasks outside the original clinical domain (Johnston & Cannon-Bowers, 1996; Thompson & McCreary, 2006). Driskell, Johnston, and Salas (2001) found that SET improved performance and reduced subjective stress in laboratory studies following exposure to novel stressors and novel tasks. Graduated stress exposure is built into the SET and SIT framework to practice stress coping skills and build resilience through repeated exposure to stressful situations.

Completion of multiple sessions of stress exposure using graduated stress exposure may familiarize users with the stress encountered in emergencies, while promoting competency and control skills. Since high fidelity training environments often require expensive mock-ups or training facilities, VR may offer a safe and controlled environment to practice task related skills. Several studies have attempted to conduct graduated stress training in VR, following the 10-session preventative SIT framework, by giving military participants presentations on stress education and coping, and then conducting VR exposure sessions with breathing training (Ilnicki, Wiederhold, & Maciolek, 2011; Kosinska et al., 2013; Maciolek et al., 2013). In the short term, the soldiers were able to reduce their arousal at the conclusion of 10 sessions (Kosinska et al., 2013), however, upon returning from a 19-month military deployment, none of

the soldiers showed long-term inoculation (Maciolek et al., 2013). While these results suggest potential for short term inoculation, the small sample sizes leaves room for future work with larger samples. Hourani et al. (2011) as well as Winslow et al. (2016) conducted experiments using increasing levels of stressors. However, their training was a modified form of graduated stress exposure with the goal of building resilience across multiple stressors and tasks, making it difficult to deduce the benefit in a single operational environment. These studies demonstrate that while SIT and SET are promoted as frameworks to improve stress for operational tasks, there may be shortcomings in validating the stressors and relating the training task to real life. In order to train with stress exposure, a reliable way of identifying and manipulating levels of stress in individuals is needed.

To identify stressors relevant to a spaceflight emergency and measure the manipulation effect, insight can be gained from standardized stress tests which have been used by researchers to reliably manipulate stress levels precisely and consistently in a laboratory setting. Several of stress tests include public speaking, mental arithmetic, and cold-pressor (Kirschbaum, Pirke, & Hellhammer, 1993; Plarre et al., 2011). These stress tests aim to use the most potent stressors possible to elicit a general physiological stress response, often unrelated to the environment the individual is in. However, the stressor used for task training should be stressors that could occur in the operational environment (Johnston & Cannon-Bowers, 1996). These stressors could be related to the environment (e.g. distractions), the task (e.g., increased difficulty), or that state of the human (e.g. fatigue). To identify stressors that are relevant to the operational environment, where the environment is more dynamic and harder to anticipate stressors, some researchers have used expert opinion to identify prominent stressors and measured the manipulation using

subjective stress ratings and physiological indices of stress (Dorneich et al., 2008; Healey & Picard, 2005).

The objective of this study was to assess the extent to which operationally relevant VR stressor levels (i.e., low, medium, high) derived from existing emergency spaceflight procedures could evoke a reliable stress response. A panel of experts in spaceflight procedures and human stress response identified relevant stressors and created an experimental emergency procedure. The stressors and procedure were then used to design the VR-ISS spaceflight environment. Participant were trained with the procedures and conducted a VR-ISS fire response for three stressor levels. Self-assessments of stress, mood, and workload were administered for each level. This work contributes exploring how virtual environments can be designed for graduated stress exposure training and how spaceflight stressors can induce different level of stress in trainees.

Environment Stressor Design

To inform how training simulations in VR should be designed, a workshop was held with a panel of subject matter experts to identify stressors that contribute to the stress in a spaceflight environment. Attending experts included a retired U.S. NASA astronaut, a retired NASA International Space Station (ISS) flight director, and experts on psychological and physiological stress. The panel was presented with existing ISS emergency fire procedures and layouts of the ISS (United States, 2013). The panel identified a series of potential stressors, how they are mapped onto the emergency response procedures, and what effects the stressors might have on astronauts. Each stressor's intensity was rated on how the subject was likely to feel, and what the effects on performance will be. In addition, it was assumed that certain stressors would increase stress levels without influencing workload, so the stressors were categorized into task-related and non-task-related (i.e., environment-related) stressors. In other words, the stressors during an

emergency fire were categorized by their ability to change the environment without substantially changing how the individual would perform the procedure.

A list of stressors was compiled by the panel (Table). The stressors were categorized by the type of stress induction, manipulation type (environment or task manipulation), within or between experimental comparison. Stressor levels and potential deviations from the selected emergency procedure were also listed. Three environmental stressors were selected, where there intensity was varied among training scenarios: alarms, flickering lights, and visibility from smoke. Priority was given to stressor manipulations that did not affect tasks requirements, to avoid later confounds in experimental design. The goal was to increase stress while keeping the task requirements the same, in order to allow direct comparisons of stress training with and without stressor manipulation. By using only environmental stressors and not task stressors, any changes from such a comparison would not be confounded by differences in task requirements. For example, intensity of noise and visible smoke may be stressful, but will not change the task procedure for locating and extinguishing a fire. The selected stressors were then used to establish three stressor levels of low, medium, and high during the VR training scenarios. A few of the task stressors (e.g., rising atmospheric contaminants) were includes in all the training scenarios to distribute task load equally.

Since a change in procedure will affect task load and consequently stress levels, the panel then developed a simplified fire procedure for trainees to follow. Figure illustrates a flowchart of the selected emergency procedure that could be used for laboratory experiments. This procedure was modified from the existing ISS Emergency Procedures by shortening the duration of the procedure down to 5-10 minutes, eliminating communication with Mission Control Center

Table 1: Stressors Identified from ISS Emergency Fire Procedure

Stressor	How to Manipulate	Type of stress induction	Intensity Scale	Levels	Manipulation (Task/Envir.)	Within/btw training session	Task Implications
Fire alarm	Magnitude	Divided Attention	Linear		Environ.	Within	Shut off alarm
	Noise type	Divided Attention	Binary	On/off	Environ.	Between	Shut off alarm
Warning alarm	Magnitude	Divided Attention	Linear		Environ.	Within	
	Noise type	Divided Attention	Binary	On/off	Environ.	Between	
Caution alarm	Magnitude	Divided Attention	Linear		Environ.	Within	
	Noise type	Divided Attention	Binary	On/off	Environ.	Between	
Lights	Flashing	Divided Attention	Linear		Environ.	Within	Fix or turn off lights
Visibility (smoke obscurity)	Magnitude	Task Difficulty, Time pressure	Linear	0 - no visibility	Task, Environ.	Within	Don emergency mask, safe to enter?
	Rate of Change	Task Difficulty, Time pressure	Linear	0 - fast rate	Task, Environ.	Within	Don emergency mask, safe to enter?
Multiple sources (of fire)	Multiple sources in module(s)	Concurrent Task Mgmt.	Binary	On/off	Task	Between	Complex fireport sampling plan
Location of smoke/fire	Source in capsule	Task Difficulty	Binary	On/off	Task	Between	Cannot retreat
	Source separating crew	Task Difficulty, Moral Dilemma, Difficult tradeoff	Binary	On/off	Task	Between	Cannot contain environment
	Source in Module(s)	Task Difficulty	Binary	On/off	Task	Between	Multiple sampling plans required
Open Flame	Magnitude	Physical threat, Task Difficulty	Linear		Task	Within	Don Emergency Mask, forgo air readings, extinguish fire
	Spread Rate	Physical threat, Task Difficulty	Linear	0 - fast rate	Task	Within	Don Emergency Mask, forgo air readings, extinguish fire
Oxygen/Emer Mask	Limited oxygen supply	Time Pressure	Linear	2 min - infinite min	Task	Within	Find new mask
	No oxygen, faulty	Task Difficulty	Binary	On/off	Task	Between	Find new mask
	Reduced peripheral vision	Task Difficulty	Linear	0 - no visibility	Task	Between	Difficulty noticing threat cues, seeing CSA-CP readings, remove mask more frequently
	Clarity of other crew's voices	Task Difficulty	Linear		Task	Within	Difficulty in accomplishing fireport sampling plan
Contaminants	Magnitude	Threat, Task Difficulty	Linear		Task	Within	Don emergency mask
	Rate	Time Pressure	Linear		Task	Within	Don emergency mask
Multiple electrical trips	Multiple sources	Concurrent Task Mgmt.	Binary	On/off	Task	Between	Complex fireport sampling plan
Team Member	Language, accent	Task Difficulty	Linear		Task	Between	Difficulty in accomplishing fireport sampling plan
	Experience of crew	Task Difficulty	Linear		Task	Between	Crew can help with fire sampling plan and source locating
	Location of crew	Task Difficulty	Linear		Task	Between	Cannot contain environment
	Sleeping crew	Task Difficulty	Binary	On/off	Task	Between	Wake sleeping crew
	Reaction of crew	Task Difficulty	Linear	calm - panic	Task	Within	More energy spent calming/tracking crew whereabouts, higher risk of crew injury
Communication with MCC	Comm Delay	Task Difficulty	Linear		Task	Between	Delayed instructions/help from MCC
	No Comm	Task Difficulty	Binary	On/off	Task	Between	No MCC, crew self-reliance
	No lights (power outage)	Concurrent Task Mgmt.	Binary	On/off	Task	Within	Determine cause of power outage, retrieve flashlight
Power Outage	Compromised Life Support	Concurrent Task Mgmt.	Binary	On/off	Task	Between	Choice between turning on power or abandoning station
MCC	Frequency of communication	Concurrent Task Mgmt.	Linear		Task	Between	More time communicating with MCC, longer fire procedure
	Inconsistent instructions	Task Difficulty	Linear		Task	Between	Possible changes to fire sampling plan, changes in procedure branches
Fire Extinguishers	Limited PFEs available	Task Difficulty	Linear		Task	Between	More time spent finding PFEs
	Limited PFE uses	Task Difficulty	Linear		Task	Between	

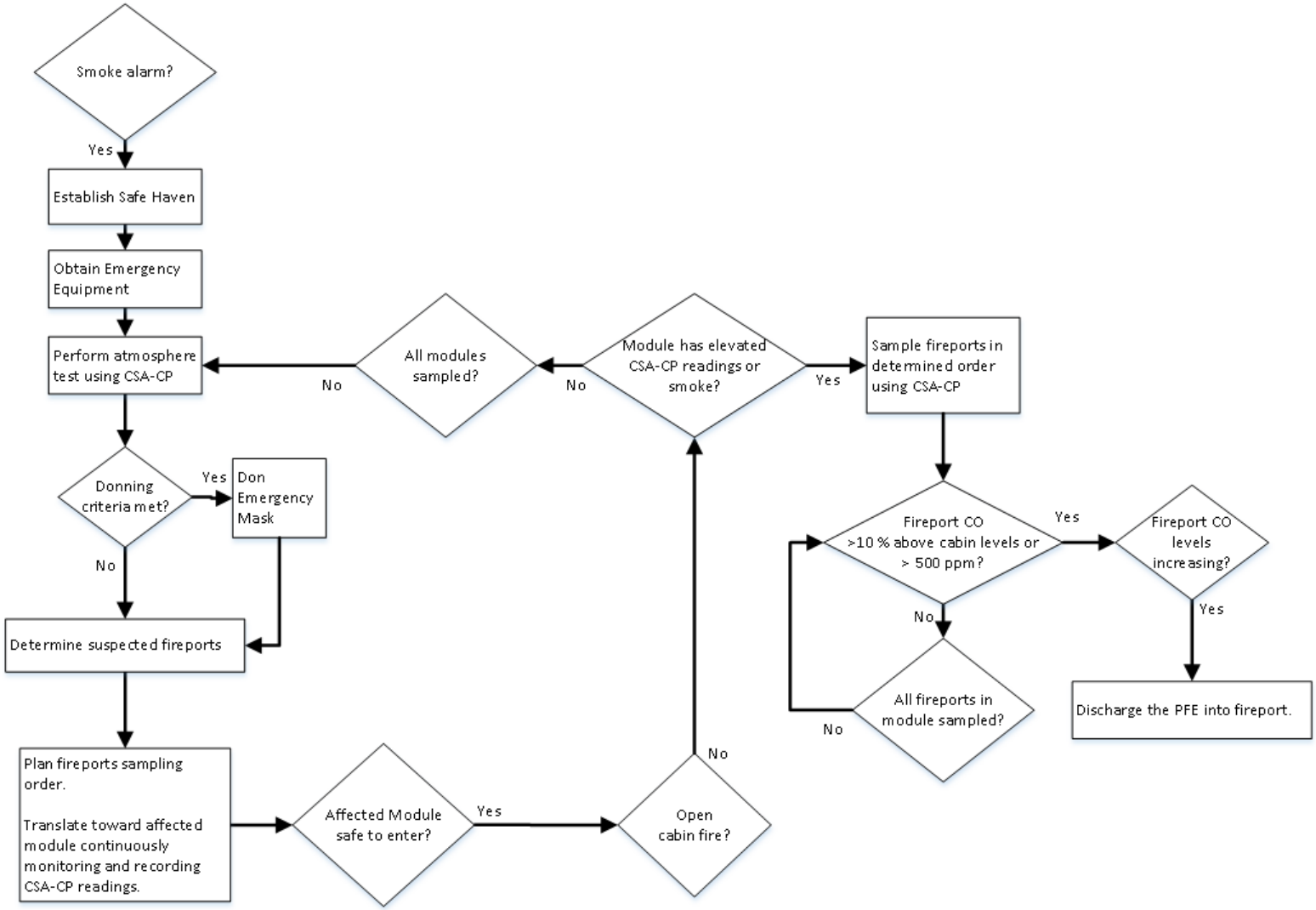

Figure 1: VR-ISS emergency fire procedure steps created by the workshop, modified from existing NASA ISS Emergency Fire Procedures.

(MCC) and crewmember induced stressors, and minimizing tangential procedure steps that do not directly help the trainee locate the fire source.

Methods and Materials

Participants

Sixty-two subjects participated (Male=47, Female=15). The study was reviewed and approved by Iowa State University Institutional Review Board (see Appendix). It is important to note that participant demographics differ from these of astronauts. Participant mean age was 20.6 years (SD = 2.6), compared to an average astronaut candidate of 40 years old (Kovacs & Shadden, 2017). The gender ratio of participants also differs from recent astronaut candidate cohorts of 50% male and 50% female. NASA requires all astronaut applicants to have STEM related education; 54 of the 61 participants in this study had a STEM backgrounds. However, it was determined that the initial training design may benefit from a boarder recruitment pool of stress appraisals found in a general population.

Experimental Design

A within-subject 1 x 3 (trial) experiment was conducted. Each participant completed the same task of locating an onboard fire, but each trial had one of three different stressor levels (low, medium, and high). The order of stressor levels was assigned via Latin square.

Task Environment

VR-ISS is the operational environment for this experiment. The VR-ISS environment is based on IGOAL (2017) and Finseth et al. (2018), but has been largely modified from its former state to be used with VR head mounted display (HMD) and facilitate spaceflight procedure training. The emergency response procedure to a fire on the VR-ISS was established based on existing NASA Emergency Procedures (United States, 2013). To simplify training participants, the VR-ISS consisted only of three of the existing U.S. Orbital Segment modules, Figure 2

illustrates the VR-ISS configuration used in the experiment herein: Only Node 1, US Lab, and Node 2 were used.

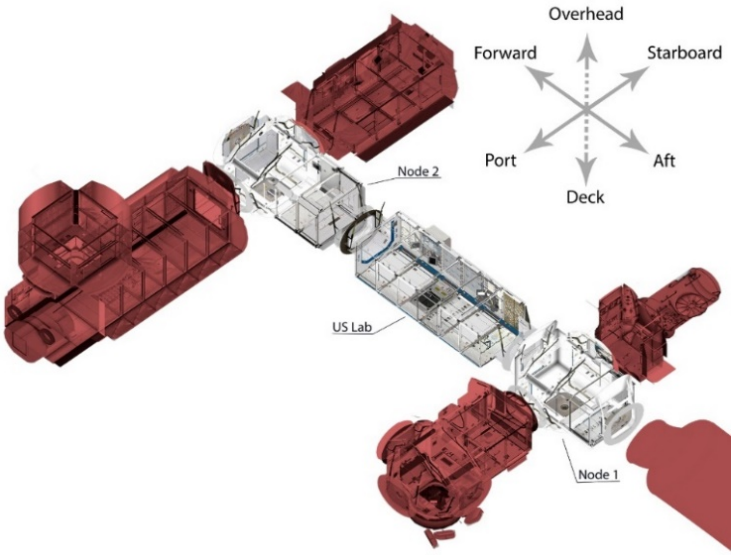

Figure 2: VR-ISS configuration. Some sections of the ISS U.S. and Russian segment were not included in the simulation.

Participants were tasked with locating and extinguishing the location of a potential fire on the VR-ISS. Several dynamic interactions were included in the VR-ISS to aid detecting and locating the source of a fire. Atmospheric contaminant levels rose as a function of time and distance from the virtual fire source. Virtual smoke changed in density as a function of time and spread in a uniform pattern, consistent with expected smoke behavior in a microgravity; therefore, participants could not rely solely on visual smoke patterns to detect the location of the source.

Participants reviewed readings of contaminant levels with a simulated NASA Compound Specific Analyzer–Combustion Products (CSA-CP). The purpose of the CSA-CP onboard the ISS is to determine the level of atmospheric contaminants. Virtual CSA-CP displayed levels of oxygen, carbon monoxide (CO), hydrogen chloride (HCl), and hydrogen cyanide (HCN) in parts per million (Figure 3). Using voice commands, a floating CSA-CP appeared in front of the

participant with the contaminant concentration values visible. The window disappeared after three seconds. Participants are expected to identify the location of the source by following the invisible path mentally established from recalling highest levels of contaminants in each VR-ISS module. Participants were instructed to retrieve a Portable Fire Extinguisher (PFE) and Portable Breathing Apparatus (PBA) when the contaminate levels are excessive (Figure 3). The PFE is used to extinguish a fire source behind a rack fireport. The PFE has the capability for two uses before the canister is empty. Five PFEs are available in cabinets in the VR-ISS. PBAs are available in the same cabinets and can be done on the participant avatar's head. A Caution and Warning (C&W) panel displayed flashing lights to alert participants to a potential fire (Figure 3).

Figure 3: VR-ISS emergency fire equipment, (A) Compound Specific Analyzer–Combustion Products, (B) Portable Fire Extinguisher, (C) Portable Breathing Apparatus, and (D) Caution & Warning Panel.

Once the participants identified where the fire source was located, they began sampling fireports within the module to locate the "rack" that caused the fire. The VR-ISS included fireport labels, accurately placed on the racks throughout the ISS. When participants identified the fireport which had the highest level of contaminant and extinguished the fire with the PFE, the contaminate levels were reset and a new randomized fire source was created. The task ended five minutes after the beginning of the simulation.

Independent Variables

The simulation had three different stress levels, each with a fire location randomized (Figure 4). The low stress level indicated a fire using increased CSA-CP contaminate values and C&W panel lights. The medium stress level indicated a fire using increased CSA-CP contaminate values, C&W panel lights, a continuous Caution alarm, and low levels of smoke (visibility of 6 ft.). The high stress level indicated a fire using increased CSA-CP contaminate values, C&W panel lights, a continuous Caution alarm, a continuous Fire alarm, flickering lights, and high levels of smoke the spread over time (visibility of 1 ft.).

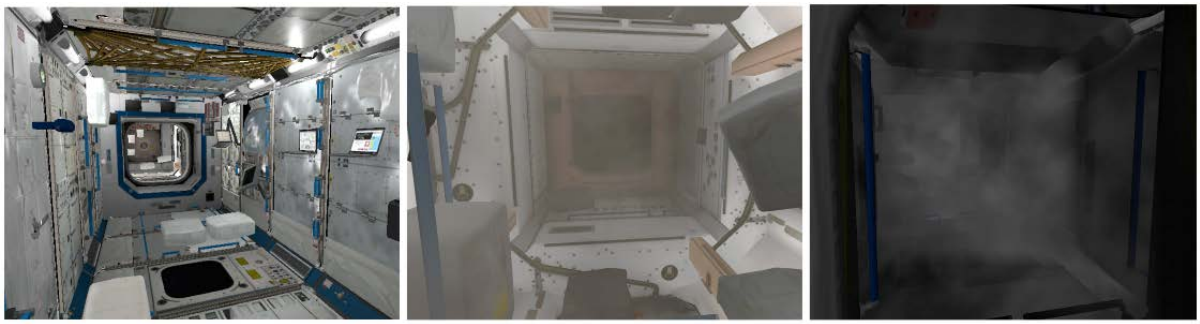

Figure 4: VR-ISS emergency fire (low, medium, high stressor scenarios).

Dependent Variables

Questionnaires were used to measure different psychological effects.

Subjective stress

Three kinds of ratings were used to assess the perception of low, medium, and high stress: Post Task Stress Reaction (PSTR), Free stress scale of events, and Short Stress State Questionnaire (SSSQ). The PSTR and Free stress scale questionnaires were used by Healey and Picard (2005; Singh, Conjeti, & Banerjee, 2013; Sharma & Gedeon, 2014). The PSTR asks participants to rate the ground truth simulations on a scale of “1” to “9” where a rating of “1” was used to represent experiencing “no stress”, “5” was used to represent “medium stress”, and

"9" was used to represent experiencing "high stress." The PSTR is intended to measure the immediate retrospective stress after completing a trial. The Free stress scale has participants rate the relative stress level on a scale of 0 to 100 (least to most stressful) in comparison to other simulations. The stress appraisal process is continuous and relative, with reappraisals of the experience happening long after the stressful exposure (Carpenter, 2016). The free stress scale was intended to measure the relative retrospective reappraisal after completing all the simulations.

The Short Stress State Questionnaire (SSSQ; Helton, 2004) assessed the subjective states pre- and post-trial to measure three state factors: task engagement, distress, and worry. Engagement refers to qualities of energetic arousal, motivation, and concentration. Distress is defined as feelings of tense arousal, hedonic tone, and confidence-control. Worry relates to self-focus, self-esteem, and cognitive interference (Matthews et al., 1999). The stress state acts as a mediator between the stressor and cognition or information processing, whereby the three aspects represent components of conscious experience during person-task-environment transactions (Helton & Näswall, 2015).

Workload

The NASA Taskload Index (TLX; Hart & Staveland, 1988) was used to assess the subjective workload during exposure. The NASA TLX measures six dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration level. NASA TLX was administered after the completion of a trial. Participant scores on the six numerical rating scales were computed in the 0 to 100 range and as an unweighted participant mean for each of the six-dimensional subscales (Nygren, 1991).

Mood

The Positive and negative affect scale (PANAS; Watson, Clark, & Tellegen, 1988) was given after each trial to assess the effects of the stressor on state affect in response to the task. The PANAS consists of 10 items for positive affect and 10 items for negative affect. Items are rated on a five-point scale ranging from 1= “very slightly or not at all,” to 5= “extremely”. The ratings were averaged to create overall scores for positive affect and negative affect.

Procedure

The experiment was completed in a single laboratory visit, lasting approximately 60 minutes. At the beginning of the experiment, participants completed a series of pre-trial questionnaires including demographic questions, the PSS-10 (to evaluate their initial stress condition when entering the lab), a SSSQ to measure the stress in response immediately before the trials, and training on how to use the NASA TLX. To acclimate to VR before the data collection tasks, participants were trained on navigating, operating and controlling the VR simulation (e.g., head mounted display, hand controls, "play-area" boundaries represented but a visual blue-grid). Participants were asked to report cybersickness.

For the VR-ISS, participants completed a VR interactive tutorial that included information about the ISS layout, how to navigate, fire equipment, and the appropriate emergency fire response. Participants practiced the procedure in the tutorial until memorized.

Participants then completed three trials: low, medium, and high stressor levels. After each trial, participants completed several questionnaires including the post-trial SSSQ, NASA TLX, PANAS, and PSTR. Participants were given 5-10 minutes between trials to complete questionnaires. After completing the three trials, participants completed the Free stress scale too.

Experiment Materials

The apparatus consisted of two parts: an HTC VIVE (professional version; HTC, 2016) consumer VR headset. The Unity (5.4.0f3, Unity Technologies, 2014) 3D game engine was used to facilitate all aspects of the VR-ISS as a virtual environment. The HTC VIVE setup consists of the HMD and two Lighthouse sensors that are responsible for tracking the headset position and orientation. For this experiment, the lighthouse sensors were positioned facing each other at opposite ends of our lab space, 8 ft high with 12x12 ft detectable play area.

Data Analysis

Data analysis was performed using SPSS software (Version 23.0; IBM Corp.). Distributions were tested for normality using skewness and kurtosis divided by the standard error and concluded to be normal if less than 1.96 (Kim, 2013). For comparison of questionnaires, repeated measure analysis of variance (RM-ANOVA) was used to calculate the fixed effect of stressor level. Significant differences were located using pair wise comparisons, and acceptance level was adjusted to control for type I errors (Bonferroni adjustment). Results were considered significant for $p \leq 0.05$. Cohen's *d* was used for assessing effect size, where r $0.2 < |d| < 0.5$ considered small effect size, medium effect size when $0.5 < |d| < 0.8$, and large effect size for $|d| > 0.8$ (Cohen, 1988).

The three factor SSSQ scale scores for pre- and post-trial were calculated for each participant. The factor scores from both pre- and post-trial are standardized against normative means and standard deviation values from a large sample of British participants (Matthews et al., 2002) and standardized using methods in Helton and Näswall (2015; Helton, Matthews, & Warm, 2009). Change scores were calculated for each factor using the z-score formula (1) which

has been used in previous studies (Helton & Näswall, 2015). The z-score then represents the change between pre- and post-trial in units of the deviation from the population mean.

$$z = (\text{standardized post-score} - \text{standardized pre-score}) \quad (1)$$

Results

Subjective Stress

The main effect of stressor level on PSTR was significant, $F(2,90) = 28$, $p < 0.001$, $d = 1.58$ (Figure 5). Pairwise comparison indicated the PSTR was significantly higher for participants in high stressor compared to low stressor ($p < 0.001$), significantly higher for high stressor compared to medium stressor ($p < 0.001$), and significantly higher for medium stressor compared to low stressor ($p = 0.026$).

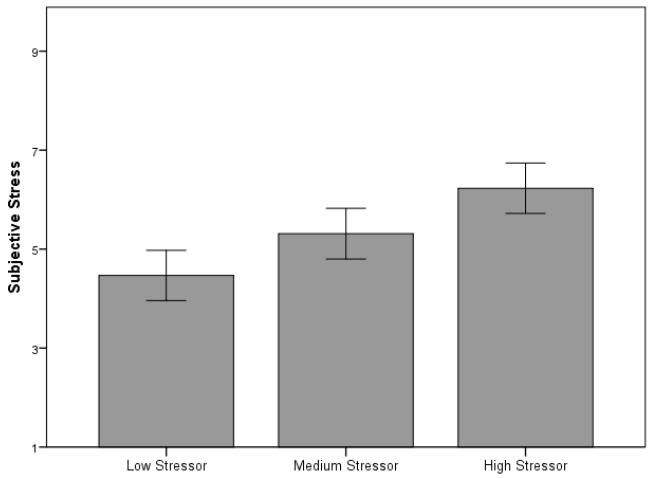

Figure 5: Subjective stress for different levels of stressors obtained by Post Task Stress Reaction (PTSR). Error bars representing 95% confidence intervals.

The main effect of stressor level on free stress was significant, $F(2,90) = 102$, $p < 0.001$, $d = 3.02$ (*Figure* 6). Pairwise comparison indicated the free stress was significantly higher for participants in high stressor compared to low stressor ($p < 0.001$), significantly higher for high stressor compared to medium stressor ($p < 0.001$); and significantly higher for medium stressor compared to low stressor ($p < 0.001$).

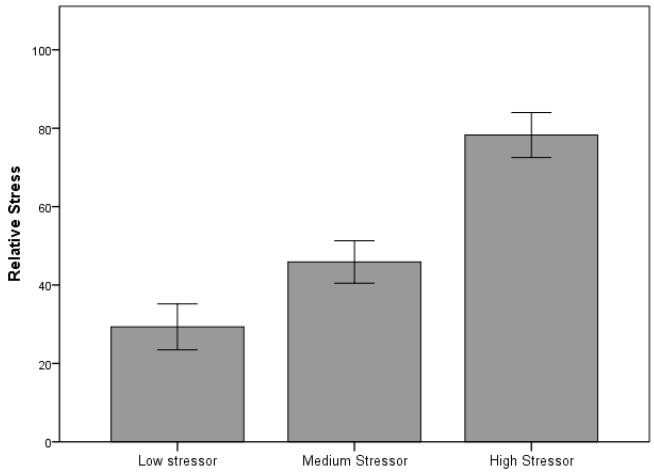

Figure 6: Relative stress for different levels of stressors obtained by Free Stress Scale. Error bars representing 95% confidence intervals.

A subset of the participants (N = 26) completed the SSSQ due to experiment alterations. The main effect of stressor level on distress was significant, $F(1.5,36.2) = 8.57$, $p = 0.002$, $d = 1.17$ (*Figure* 7). Pairwise comparison indicated the distress was significantly higher for participants in high stressor compared to low stressor ($p = 0.026$), significantly higher for high stressor compared to medium stressor ($p = 0.009$), but not significantly different for medium stressor compared to low stressor ($p = 0.62$). The main effect of stressor level on *engagement* and *worry*, along with the pairwise comparisons, were not significant.

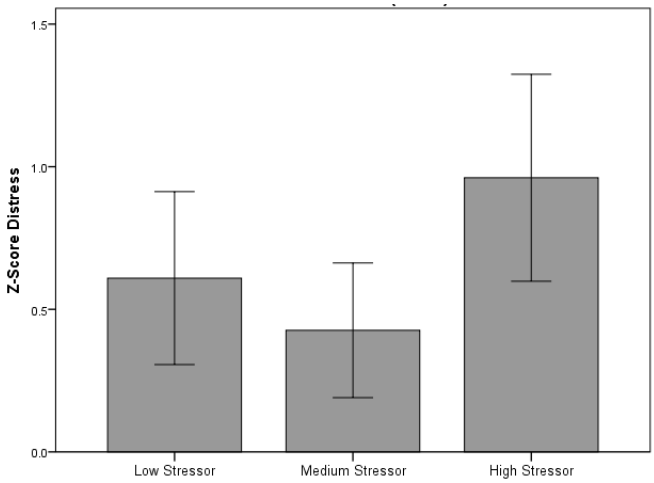

Figure 7: Pre-post change in distress for different levels of stressors obtained by Short Stress State Questionnaire (SSSQ). Error bars representing 95% confidence intervals.

Workload

The main effect of stressor level on workload was significant, $F(2,90) = 28$, $p < 0.001$, $d = 1.18$ (*Figure* 8). Pairwise comparison indicated the workload was significantly higher for participants in high stressor compared to low stressor ($p < 0.001$), significantly higher for high stressor compared to medium stressor ($p < 0.001$), but not significantly different for medium stressor compared to low stressor ($p = 0.929$).

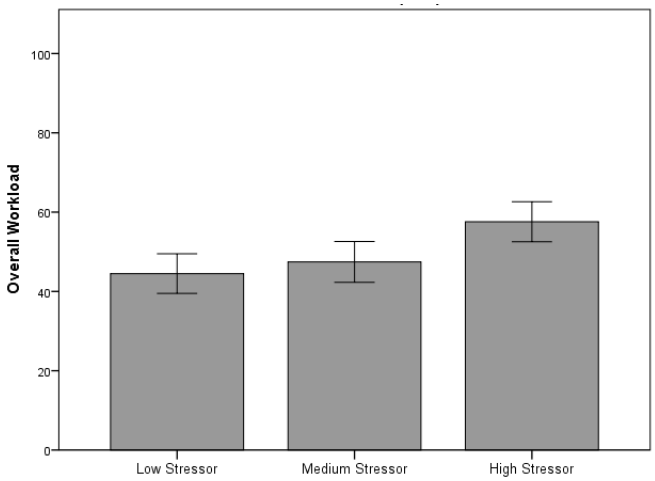

Figure 8: Overall workload for different levels of stressors obtained by NASA Task Load Index (TLX). Error bars representing 95% confidence intervals.

Within the TLX subscales, *mental workload* was significantly different, $F(1.6,74.9) = 5.98$, $p = 0.006$, $d = 0.73$, with the high stressor being significantly higher than the medium stressor ($p = 0.044$) and low stressor ($p = 0.021$). *Physical workload* was significantly different, $F(2,90) = 7.78$, $p = 0.001$, $d = 0.84$, with the high stressor being significantly higher than the low stressor ($p = 0.003$). *Temporal workload* was significantly different, $F(2,90) = 16.36$, $p < 0.001$, $d = 1.2$, with the high stressor being significantly higher than the low stressor ($p < 0.001$), high stressor being significantly higher than the medium stressor ($p = 0.02$), and medium stressor being significantly higher than the low stressor ($p = 0.008$). *Performance* was significantly different, $F(1.7,74.2) = 5.37$, $p = 0.009$, $d = 0.71$, with the high stressor being significantly

higher than the medium stressor ($p = 0.006$). *Effor*t was significantly different, $F(1.6,72.6) = 5.79$, $p = 0.008$, $d = 0.72$, with the high stressor being significantly higher than the medium stressor ($p < 0.001$) and low stressor ($p = 0.047$). *Frustration* was significantly different, $F(1.6,72.6) = 8.45$, $p = 0.001$, $d = 0.95$, with the high stressor being significantly higher than the medium stressor ($p = 0.003$) and low stressor ($p = 0.014$).

Mood

Only several of the participants (N = 24) completed the PANAS questionnaire due to experiment alterations. The main effect of stressor level on positive affect was not significant, $F(2,48) = 0.134$, $p = 0.86$ (*Figure* 9). The main effect of stressor level on negative affect was not significant, $F(2,46) = 2.55$, $p = 0.089$. None of the pairwise comparisons achieved significance.

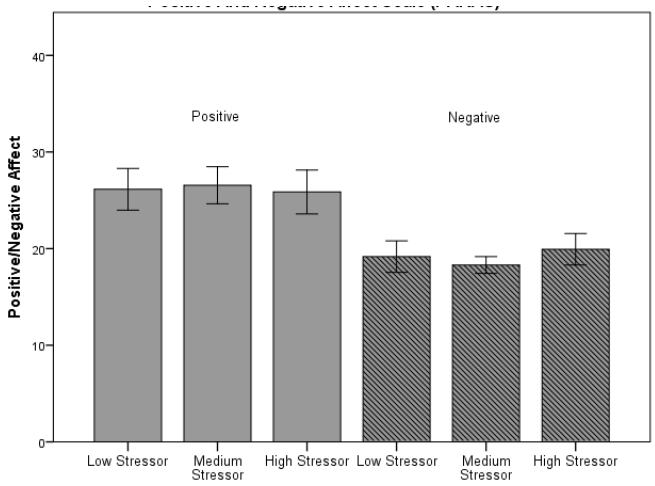

Figure 9: Positive and negative affect for different levels of stressors obtained by the Positive and Negative Affect Scale (PANAS). Error bars representing 95% confidence intervals.

Discussion

This study investigated whether manipulating different levels of VR stressors based on existing spaceflight procedures can induce different levels of stress on participants engaged in implementing an emergency response in a simulated VR-ISS. The results demonstrated that

levels of subjective stress and workload were significantly different for those training simulations.

The subjective stress was found to be different for the stressor levels based on immediate ratings after each trials (PSTR), ratings after the experiment (free stress scale), and changes in distress during the trials (SSSQ). The results suggest that design of the VR simulations using environmental stressors was successful at manipulating trainee stress levels. However, the negative affect and distress results show that low and medium were hard to distinguish between. The cause may be attributed to which stressors were selected for the simulation, the magnitude of each environmental stressors, or the combination of different stressors (e.g., noise, smoke). Previous research has shown that different stressors can elicit varying stress responses and can have a cumulative effect that may be greater than the individual effects of the stressors alone (Abdelall et al., 2020; Pedrotti et al., 2014). The present experiment used expert opinion to inform how the stressors were included in the training simulations, but had little empirical support. By changing the stressors, magnitude, or combination, it may be possible to have the VR stressor scenarios the show difference in affect and distress in future research. Nevertheless, the large effect of the stressor levels on subjective stress demonstrates distinguishable training levels that can be used to conduct graduated stress training.

The results show that the workload was higher for the high stressor compared to the other stressor levels. Because the simulations were designed to only change environment stressors, it was expected that workload would not change between the three stressor levels because the procedure was the same for each. Since workload can be thought of as a stress derived from the stressor of task load, it was predicted that the TLX rating for frustration and performance may be different, but not enough to impact the overall workload. However, it is unexpected to find a

difference for the high stressor level from the TLX mental workload, physical workload, temporal, and effort. There are two possible explanations: the high stressor level was unintentionally made with a higher task load, or the increased environmental stressors resulted in more stress, less resources for emotion regulation, and thereby, increases in the perceived demand of the workload. The latter explanation can be supported by previous research that found heightened stress reactivity and threat sensitivity when individuals in conditions with high amount of stress verses a lower threat sensitivity and reactivity in conditions with no stress (Akinola & Mendes, 2012). The implications are that the subjective workload may increase concurrently with increasing stress levels, even when task load stays constant.

This study had several limitations that should be considered with regards to interpretation of results and future work. First, the sample size for the PANAS and SSSQ was small and statistically underpowered. This might also explain why the SSSQ Distress for the low stressor was higher than expected. Alterations were made during the experiment to suspend the PANAS and SSSQ because the authors speculated that the duration and number of questionnaires were eliciting stress between trials, which could possibly confound perceived stress evoked during the trial. The time duration to complete the original questionnaires was 5-10 minutes, however, removing the PANAS and SSSQ halved the duration. The distress and negative affect for the high stressor VR scenario was trending toward being different than the medium stressor, but did not reach the level of significance. Running more participants would increase the statistical power and confidence in the results. Second, the study recruited participants from the general population rather than astronauts. The general population is less familiar with the ISS layout and procedures, which could possibly lead to stress or confusion from being trained in a short period of time. Further, astronauts may have stronger associations between threat cues and spaceflight

hazards, and simultaneously, possible development of coping skills to manage the threat appraisal. Future work will include evaluating the simulations with a participant sample similar to the age range, education level, and demographic of astronauts. Future investigation of other training effects would be beneficial and include memory consolidation/retention, task performance, and physiological habituation over multiple sessions. Objective measures of stress were measured for a subset of the participants in response to different stressor levels and will be analyzed in future work. By validating the stressor levels with subjective and objective measures, the simulation can then be used to aid development of a stress detection system to train individuals over multiple sessions and investigate the efficacy of graduated stress exposure in building resilience.

Conclusion

Graduated stress exposure has shown potential benefits for preparing individuals for stressful operational tasks. To develop training, it is important to identify of relevant stressors, create training simulations based on existing procedures, and measure that the stressor levels within the Results from this experiment show that the stressor levels have distinguishable subjective stress and workload. By validating stressor levels, graduated stress exposure may someday be used to train astronauts for emergency fires and other stressful spaceflight procedures.

Acknowledgements

This work was funded by the National Aeronautics and Space Administration [grant number 80NSSC18K1572]. The authors thank Clayton C. Anderson, Tomas GonzalezTorres, and Elizabeth Shirtcliff for providing expertise on spaceflight procedures and the human stress response. The authors also thank Robin Gillund, Silvia Verhofste, Matthew Kreul, and Kelly

Thompson for their laboratory assistance with research participants. For their help developing the VRISS and fire equipment models, the authors thank Pete Evans, Grant Leacox, Peter Carlson, and Robert Slezak.

References

Abdelall, E. S., Eagle, Z., Finseth, T., Mumani, A. A., Wang, Z., Dorneich, M. C., & Stone, R. T. (2020). The interaction between physical and psychosocial stressors. *Frontiers in behavioral neuroscience*, *14*.

Akinola, M., & Mendes, W. B. (2012). Stress-induced cortisol facilitates threat-related decision making among police officers. *Behavioral neuroscience*, *126*(1), 167.

Carpenter, R. (2016). A review of instruments on cognitive appraisal of stress. *Archives of Psychiatric Nursing*, *30*(2), 271-279.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* Hillsdale, NJ: Laurence Erlbaum Associates.

Dorneich, M. C., Mathan, S., Ververs, P. M., & Whitlow, S. D. (2008). Cognitive state estimation in mobile environments. *Augmented cognition: A practitioner's guide*, 75-111.

Driskell, J. E., Johnston, J. H., & Salas, E. (2001). Does stress training generalize to novel settings?. *Human factors*, *43*(1), 99-110.

Finseth, T. T., Keren, N., Dorneich, M. C., Franke, W. D., Anderson, C. C., & Shelley, M. C. (2018). Evaluating the Effectiveness of Graduated Stress Exposure in Virtual Spaceflight Hazard Training. *Journal of Cognitive Engineering and Decision Making*, *12*(4), 248-268.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, *6*(2), 156-166.

Helton, W. S. (2004, September). Validation of a short stress state questionnaire. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 11, pp. 1238-1242). Sage CA: Los Angeles, CA: SAGE Publications.

Helton, W. S., Matthews, G., & Warm, J. S. (2009). Stress state mediation between environmental variables and performance: The case of noise and vigilance. *Acta psychologica*, *130*(3), 204-213.

Helton, W. S., & Näswall, K. (2015). Short stress state questionnaire. *European Journal of Psychological Assessment*.

Hourani, L. L., Kizakevich, P. N., Hubal, R., Spira, J., Strange, L. B., Holiday, D. B., Bryant, S., & McLean, A. N. (2011). Predeployment stress inoculation training for primary prevention of combat-related stress disorders. *Journal of CyberTherapy & Rehabilitation 4*(1), 101-116.

IGOAL, NASA/JSC. (2017, March 27). *NASA 3D Resources*. **https://nasa3d.arc.nasa.gov/detail/iss-internal.**

Ilnicki, Wiederhold, & Maciolek. (2011). Effectiveness evaluation for short-term group pre-deployment VR computer-assisted stress inoculation training provided to Polish ISAF soldiers. *Annual Review of Cybertherapy and Telemedicine 2012*, 113.

Johnston, J. H., & Cannon-Bowers, J. A. (1996). Training for stress exposure. *Stress and human performance*, 223-256.

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, *38*(1), 52.

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76-81.

Kosinska, L., Ilnicki, S., Wiederhold, B. K., Maciolek, J., Szymanska, S., Opalko-Piotrkiewicz, E., Siatkowska, A., Ilnicki, P., Zbyszewski, M., & Glibowska, A. (2013). VR stress inoculation training results for Polish ISAF soldiers–A study of 4 cases. *New tools to enhance posttraumatic stress disorder diagnosis and treatment*, *108*, 149-60.

Kovacs, G. T., & Shadden, M. (2017). Analysis of age as a factor in NASA astronaut selection and career landmarks. *PloS one*, *12*(7), e0181381.

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer publishing company.

Maciolek, J., Ilnicki, S., Wiederhold, B. K., Kosinska, L., Szymanska, S., Zbyszewski, M., Opalko-Piotrkiewicz, E., Siatkowska, A., Glibowska, A., Borzetka, D., Ilnicki, P., Filarowska, M., Pleskacz, K., & Murawski, P. (2013). The influence of pre-deployment VR computer-assisted stress inoculation training on the anxiety level in the Polish ISAF soldiers. *New tools to enhance posttraumatic stress disorder diagnosis and treatment*, *108*, 161-1.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., Grier, R., & Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*(4), 315.

Matthews, G., Joyner, L., Gilliland, K., Campbell, S., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire: Towards a state big three. *Personality psychology in Europe*, *7*, 335-350.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human factors*, *33*(1), 17-33.

Pedrotti, M., Mirzaei, M. A., Tedesco, A., Chardonnet, J. R., Mérienne, F., Benedetto, S., & Baccino, T. (2014). Automatic stress classification with pupil diameter analysis. *International Journal of Human-Computer Interaction*, *30*(3), 220-236.

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., Al'Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., Smailagic, A., & Wittmers, L. E. (2011, April). Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks* (pp. 97-108). IEEE.

Sharma, N., & Gedeon, T. (2014). Modeling observer stress for typical real environments. *Expert Systems with Applications*, *41*(5), 2231-2238.

Singh, R. R., Conjeti, S., & Banerjee, R. (2013). A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, *8*(6), 740-754.

Thompson, M. M., & McCreary, D. R. (2006). *Enhancing mental readiness in military personnel*. Defence Research And Development Toronto (Canada).

United States. National Aeronautics and Space Administration (NASA). (2013). International Space Station, Emergency Procedures 1a: Depress, Fire, Equipment Retrieval (No. JSC-48566). Houston, TX: NASA Johnson Space Center.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, *54*(6), 1063.

Winslow, B. D., Chadderdon, G. L., Dechmerowski, S. J., Jones, D. L., Kalkstein, S., Greene, J. L., & Gehrman, P. (2016). Development and clinical evaluation of an mHealth application for stress management. *Frontiers in psychiatry*, *7*, 130.

Appendix. Approval for Research (IRB)

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 11/16/2018

To: Tor Finseth Michael Dorneich, Ph.D.

From: Office for Responsible Research

Title: **Developing and testing a stress gauge using virtual reality**

IRB ID: **18-432**

Submission Type: Initial Submission **Review Type**: Full Committee

Approval Date: 11/15/2018 **Date for Continuing Review: 11/14/2020**

CHAPTER 6. PHYSIOLOGICALLY BASED STRESS DETECTION FOR HAZARDOUS OPERATIONS USING APPROXIMATE BAYES ALGORITHM

Tor Finseth[1], Michael C. Dorneich[2], Stephen Vardeman[2], Nir Keren[3], Warren Franke[4]

Modified from a manuscript submitted to *Journal of Biomedical Informatics* (Finseth et al., 2021a).

Statement of Authorship

As the lead author on this paper, I developed the theoretical background, designed and conducted the experiment, and performed the analysis. Dr. Dorneich was involved in detailed supervision. Dr. Vardeman conceived the idea for the Approximate Bayes algorithm and wrote the classification methods section. Dr. Dorneich, Dr. Keren, Dr. Franke, Dr. Vardeman helped supervise the project and interpret the results. I was the primary writer of the manuscript with review and input from all authors.

Abstract

When training for hazardous operations, stress detection can be used to optimize task performance and build resilience. Stress detection systems use physiological signals that indicate stress to teach a machine-learning model how to predict the stress class of future data. Unfortunately, several challenges must be addressed before stress detection can be used for real-time monitoring. First, the subjective appraisal and individualized physiological response to stressors present challenges for creating generalizable and robust systems. Second, the time-series nature of physiological signals creates temporal correlations and exacerbate the limitations

[1] Dept. of Aerospace Engineering, Iowa State University, Ames, IA, 50011, USA
[2] Dept. of Industrial Manufacturing and Systems Engineering, Iowa State University, Ames, IA, 50011, USA
[3] Dept. of Agricultural and Biosystems Engineering , Iowa State University, Ames, IA, 50011, USA
[4] Dept. of Kinesiology, Iowa State University, Ames, IA, 50011, USA

of standard machine learning approaches to stress detection. Third, traditional machine learning algorithms make approximations of class conditional probabilities, while a more direct approximation may be ascertained using Bayes theorem. This study was designed to assess the extent to which an individualized stress detection system can be effective in classifying multiple levels of stress in a domain-relevant task. A Bayes classifier, known as Approximate Bayes (ABayes), was evaluated as a method of predicting stress. Healthy participants completed a task with three levels of stressors (low, medium, high), either a complex task in virtual reality (responding to a spaceflight emergency fire, n=27) or a simple laboratory-based task (N-back, n=14), while heart rate, blood pressure, electrodermal activity, and respiration were assessed. A machine learning pipeline was developed to collect sensor signals, extract features, select features, train machine learning models, and classify the three stress levels. Features were selected for each participant and different interval window sizes were compared. ABayes classification performance was compared to the traditional classifiers of support vector machine, decision tree, and random forest. The results demonstrate that three levels of stress can be classified by approximating Bayes theorem with time series intervals, and the ABayes model has comparable cross-validation accuracy to traditional classifiers and between tasks during validation testing. Results also demonstrate that wrapper feature selection can result in better accuracy, but also create biases.

Introduction

Despite extensive training in responding to an emergency, a person's response to an actual emergency can be negatively affected by the stressfulness of the situation. Stress can result in a cascade of physiological changes that may alter behavioral patterns, situational awareness, decision making, and cognitive resources (Driskell et al., 2008). An inability to cope

with the stress of a high-stress condition can decrease task performance and thereby risk mission failure, injury, or death (Barshi & Dempsey, 2016). Consequently, developing resiliency to this situational stress through improved training may lead to better outcomes. To that end, using real-time monitoring of a person's stress responses to customize the stressfulness of training scenarios may, in turn, lead to more appropriate handling of actual hazardous operation (Gjoreski et al., 2017; Zahabi & Razak, 2020).

For a number of reasons, stress detection using machine learning has been challenging. First, there are individual differences in the appraisal of, and physiological responses to, stressful situations. Numerous stress detection approaches have attempted to reduce technical complexity by generalizing their models to a broad population, or the "average" response (Gjoreski et al., 2017). However, the stress response to a unique situation is largely subjective, personalized stress detection models may be more robust to individual differences (Can, Arnrich, & Ersoy, 2019a; Akmandor & Jha, 2017). Second, the time series nature of physiological signals can be problematic. The physiological stress response has temporal correlations and feature correlations; these correlations may violate the machine learning assumption that the data are independently and identically distributed, thereby leading to biased results (Verleysen & François, 2005). Finally, an additional challenge is how the conditional probabilities of the subject's stress levels are estimated. Stress prediction models are reliant on traditional machine learning algorithms that make data-driven approximations. However, these approximations can have varying accuracy. A classifier based on the Bayes theorem is theoretically the optimal solution and will have the lowest probability of error (Tong & Koller, 1999).

To achieve real-time and continuous monitoring of stress levels, new approaches are needed to analyze time series for physiologically-based stress detection (Smets, Raedt, Van

Hoof, 2019). Real-time stress detection could enable closed-loop automation to either modify the training environments to better match the trainee's responses or better assess individual stress during staged or real operations (Jones & Dechmerowski, 2016). Further, an approximation of the Bayes Optimal Classifier may be a benchmark to compare traditional machine learning classifiers in future applications.

To address these challenges, the goal of this research is to assess the extent to which a time-series interval approach to stress detection can accurately detect participant stress based on the physiological responses to data collected during stressful situations. The present study consisted of evaluating classification performance of the system, as well as analyzing the impact of the choices made for physiological feature selection and data window sizes. A Bayes classifier, known as Approximate Bayes (ABayes), was developed to address the limitations of existing supervised machine learning methods and approximate optimal stress predictions within the stress detection system. Testing ABayes with unseen data may be a more accurate method of predicting stress.

The next section of this paper describes the theoretical background of the classification and evaluation approach. The approach is then validated using a holdout method to assess the ability of the approach to classify unseen data, and compared to other classification approaches. Cross-validation results are also presented for completeness.

Background

Stress detection is challenging due to the individual differences in the response to stress and the time-series nature of the physiological stress response. Stress detection systems rely on classifying physiological signals into multiple stress classes using machine learning. This section describes the physiological responses, stress detection research and physiological sensors,

approaches to classifying time series data, a deeper look at the challenges facing stress detection, and the current research approach.

Stress

The physiological stress response involves the interaction between the nervous system and the endocrine system that aims to maintain physiological integrity under changing environmental demands. The time course of the physiologic responses to stress varies by system and by the intensity and duration of the stressor; they are neither physiologically independent nor statistically orthogonal. After the psychological appraisal of a stressor, neural ganglia pathways are activated almost instantaneously to evoke very rapid responses via local neurotransmitters. For example, disinhibition of heart rate via vagal withdrawal occurs within milliseconds while a sympathetically-mediated increase in heart occurs after a few seconds (5-10 s; Shaffer et al., 2014). Sympathetic and sudomotor activity results in the opening of eccrine sweat glands on hands and feet, which occur about 1-5 seconds after stimuli (Posada-Quintero & Chon, 2020). On the other hand, the physiologic responses due to circulating chemicals take longer to manifest. Epinephrine is secreted from the adrenal medulla and range from milliseconds to minutes to exert their cardiovascular effects. These processes can act exclusively or in conjunction on target organs to potentiate (e.g., memory, muscle activation) or attenuate organ function (e.g., digestion, reproduction).

There is increasing support that the physiological systems activated are those best suited to cope with the type of stressor, rather than the prior theories that certain systems are activated if the stressor magnitude surpasses a threshold (Bowers et al., 2008). Therefore, the same stressor can differentially affect individuals via leading to the activation of varying physiological systems, with each system having individualized, and differing, times-scales to respond and

recovery. The different individual stress response and system time-scales present challenges in detecting and classifying levels of stress.

Stress detection

Stress detection, by means of classifying these physiological responses into levels of stress via machine learning, continues to evolve and is motivated by the potential utility of continuously monitoring stress levels in real-time (Smets, Raedt, Van Hoof, 2019; Reimer et al, 2017). Stress detection systems have been developed for drivers in semi-urban scenarios (Singh, Conjeti, & Banerjee, 2014; Healey & Picard, 2005), patients undergoing virtual reality therapy (Tartarisco et al., 2015), individuals in working environments (Betti et al., 2018), and people that need help managing daily stress (Sun et al., 2010; Hovsepian et al., 2015; Reimer et al., 2017; Martinez et al., 2017; Alexandratos, Bulut, & Jasinschi, 2014; Plarre et al., 2011).

These detection systems collect information about stress responses from either objective physiological sensors or subjective psychological metrics, in the form of independent variables called features, which they then use to classify the stress level. Commonly used sensors include electrodermal activity (EDA), electrocardiogram (ECG), electroencephalogram (EEG), respiration (RSP), skin temperature (ST), and blood volume pulse (BVP; Giannakakis et al., 2019). For an ECG signal, stress indices have been primarily inferred from changes in the time intervals between heartbeats, which measure Heart Rate Variability (HRV) using a time-domain, frequency-domain, or non-linear analysis. HRV metrics have been associated with sympathetic and parasympathetic activation. However, attempting to detect stress levels from signal amplitude only neglects the time series nature of physiological data. Physiological systems may be simultaneous and coupled (e.g., breathing can modulate heart rate), may contain both deterministic and stochastic components, and may be correlated when measured over long

periods of time (Novak et al., 1993). Stress sensor signals are continuous ordered attributes; therefore, they are best characterized by features that quantify the distribution of data points, variation, correlation properties, stationarity, entropy, and nonlinear properties (Fulcher, 2018).

Approaches to time series classification

To address the time series nature of physiological signals, common time series classification methods include a) comparing whole series data by employing distance-based algorithms like Dynamic Time Warping (DTW), b) performing high-level feature extraction from successive, sequential time intervals and classifying intervals with a model, c) judging the presence or absence of short patterns (i.e., shapelets) in the whole series, d) frequency counts of recurring patterns to form a "dictionary" that defines the classes, e) combinations of the aforementioned methods, or f) model-based learning methods like those relying on auto-regressive models or hidden Markov models (Bagnall et al., 2017). Each of these methods has advantages and disadvantages with regard to physiological stress detection.

DTW is highly effective with a nearest-neighbor classifier for time series data, such as repeated patterns in ECG due to heart arrhythmias and sleep apnea. However, there are few examples of it being applied to stress detection (Zarei & Asl, 2020). This is likely due to acute stress having temporal and pattern variation, which make it difficult for whole series, shapelets, or dictionary methods to be effective. Model-based learning methods, like hidden Markov models, fit multiple models to the data in order to determine the best model to use. This type of framework has been seldom studied for physiological stress detection unless paired with stress speech analysis (Giannakakis et al., 2019). Interval characteristics is the most common classification method for stress detection. It is implemented by extracting features for windows/epochs, which is a highly reliable analytical method for quantifying stress through

features like HRV (Pourmohammadi & Maleki, 2020; Shaffer & Ginsberg, 2017). Normally a time window size is predetermined. However, some features may behave differently depending on the window size. For example, HRV frequency features are recommended to have windows in the order of minutes and smaller time intervals may increase error (Shaffer & Ginsberg, 2017). An evaluation of window sizes could help identify which features work best for interval methods.

Neural networks have become a popular classifier choice for interval methods, due to highly accurate frameworks such as convolutional or recurrent neural networks (Saeed et al., 2017). The success of a neural net is partly due to its ability to handle unequal time series lengths and optimize model parameters over time (Hidasi & Gáspár-Papanek, 2011). However, neural networks simultaneously extract features and many of the classification rules are created by the model rather than programmers. These classification rules can be hidden within interconnected layers (Smets, Raedt, & Van Hoof, 2019). The net effect is that the logic used in the classifications is often implicit and uninterpretable. For this reason, traditional machine learning models that classify interval features from a time-series are more informative and interpretable as to how data points are assigned to classes. Interval classification often uses supervised learning, where classification models are trained using interval features separated into classes/states (e.g., low, medium, high stress levels) and the model is subsequently used to make future predictions on a test dataset's class/state probability. Traditional supervised machine learning algorithms include support vector machine (SVM), decision tree, and random forest.

Challenges of physiological stress classification

A major challenge in using physiologic signals to detect stress is the uncertainty in the model's ability to predict an individual's stress level. Stress responses vary among individuals

and depend on the individuals' appraisals of the stressor; therefore, individualized models may be more accurate than generalized classifiers (Smets, Raedt, & Van Hoof, 2019; Gjoreski et al., 2017). For example, an EDA-based generalized classifier that is deployed and tested on multiple people may have higher classification error among a subset of this group, since as much as 25% of the population are EDA non-responders or hypo-responders (Braithwaite et al., 2013). Consequently, an individualized classifier may result in higher accuracy than the generalized, if the model accounts for the individual's respective EDA signal. Supervised classifiers can be individualized by having the stress detection system create a model using training data from the individual and by selecting relevant features for the individual.

Another challenge is that supervised classifiers have a degree of uncertainty depending on how they estimate probability distributions in order to make stress level predictions. Supervised models produce a probability distribution for each stress level (class) for a set of physiological signal data points (vectors); this distribution determines which class is most probable at a given time. However, rather than creating a distribution directly from the dataset, the probability distribution is created indirectly (and often ad hoc) based on the technical specifics of a classification method. For example, decision tree classifiers produce rectangles that partition the input space and calculate the approximate class probabilities based on the number of vectors located within each rectangle. Thus, the class probability is constant for each rectangle and always discontinuous at the rectangle boundaries, leading to a probability that is more defined by how the rectangles are positioned within the input-space rather than the vector distribution across the entire input-space. Similarly, SVMs create a hyper-planes intended to produce maximum separation between class vectors in the input space. Ad hoc "approximate class probabilities" are often created using softmax functions of distances from vectors to

hyperplanes—a practice that may not match empirical probability estimates (Zadrozny & Elkan, 2002). The process by which these ad hoc methods approximate class probabilities does not easily translate to meaningful cause/effect insights related to either changes in the environment or the measured changes in physiological measurements.

Bayes theorem provides more direct estimations of conditional probabilities. Thus, it would be useful to compare the aforementioned traditional classifiers to a classifier that uses Bayes Theorem. This can be done by implementing Bayes theorem in a new approximately Bayes classifier. To that end, the purpose the present research is to assess the extent to which an Approximate Bayes (ABayes) classifier and class probability estimation methodology can overcome the limitations of existing supervised machine learning methods when used for stress detection in systems that will need to classify on unseen data. It is hypothesized that since Bayes theorem results in the optimal probability, a classifier based on Bayes theorem would show greater classification accuracy than the traditional machine learning methods (decision tree, support vector machine, and random forest classifiers) for time-series data when testing on unseen data.

Approach

This paper describes the development of a physiological-based stress detection system to classify acute stress using an ABayes classifier. Participant physiological signals were collected for three stressor levels during either a spaceflight emergency fire procedure on a VR International Space Station (VR-ISS; Finseth et al., 2018; Finseth et al., 2020) or on a well-validated and less-complex N-back mental workload task (Herff et al., 2014). An individualized machine-learning pipeline was developed using feature selection on intervals of the time-series data. Each participant had features selected at different interval window sizes, then those

personalized features trained the classifier model, and subsequently tested the classifier's predicative accuracy. ABayes classification performance was assessed using both validation techniques of holdout and cross-validation. The approach was also compared to decision tree, support vector machine, and random forest classifiers.

Methods

Participants

Forty-one healthy participants (83% male, 17% female) experienced a complex task in virtual reality (spaceflight emergency fire, N=27) or a laboratory-based task (N-back, N=14). The mean age was 20.9 years (*SD* = 6.5). The demographic distribution included 76% Caucasian, 12% Asian or Asian American, and 7% Hispanic or Latino. All study procedures were approved by the Iowa State University Institutional Review Board (see Appendix B).

Experimental Design

The evaluation had two types of tasks and three stressor levels within each task. Task was a between-subjects variable: participants either conducted a fire response task aboard a VR International Space Station (VR-ISS) or a computer-based N-back task. These tasks were selected since it is possible to facilitate varying degrees of task complexity. Stressor level was a within-subjects variable: each task consisted of three stressor levels (low, medium, and high), where trials were counterbalanced via Latin Squares.

One task, *VR-ISS,* is the virtual reality environment of the ISS specifically designed for participants to implement an emergency fire response procedure by locating and extinguishing a fire source (Finseth et al., 2018). The VR-ISS task is highly dynamic and in a complex environment with many stimuli and task steps. The task is based on existing NASA Emergency Procedures (NASA, 2013) but simplified to reduce the amount of needed training. A number of

dynamic interactions were included in the VR-ISS to aid detection and location of the source of the fire. To locate the fire, participants evaluated atmospheric contaminant levels; these levels changed as a function of time and distance from the fire source. The highest contaminant value would indicate the approximate location of the fire source. Thus, participants would have to monitor and recall the local contaminant levels. When needed, participants used virtual oxygen masks and fire extinguishers.

Stressor levels in the VR-ISS were created with a combination of environmental stressor intensities that were independent from the task procedure: smoke, alarm noise, and flickering module lights (Finseth et al., 2020). The low stressor level did not contain any stressors; therefore, a voice recording announced a fire situation at the beginning of the simulation. The medium stressor level included a continuous caution alarm, low smoke density (visibility limit of 6 ft.) and flashing lights in one of the three ISS modules. The high stressor level involved a continuous caution alarm, a continuous fire alarm, flickering lights in all modules, and dense smoke (visibility limit of 1 ft.). Figure presents smoke density in the VR-ISS for each stressor level. Prior research verified that the three stressor levels produced different levels of subjective stress (Finseth et al., 2020).

Figure 1: VR-ISS emergency fire (low, medium, high stressor scenarios).

The other task, *N-back task,* is presented with a sequence of colored squares on a computer screen; participants need to recall the location of the square that was shown *n* steps earlier in the sequence. The N-back task is a well-validated stressor (Herff et al., 2014) where low-complexity can be induced through manipulating the one primary stressor of working memory demand. Stress is manipulated by asking participants to recall 1-back (low-demand), 2-back (medium-demand), and 4-back (high-demand).The N-back task is a measure of working memory capacity that is associated with executive functioning which can affect physiological stress indices (Shields, Sazma, & Yonelinas, 2016).

Stress Manipulation Measures

In order to verify that the three stressor levels produced variable levels of stress, participants completed the Free Stress scale after the third trial. This scale was used to rate the subjective stress level on a scale of 0 to 100 (least to most stressful) (Healey & Picard, 2005; Singh, Conjeti, & Banerjee, 2013; Sharma & Gedeon, 2012). The stress appraisal process is continuous and relative, with reappraisals of the experience happening long after the stressful exposure (Carpenter, 2016). The Free Stress scale was intended to relatively measure the subjective stress by comparing all three trials at the same time.

Procedure

The experiment was completed in a single laboratory visit, lasting approximately 120 minutes (Figure 2). After participants gave written consent, they completed a demographic questionnaire. To acclimate to VR before the data collection tasks, participants were placed in the VIVE Virtual Reality Home simulation (HTC, 2016) to train them on navigating, operating and controlling the VR simulation. Participants were asked about cybersickness. Participants

were equipped with physiological sensors and a baseline recording was taken to verify sensors were working properly.

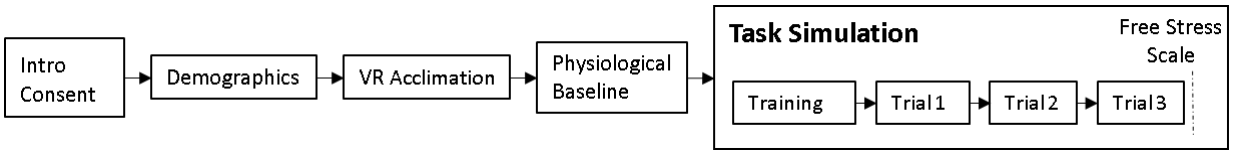

Figure 2: Design and procedure of the study.

For the task simulation (Figure 2), participants were then assigned to either the VR-ISS or N-back tasks. For the VR-ISS, participants completed a 20-30 minute VR tutorial that included information about the VR-ISS layout, how to navigate, fire equipment, and the emergency fire response. For N-back, participants completed a 3 minute tutorial on how to indicate if the current stimulus is the same as the one presented N trials ago. Participants then completed three trials: low, medium, and high stressor levels. Participants were given 5 minutes between trials to recover to physiological baseline. The Free Stress scale was given after the last trial.

Overview of the Stress Detection System

To aid in the development of the stress detection system, a machine learning pipeline was developed to detect and classify the three stress levels from the physiological measures and to evaluate ABayes as a classifier against other supervised classifiers. The pipeline consisted of several steps including data collection, preprocessing, feature extraction, feature selection, and classification as presented in Figure 3. The methods employed in each of these steps are described in sections 0 to 0. Data were collected through multiple sensors that measure physiological responses. A time-series classification approach was implemented by segmenting the data into multiple intervals and using summary measures as features. A feature extraction process was then used to find a high-level subset of features that may have class discrimination

with respect to a single individual, which was reduced into a low-dimensional feature subset (feature selection) by means of classification of a random holdout. A supervised approach was then taken to train the classifiers with the selected feature subset comprised of physiological data from three stress trials for each participant, investigated only as a subject-specific individualized model. Lastly, the classifiers were evaluated on their ability to predict participant stress levels.

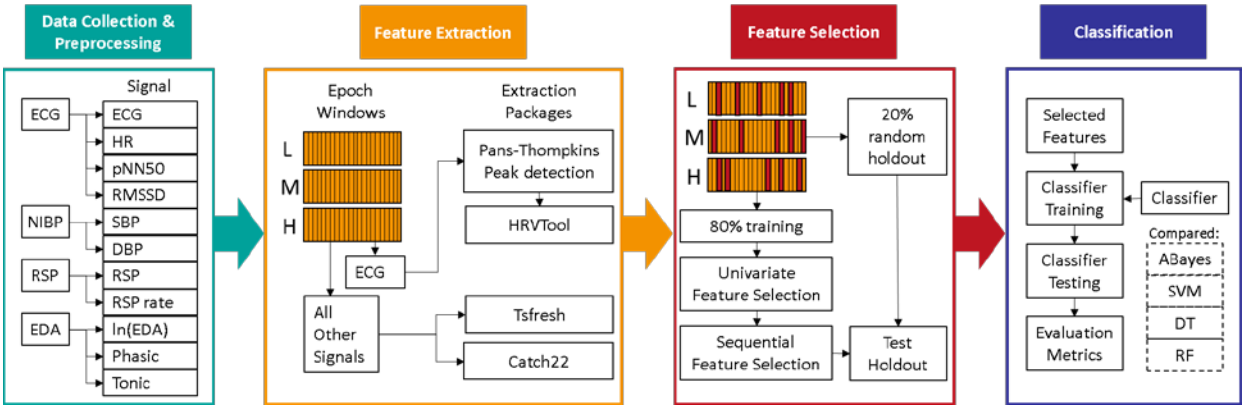

Figure 3: The machine learning pipeline of stress detection and classification.

Data Collection

Data were collected for the machine-learning pipeline using four physiological signals that were acquired simultaneously: electrocardiogram (ECG), Electrodermal Activity (EDA), Respiration (RSP), and Noninvasive Blood Pressure (NIBP).

Biopac's MP150 system (Biopac Systems Inc., Santa Barbara, CA) was used to measure ECG, and was equipped with an ECG100C module (Greene, Thapliyal, & Caban-Holt, 2016). The ECG signal was used to calculate Heart Rate and two time-domain Heart Rate Variability (HRV) signals of root mean squared of successive differences (RMSSD) and percent of peak-to-peak intervals exceeding 50 milliseconds (pNN50). Increasing values of RMSSD and pNN50 indicate relaxation (vagal activation) and decreasing values indicate arousal (vagal inhibition). Respiration was also measured as an indicator of autonomic nervous system (ANS) activity.

ECG and RSP were sampled using Biopac MP150 (125 Hz) and Bionomadix Bioshirt with the Heart Rate (HR), RMSSD, and pNN50 features extracted by the Acqknowledge software (Version 5.0.1, Biopac Systems Inc.) provided by the manufacturer.

Systolic blood pressure (SBP) and diastolic blood pressure (DBP) were collected as another measure of cardiovascular reactivity. DBP and SBP can reflect changes in the total peripheral resistance of blood vessels. Increases in local sympathetic activity cause constriction of blood vessels, while reductions in sympathetic activity lead to dilation. In the absence of changes in cardiac output, decreases in blood vessel constriction are usually reflected by decreases in DBP. In the present study, beat-to-beat blood pressure data were collected. An oscillometric noninvasive blood pressure (NIBP) fingercuff was placed on the participants' nondominant hand over the middle phalanx of the long and ring finger (CNAP Monitor 500, CNSystems Medizintechnik AG). The nondominant arm was placed in an arm sling to standardize the position of the hand relative to the heart between all participants. To calibrate the finger cuff, an NIBP cuff (CNAP Monitor 500, CNSystems Medizintechnik AG) was placed on the participant's dominant upper arm and measured periodically to minimize potential hydrostatic pressure differences between the fingers and heart level. NIBP was sampled with the Biopac MP150 at 125 Hz with the SBP and DBP features extracted within the Acqknowledge software.

Electrodermal activity (EDA) measures changes in electrical conductivity in the skin due to production of sweat by activation of the ANS. Increased arousal during stress will elicit higher EDA. EDA can be parsed into slower tonic-level and faster changing phasic-level components. Skin Conductance Level (SCL) is a measure of tonic EDA and reflects the general changes in autonomic activity. Skin Conductance Response (SCR) is discrete, short, phasic fluctuations that

reflect higher frequency variability of the signal as a response to immediate stimuli (Figner & Murphy, 2011). Electrodes were placed on the intermediate phalanges on the index and middle fingers of the non-dominant hand. EDA was sampled with the Biopac MP150 (125 Hz) with SCL and SCR features being extracted within the Acqknowledge software.

Preprocessing

As expected, the data contained different types of noise and artifacts associated with subject movement, power line, and electromagnetic interference. NIBP was corrected for motion artifacts using an IIR band pass filter with cut-off frequencies at 1 Hz and 10 Hz. The ECG signal was filtered for electrical noise by an internal 50-60 Hz notch filter. The EDA signal was corrected with an IIR low pass 2nd order Butterworth filter fixed at 5 Hz (Karthikeyan, Murugappan, & Yaacob, 2014; Bong, Murugappan, Yaacob, 2013; Zheng, Murugappan, & Yaacob, 2012). The EDA signal was decomposed into two components: phasic and tonic. The tonic component was extracted by low pass filtering with a cut-off frequency of 0.16 Hz, while the phasic component was extracted with a band-pass filter of 0.16 Hz and 2.1 Hz (Singh, Conjeti, & Banerjee, 2013). RMSSD and pNN50 features were extracted from the ECG signal. A smoothing window of 5-second averages was used on all derived features (e.g., HR, RMSSD, EDA, SBP, DBP). The sensor signals and features were saved at 125Hz. The first 60 seconds were deleted from each trial to remove effects of anticipatory stress and initial stress reactivity, which may result in weakly labeled time series data.

Feature Extraction

The feature extraction process was intended to improve information density by extracting a variety of features that characterize the time-series data. Features were chosen with the intent to characterize distribution of data points, variation, correlation properties, stationarity, entropy,

and nonlinear properties (Fulcher, 2018). Signals and features calculated by Biopac were binned into epoch windows from which other signals were extracted. Since the goal of this study was to build an automatic stress classification model with the potential to be applied to real-time applications, small window sizes were selected for evaluation: 10 sec, 20 sec, 30 sec, and 40 sec.

To extract features from the time series, two toolset packages were used: Tsfresh (Christ et al., 2018) and Catch22 (Lubba et al., 2019) for automatic feature extraction of time series characteristics, including absolute energy, absolute sum of changes, autocorrelation, entropy, and number of values above and below the average. Since ABayes is designed for probability density estimation in a real-time system, Tsfresh features were excluded if they were Boolean data types or were previously reported to take longer than 10^{-2} second to compute (see Christ et al., 2018). The Catch22 toolset is a set of 22 time series features from the much larger MATLAB toolbox, called *hctsa*, which has high accuracy in predicting different types of time series data (Lubba et al., 2019). The features extracted by Catch22 and Tsfresh do not overlap. From the ECG signal, the inter-beat interval (RR) signal was extracted via Pans-Tompkins peak detection (Sedghamiz, 2018). Time-series and spectral HRV features were then extracted from the RR signal via HRVTool (Vollmer, 2019). The final extracted features are listed in the Appendix A.

Feature Selection

Feature selection is the process of reducing the dimensionality of the classification problem by finding an optimal subset of available features that provide class discrimination. The best subset contains the fewest number of dimensions that most contribute to the classifier performance; the remaining, less contributing features are discarded (Mar et al., 2011).

To improve the robustness of the feature selection, a portion of the dataset was held out for validating the selection process. For the wrapper holdout, 20% of epochs were randomly

selected and stratified from each class. These were set aside to ensure that a test set of equal class sizes remained unseen while the other 80% was used to select features. A hybrid method of feature selection was conducted on the remaining 80% of data in a two-step process involving a univariate feature selection (UFS) filter method and sequential feature selection (SFS) wrapper method. First, UFS was used to find the best features for classification by quantifying their discriminative power using a univariate statistical test. The features were then ranked according to their mean one-way ANOVA F-value to prioritize features that explain large amounts of variance. In the second step, the SFS method started adding features from the UFS iteratively in a forward search to measure performance gain. SFS starts with the most discriminate feature identified by UFS and then adds features, one-by-one, according to their F-value rank and stopped after 15 iterations. A SVM classifier was employed as the objective function in a 10-fold cross-validation where misclassification rate was used as a criterion to opt for the best subset of features. Since the wrapper holdout was selected randomly and the datasets were small, this entire process was repeated six times and the final features were those that appeared in the optimal features sets multiple times. After the features were selected, the wrapper holdout was again included in the dataset to prepare for classification training and testing.

Classification

The ABayes classifier was formulated to address the indirect ways that standard machine learning algorithms typically estimate probability distributions across classes for given input vectors. When considering traditional machine learning classifiers, all standard classifier development and performance evaluation is implicitly or explicitly done using a probability model that generates class-conditional probabilities. To create a distribution of probability densities, the probability model says that for classes $k = 1, 2, \ldots, K$, observable data vectors $\mathbf{x}$ are

generated for each random choice of a class $y = k$ by using probability distributions. The probability distribution is specified by class probabilities $\pi_1, \pi_2, \ldots, \pi_K$. Then, data vectors $\mathbf{x}$ are generated for a given corresponding distribution of features specified by a class-conditional probability density $g_k(\mathbf{x})$. For a classifier $f(\mathbf{x})$ that maps observations to classes ($f(\mathbf{x}) \in \{1, 2, \ldots K\} \quad \forall \mathbf{x}$), the conditional probabilities can be calculated (Eq. 1).

$$P\left[f(\mathbf{x}) = y \right] = \sum_{k=1}^{K} \pi_k \int_{f(\mathbf{x})=k} g_k(\mathbf{x})\, d\mathbf{x} \tag{1}$$

In an optimal situation, the classifier would assume densities $g_k(\mathbf{x})$ and probabilities π_k are perfectly known, state k is classified with maximum $\pi_k g_k(\mathbf{x})$ (Eq. 2), which is equivalent to the maximum conditional probability of state/class k given the observation $\mathbf{x}$ (Eq. 3), which is also equivalent to the optimal classifier with minimum-error-rate, also known as Bayes Optimal Classifier (Eq. 4; Rish, 2001).

$$\underset{k}{\operatorname{argmax}}\ \pi_k g_k(\mathbf{x}) \tag{2}$$

$$\underset{k}{\operatorname{argmax}} \left(\frac{\pi_k g_k(\mathbf{x})}{\sum_{l=1}^{K} \pi_l g_l(\mathbf{x})} \right) \tag{3}$$

$$f^{\text{opt}}(\mathbf{x}) = \underset{k}{\operatorname{argmax}}\, P[\text{class is } k \mid \mathbf{x} \text{ is observed}] \tag{4}$$

However, standard machine learning classifiers are not optimal and are limited because the densities $g_k(\mathbf{x})$ are not known. Subsequently, some classifiers attempt to make approximations of the post-data weights π_k and densities $g_k(\mathbf{x})$ for each state/class, while other classifiers refrain from estimating the distributions entirely and attempt to approximate $f^{\text{opt}}(\mathbf{x})$. For example, in tree algorithms, the relative frequencies of the training set class/state in

rectangles serve as weight estimates of conditional probabilities of classes given that the input vector falls in given rectangles, whereas SVM directly learns a decision boundary without estimating data generating distributions. Even in Naive Bayes, the classifier estimates all marginal distributions and uses the product as a density, while making a generally poor assumption that the input vectors for each class have independent components (Rish, 2001). In these cases, machine learning classifiers use data-derived functions of **x** as a substitute for the optimal classifier. Incidentally, dataset sizes may not be in the same proportions as the π_k *s*, causing vague density distributions. Probability densities can be adjusted so that training set class relative frequencies match desired weights (π_k), but only using the training dataset. If adjustments are made to the training set class frequencies using knowledge of full dataset, this will result in data leakage where the model may be over-fit and overestimate the model performance when deployed (Samala et al., 2020). However, parametrizing to class frequencies is typically unrealizable in practice in that it depends upon the model weights π_k and densities $g_k(\mathbf{x})$. Therefore, no classifier can improve on this optimal classifier if the posterior distribution (i.e., densities) are known.

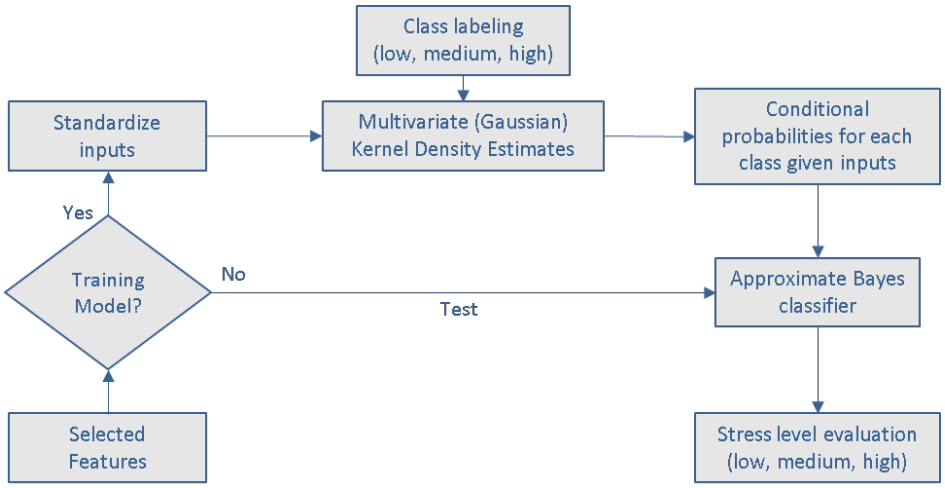

Figure 4: The Approximate Bayes for stress level classification.

The optimal classifier is a derived from of Bayes theorem, which provides a direct approximation of conditional class probabilities. Hence, the Approximate Bayes (ABayes) classifier is a statistical approach that attempts to optimally discriminate states/classes based on estimated conditional probabilities determined through a direct approximation for the K multivariate kernel density estimates (Figure 4). That is, training sets of multivariate observations $\mathbf{x}$ from classes k have been processed through a (Gaussian) kernel density estimation routine to produce functions $\hat{g}_k(\mathbf{x})$ approximating the class conditional densities $g_k(\mathbf{x})$. These are used to produce the Approximate Bayes classifier (Eq. 5), whereby an observation is classified to the class that gives it the largest estimated probability density. Recognizing the limitations of other standard machine learning classifiers, Bayes should result in the most accurate probability estimate, all things being equal.

$$\widehat{f^{\text{opt}}}(\mathbf{x}) = \underset{k}{\text{argmax}}\ \hat{g}_k(\mathbf{x}) \tag{5}$$

The input variables of the classifier are the extracted features from physiological sensors, standardized in the range of [0, 1]. An estimate is made for the probability density per class by applying multivariate (Gaussian) kernel density estimates. The training phase initializes weights based on assumed frequency of occurrence. The test phase is used to automatically classify an unknown input vector.

Data Analysis

The stress detection capabilities of the novel classifier ABayes and three common supervised machine learning classifiers were compared: support vector machine (SVM), decision tree (DT), and random forest (RF). The methods were compared using two validation methods: cross-validation and holdout. The classifiers were implemented with the MATLAB Statistics and

Machine Learning Toolbox. The performance of the classifiers was evaluated using two different validation techniques: Cross-Validation and Holdout (Figure 5). Cross-validation is included for completeness. Holdout tests the ability to classify on unseen data, and principle goal of the ABayes-based approach. Cross-validation was performed by 2-Folds and 10-Folds, with folds constructed consecutively in time. The holdout consisted of 20% of data from the end of each class being used as "unseen" testing data, without being used to select features. The cross-validation uses all of the participant's dataset for the feature extraction and selection, whereas the holdout technique only extracts and selects features from the first 80% of each trial's dataset. Therefore, the validation technique was implemented prior to the feature extraction within the proposed machine learning pipeline (*see* Figure 3). All data was standardized prior to classification per the physiological signal per individual, with the holdout data being standardized with respect to the means and standard deviations of the training data.

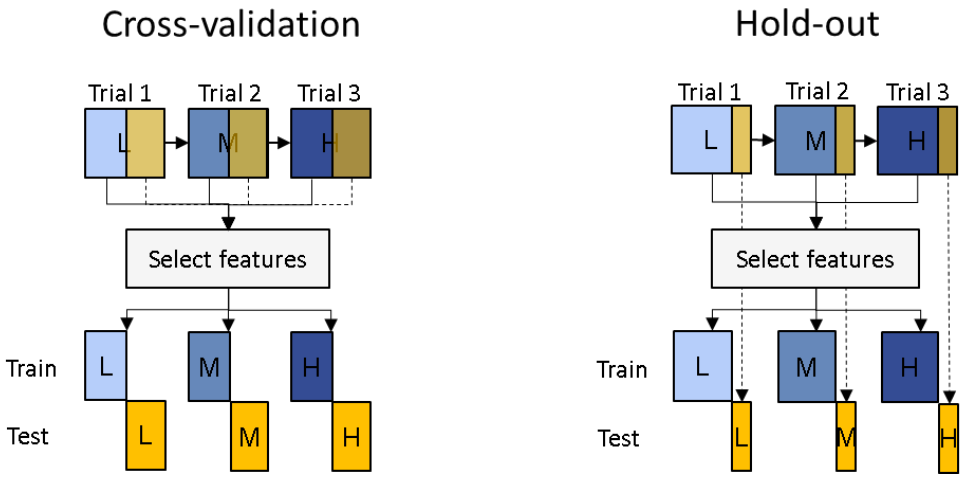

Figure 5: Examples of the feature extraction and selection for cross-validation and holdout. L, M, and H refer to the low, medium, and high stressor scenarios, respectively.

The evaluation process of classifiers involves calculating the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN; Plarre et al., 2011). Classifier performance was measured using accuracy, precision, recall, F1-score, and specificity.

Table provides the measures of multi-class classification and the corresponding equation and definition, where *k*=1,2,3,…*K* are the number of class and use macro-averaging techniques (Singh, Conjeti, & Banerjee, 2014).

Table 1: Performance measures for multiclass classification (Adapted from Singh, Conjeti, & Banerjee, 2014)

Measure	Formula	Evaluation focus
Accuracy	$\left(\sum_k^K \frac{\mathrm{TP}_k + \mathrm{TN}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k + \mathrm{TN}_k}\right)/K$	Overall effectiveness of getting a true result
Precision	$\left(\sum_k^K \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}\right)/K$	How often a positive prediction is correct
Sensitivity(Recall)	$\left(\sum_k^K \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}\right)/K$	How often a true positive is correctly predicted
F1-score	$\left(\sum_k^K 2\frac{\mathrm{precision}_k * \mathrm{sensitivity}_k}{\mathrm{precision}_k + \mathrm{sensitivity}_k}\right)/K$	Overall accuracy, but balancing precision and sensitivity
Specificity	$\left(\sum_k^K \frac{\mathrm{TN}_k}{\mathrm{FP}_k + \mathrm{TN}_k}\right)/K$	How often a negative prediction is correct

Accuracy is one of the main performance indicators and is defined as the number of correctly classified labels divided by the total number of labels. The Precision, Sensitivity, and F1-score reflect the importance of the retrieval of positive labels, while the Specificity reflects the correct classification of negative labels. F1-score is regarded as a more reliable classifier performance metric in comparison to accuracy in some circumstances, because the accuracy metric does not account for imbalanced class datasets (Chicco & Jurman, 2020).

The class size balance for each participant was evaluated with the imbalance ratio (He & Garcia, 2009) and likelihood ratio imbalance degree (LRID; Zhu et al., 2018) listed in Table 2. Imbalanced data can have harmful effects on classification and interpretation of results. The imbalance ratio (IR) is the most commonly adopted metric for class-imbalance extent, but it only for binomial datasets because it considers the ratio of the distribution of observations in the largest ($\hat{p}_{max}$) and smallest ($\hat{p}_{min}$) classes while ignoring information of other minority classes (Zhu et al., 2018). The frequency is estimated as the faction of observations in a given class (n_k) divided by total number of observations (*N*). An IR of one suggests and equal dataset. LRID

offers more resolution into multiple class distributions where the data may be overlapped or where there is ambiguity about the level of data separation. However, the score can vary by many magnitudes depending on the number of minority classes. Likelihood ratios can range from zero to infinity, with a score of zero for balanced data, while imbalanced data will result in a score larger than zero. Likelihoods were converted to probabilities (McGee, 2002).

Table 2: Imbalance measures for multiclass datasets

Measure	Formula	Evaluation focus
Imbalance Ratio (IR)	$\frac{\hat{p}_{max}}{\hat{p}_{min}}$	Overall effectiveness of a classifier
Likelihood ratio imbalance degree (LRID)	$-2\sum_{k=1}^{K} n_k \ln \frac{N}{Kn_k}$	Class agreement of the data labels with the positive labels given by the classifier

Data analysis on the subjective stress measure was performed using SPSS software (Version 23.0; IBM Corp.). Repeated measure analysis of variance (RM-ANOVA) was used to calculate the fixed effect of stressor level and pair wise comparisons that were adjusted to control for type I errors (Bonferroni adjustment). Results were considered significant for $p \leq 0.05$. Cohen's *d* was used for assessing effect size, where $0.2 < |d| < 0.5$ was considered a small effect size, medium effect size when $0.5 < |d| < 0.8$, and large effect size for $|d| > 0.8$ (Cohen, 1988).

Results

Subjective stress manipulation verification

The main effect of stressor level on subjective stress was significant for the VR-ISS, $F(2,90) = 102$, $p < .001$, $d = 3.02$. All pairwise comparisons indicated the subjective stress was significantly different ($p < .001$) between the stressor levels (see Fig. 6a). Similarly, the main effect of stressor level on subjective stress was significant for the N-back, $F(2,24) = 47.5$, $p < .001$, $d = 3.98$ (Fig. 6b). Pairwise comparisons indicated subjective stress was significantly

higher for participants in 4-Back compared to 1-Back (p <.001) and the 2-Back (p <.001). Subjective stress was significantly higher for 2-Back compared to 1-Back (p =.018).

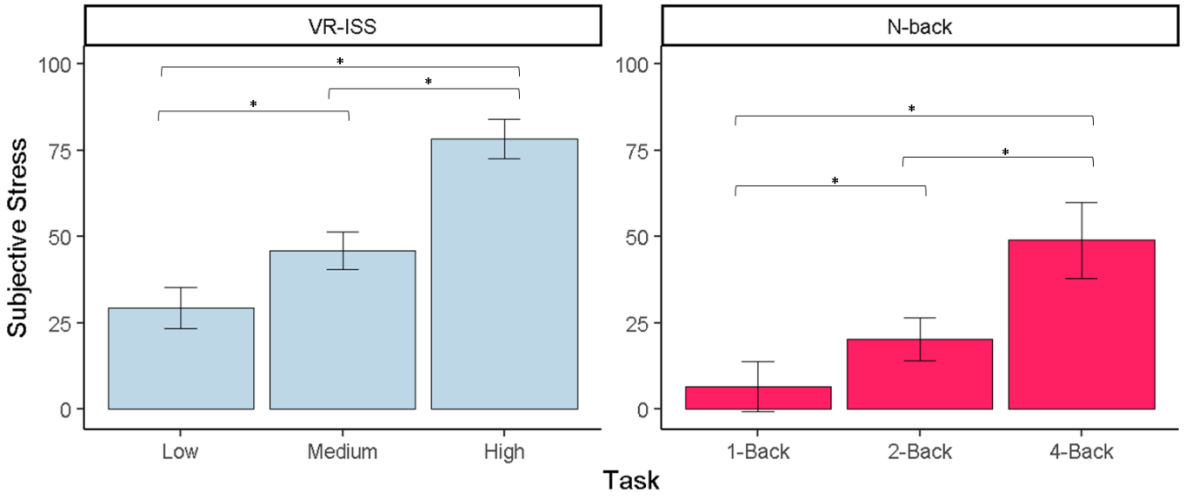

Figure 6: Subjective stress for different levels of stressors obtained by Free Stress Scale. Error bars represent 95% confidence intervals.

Machine Learning Results

The physiological data obtained from the VR-ISS and N-back were analyzed to provide insight into the features chosen by SFS, comparing the performance of ABayes between different tasks (VR-ISS, N-back), between different evaluation strategies (2-Fold, 10-Fold, holdout), and compared to the three standard machine learning classifiers. The characteristics for each task dataset (measured for 10 sec windows) are listed in Table 3.

Table 3: Details of the multiclass datasets, M (±SD).

Task	Total Size	Class Observations (L,M,H)	Imbalance Ratio	LRID
VR-ISS	76.3 (±41.6)	25.8 (±17.8), 25.5 (±19.3), 25.0 (±16.4)	2.37 (±2.39)	11 (±17.4)
N-back	62.9 (±17.0)	20.4 (±7.26), 20.5 (±6.0), 21.9 (±5.23)	1.52 (±1.71)	1.65 (±5.08)

When considering multiple classes, the LRID shows the VR-ISS is eleven times more likely to be imbalanced which is equivalent to 46% probability of being imbalanced. The N-back was only 1.65 times more likely to be imbalanced, which is 9.5% probability. Due to the

probability of imbalance, the F1-score is prioritized over the accuracy metric in the subsequent analyses (Chicco & Jurman, 2020).

Analysis of features selected over different window sizes

Before evaluating the performance with the validation techniques, the SFS output was evaluated based on different epoch window sizes: 10, 20, 30, 40 seconds. Table 4 lists the average amount of features selected by SFS for varying windows sizes and tasks, which shows the SFS had optimal performance when 4-5 features were selected on average.

Table 4: Number of features selected by SFS for each task.

	VR-ISS		N-back	
	Mean	**STE**	**Mean**	**STE**
Window				
10 sec	5.16	0.23	5.57	0.34
20 sec	4.21	0.26	5.29	0.51
30 sec	5.00	0.30	4.54	0.31
40 sec	4.41	0.28	4.25	0.33

The frequency for VR-ISS and N-back features selected by SFS for varying epoch window sizes is illustrated in Figure 7. Within the VR-ISS window sizes (10, 20, 30, 40 seconds), SBP mean (32%, 30%, 0%, 17%), SBP median (26%, 24%, 0%, 10%), and SBP power spectral density second coefficent (24%, 24%, 0%, 17%) were selected for the most participants. The 30-seond window deivated from the other windows with different features being selected. The most frequent feature in any window was HR augmented Dickey-Fuller (45%) followed by RSP periodicity measure (42%), which were in the 30-second window. Comparing the N-back window sizes, Heart Rate Aggreagate Partial Auto-Correlation (36%, 36%, 21%, 21%), RMSSD sum of reoccuring values (21%, 29%, 21%, 14%) were selected for the most participants. The most frequent feature in any window was Heart Rate Aggregate Partial Auto-Correlation (36%) which was in the 30-second window.

Figure 7: Frequency of SFS selection for each window size for VR-ISS and N-back. Features with less than 10% in every column were excluded from this figure for brevity. See appendix for the feature description and software package.

Task and window comparison for ABayes validation techniques

The validation techniques were compared for the VR-ISS and N-back (Figure 8). The window with the highest F1-score for the VR-ISS was 40 seconds (80%) for 10-Fold, 40-seconds (78%) for 2-Fold, and 20-30 seconds (80%) for holdout. For the N-back task, the window with highest F1-score was 30-seconds (85%) for 10-Fold, 40-seconds (83%) for 2-Fold, and 20-seconds (87%) for holdout.

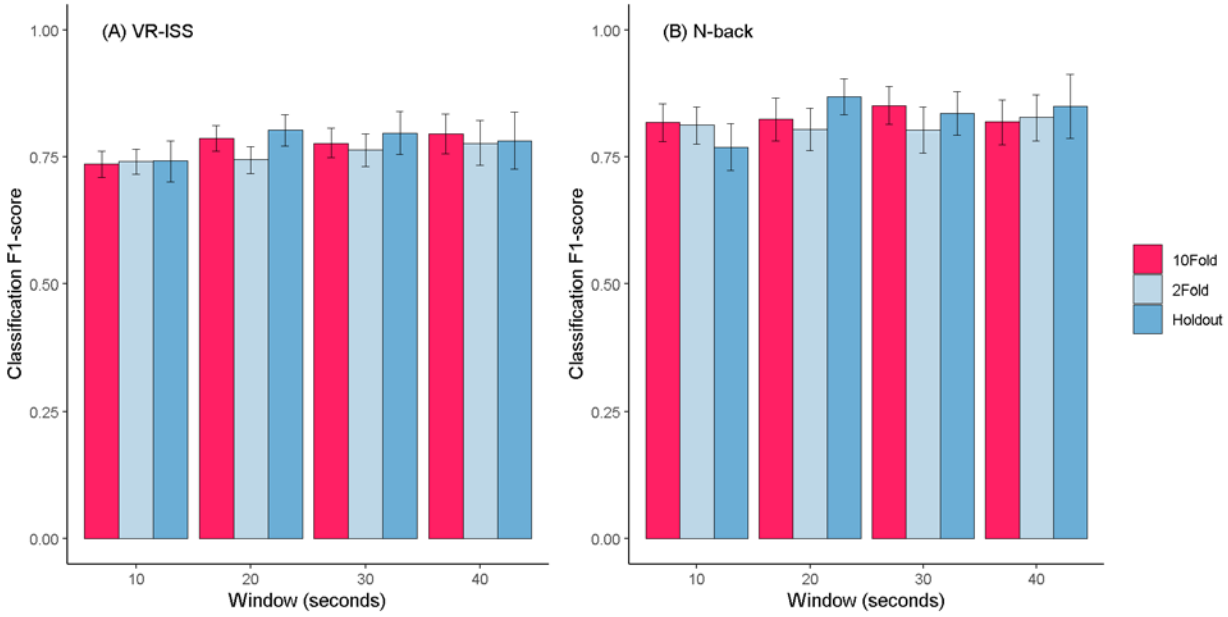

Figure 8: Validation technique comparison for (A) VR-ISS, and (B) N-back tasks. Error bars in standard error.

Classifier Comparison for the tasks

The cross-validation and holdout results for the VR-ISS task with the various classifiers trained with physiological signal segments of different window sizes are summarized in Table 5. The SVM classifier had the highest F1-score for 2-Fold and 10-Fold cross-validation. For 10-Fold, the best F1-score for the SVM model was 91% for a window size of 40-seconds. In comparison, the ABayes F1-score was 9% lower than the SVM for 40-seconds. The 2-Fold showed trends similar to the 10-Fold. The best average validation F1-score was for a window

size of 40 seconds, with an F1-score for SVM of 87%. In comparison, ABayes was 9% lower than the SVM for 40-seconds. For the holdout, the highest F1-score of 89% occurred for SVM with a window size of 40-seconds, while ABayes scored 11% lower.

Table 5: Results of the VR-ISS stress classification for different window sizes, classifiers, and validation techniques. Highest window F1-scores are highlighted.

		10Fold								2Fold								Holdout							
		ABayes		DT		RF		SVM		ABayes		DT		RF		SVM		ABayes		DT		RF		SVM	
		Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE
Window																									
10 sec	F1_score	0.74	0.03	0.75	0.02	0.76	0.02	0.83	0.02	0.74	0.02	0.72	0.02	0.74	0.03	0.81	0.02	0.74	0.04	0.79	0.03	0.76	0.03	0.75	0.04
	Accuracy	0.74	0.02	0.76	0.02	0.77	0.02	0.82	0.02	0.74	0.02	0.74	0.02	0.77	0.02	0.79	0.02	0.68	0.04	0.74	0.04	0.75	0.03	0.73	0.04
	Sensitivity	0.73	0.02	0.75	0.02	0.75	0.03	0.79	0.02	0.73	0.02	0.72	0.02	0.74	0.03	0.76	0.02	0.69	0.04	0.75	0.03	0.74	0.03	0.72	0.04
	Specificity	0.86	0.01	0.87	0.01	0.88	0.01	0.90	0.01	0.86	0.01	0.86	0.01	0.87	0.01	0.88	0.01	0.84	0.02	0.88	0.02	0.88	0.02	0.85	0.02
	Precision	0.79	0.02	0.75	0.02	0.76	0.03	0.83	0.02	0.76	0.02	0.72	0.02	0.75	0.03	0.80	0.02	0.68	0.04	0.77	0.04	0.77	0.04	0.73	0.04
20 sec	F1_score	0.79	0.03	0.78	0.02	0.77	0.03	0.87	0.02	0.74	0.03	0.73	0.03	0.76	0.03	0.84	0.02	0.80	0.03	0.82	0.04	0.85	0.03	0.82	0.03
	Accuracy	0.79	0.02	0.78	0.02	0.78	0.03	0.86	0.02	0.76	0.02	0.67	0.04	0.77	0.03	0.83	0.02	0.69	0.05	0.68	0.06	0.75	0.04	0.72	0.03
	Sensitivity	0.77	0.03	0.77	0.02	0.76	0.03	0.84	0.02	0.76	0.03	0.63	0.04	0.76	0.03	0.81	0.02	0.72	0.05	0.63	0.06	0.73	0.05	0.70	0.04
	Specificity	0.89	0.01	0.89	0.01	0.88	0.01	0.92	0.01	0.88	0.01	0.82	0.02	0.88	0.01	0.91	0.01	0.86	0.02	0.88	0.02	0.89	0.02	0.87	0.02
	Precision	0.82	0.02	0.78	0.03	0.78	0.03	0.87	0.02	0.79	0.02	0.68	0.04	0.76	0.03	0.84	0.02	0.73	0.05	0.76	0.05	0.81	0.04	0.76	0.03
30 sec	F1_score	0.78	0.03	0.76	0.03	0.77	0.03	0.87	0.02	0.76	0.03	0.65	0.03	0.73	0.03	0.84	0.03	0.80	0.05	0.69	0.05	0.79	0.04	0.82	0.04
	Accuracy	0.77	0.03	0.70	0.03	0.77	0.03	0.86	0.02	0.76	0.03	0.55	0.04	0.74	0.04	0.84	0.03	0.56	0.07	0.57	0.06	0.64	0.06	0.61	0.06
	Sensitivity	0.75	0.03	0.67	0.04	0.76	0.03	0.83	0.03	0.75	0.04	0.51	0.04	0.72	0.04	0.81	0.03	0.57	0.07	0.53	0.07	0.62	0.06	0.58	0.07
	Specificity	0.88	0.02	0.84	0.02	0.88	0.02	0.92	0.01	0.87	0.02	0.76	0.02	0.86	0.02	0.91	0.02	0.78	0.04	0.80	0.03	0.83	0.03	0.79	0.03
	Precision	0.80	0.03	0.70	0.04	0.76	0.03	0.87	0.02	0.80	0.03	0.56	0.04	0.72	0.04	0.85	0.03	0.59	0.07	0.57	0.06	0.64	0.06	0.57	0.07
40 sec	F1_score	0.80	0.04	0.76	0.04	0.80	0.04	0.91	0.02	0.78	0.04	0.68	0.04	0.77	0.06	0.87	0.04	0.78	0.06	0.65	0.06	0.78	0.06	0.89	0.04
	Accuracy	0.79	0.04	0.69	0.05	0.79	0.04	0.88	0.03	0.78	0.04	0.58	0.05	0.76	0.06	0.85	0.04	0.65	0.08	0.53	0.08	0.65	0.07	0.79	0.06
	Sensitivity	0.77	0.04	0.64	0.06	0.77	0.05	0.84	0.05	0.76	0.04	0.52	0.05	0.74	0.06	0.81	0.05	0.64	0.08	0.51	0.08	0.63	0.07	0.76	0.07
	Specificity	0.89	0.02	0.83	0.03	0.88	0.02	0.93	0.02	0.88	0.02	0.77	0.03	0.87	0.03	0.91	0.02	0.83	0.04	0.79	0.04	0.85	0.03	0.89	0.03
	Precision	0.85	0.03	0.66	0.06	0.77	0.05	0.91	0.03	0.84	0.03	0.60	0.06	0.76	0.07	0.88	0.04	0.68	0.08	0.61	0.09	0.72	0.07	0.81	0.06

The validation technique results for the N-back task are summarized in Table 6. The SVM classifier had the best accuracy and F1-score for 2-Fold and 10-Fold cross-validation. For 10-Fold, the SVM model's best F1-score was 92% for a window size of 40-seconds, with ABayes scoring 10% lower. The 2-Fold showed similar trends to the 10-Fold, with the SVM scoring 91% for a window size of 10-seconds., with ABayes scoring 10% lower. For the holdout, the best F1-score was 88% for Random Forest with a window size of 40-seconds, with ABayes only 3% lower. However, ABayes out-performed the other classifiers for the 20-second window with a F1-score of 87%.

Table 6: Results of the N-back stress classification for different window sizes, classifiers, and validation techniques. Highest window F1-scores are highlighted.

		10Fold								2Fold								Holdout							
		ABayes		DT		RF		SVM		ABayes		DT		RF		SVM		ABayes		DT		RF		SVM	
		Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE	Mean	STE
Window																									
10 sec	F1_score	0.82	0.04	0.88	0.03	0.88	0.02	0.91	0.02	0.81	0.04	0.88	0.02	0.87	0.02	0.91	0.02	0.77	0.05	0.79	0.05	0.78	0.05	0.80	0.06
	Accuracy	0.84	0.03	0.88	0.02	0.88	0.02	0.91	0.02	0.83	0.03	0.87	0.03	0.87	0.02	0.91	0.02	0.75	0.05	0.74	0.06	0.72	0.06	0.77	0.06
	Sensitivity	0.84	0.03	0.88	0.03	0.88	0.02	0.91	0.02	0.83	0.03	0.87	0.03	0.87	0.02	0.91	0.02	0.74	0.05	0.73	0.07	0.71	0.06	0.76	0.06
	Specificity	0.92	0.01	0.94	0.01	0.94	0.01	0.96	0.01	0.92	0.02	0.94	0.01	0.94	0.01	0.95	0.01	0.87	0.03	0.87	0.03	0.86	0.03	0.89	0.03
	Precision	0.89	0.02	0.88	0.03	0.89	0.02	0.92	0.02	0.87	0.03	0.89	0.02	0.88	0.02	0.92	0.02	0.78	0.05	0.75	0.07	0.73	0.06	0.81	0.05
20 sec	F1_score	0.82	0.04	0.87	0.02	0.88	0.03	0.89	0.03	0.80	0.04	0.81	0.04	0.85	0.03	0.87	0.03	0.87	0.04	0.73	0.05	0.79	0.05	0.81	0.04
	Accuracy	0.81	0.04	0.85	0.03	0.89	0.03	0.90	0.03	0.81	0.04	0.79	0.05	0.86	0.03	0.88	0.03	0.72	0.06	0.63	0.06	0.70	0.07	0.67	0.06
	Sensitivity	0.81	0.04	0.84	0.03	0.89	0.03	0.90	0.03	0.80	0.04	0.78	0.06	0.85	0.03	0.88	0.03	0.73	0.06	0.65	0.06	0.71	0.07	0.67	0.06
	Specificity	0.91	0.02	0.92	0.01	0.94	0.01	0.95	0.01	0.90	0.02	0.89	0.03	0.93	0.01	0.94	0.02	0.86	0.03	0.82	0.03	0.85	0.03	0.83	0.03
	Precision	0.83	0.05	0.86	0.03	0.89	0.03	0.90	0.03	0.84	0.04	0.80	0.06	0.86	0.03	0.89	0.03	0.73	0.07	0.68	0.06	0.73	0.07	0.71	0.06
30 sec	F1_score	0.85	0.04	0.82	0.04	0.84	0.04	0.90	0.04	0.80	0.05	0.74	0.04	0.82	0.04	0.89	0.04	0.84	0.04	0.83	0.04	0.76	0.05	0.85	0.04
	Accuracy	0.83	0.04	0.84	0.04	0.84	0.04	0.89	0.04	0.81	0.05	0.63	0.03	0.82	0.04	0.88	0.04	0.77	0.05	0.75	0.06	0.69	0.06	0.80	0.06
	Sensitivity	0.83	0.04	0.83	0.04	0.84	0.04	0.89	0.04	0.81	0.05	0.61	0.03	0.82	0.04	0.87	0.04	0.76	0.05	0.75	0.06	0.68	0.06	0.81	0.05
	Specificity	0.92	0.02	0.92	0.02	0.92	0.02	0.95	0.02	0.90	0.02	0.81	0.02	0.91	0.02	0.94	0.02	0.89	0.03	0.88	0.03	0.85	0.03	0.91	0.03
	Precision	0.87	0.04	0.84	0.04	0.84	0.04	0.89	0.05	0.85	0.04	0.67	0.04	0.83	0.04	0.88	0.04	0.79	0.05	0.77	0.06	0.73	0.06	0.82	0.06
40 sec	F1_score	0.82	0.04	0.80	0.03	0.87	0.04	0.92	0.03	0.83	0.05	0.53	0.01	0.85	0.04	0.89	0.04	0.85	0.06	0.78	0.06	0.88	0.05	0.87	0.05
	Accuracy	0.82	0.04	0.74	0.05	0.87	0.04	0.92	0.03	0.82	0.05	0.36	0.01	0.85	0.04	0.89	0.04	0.77	0.09	0.50	0.09	0.73	0.08	0.70	0.09
	Sensitivity	0.81	0.04	0.73	0.05	0.87	0.04	0.92	0.03	0.81	0.05	0.33	0.00	0.85	0.04	0.89	0.04	0.77	0.09	0.52	0.09	0.73	0.08	0.70	0.09
	Specificity	0.91	0.02	0.87	0.02	0.94	0.02	0.96	0.01	0.91	0.02	0.67	0.00	0.93	0.02	0.94	0.02	0.88	0.04	0.78	0.04	0.87	0.04	0.85	0.05
	Precision	0.85	0.04	0.73	0.05	0.88	0.04	0.93	0.03	0.85	0.04	0.36	0.01	0.87	0.04	0.90	0.04	0.79	0.08	0.54	0.10	0.77	0.08	0.72	0.10

Discussion

With the long-term goal of developing a real-time stress detection system, the purpose of this research was to evaluate different parameters of a time-series interval approach with a novel ABayes classifier, after physiological data were collected during stressful situations. The system was designed to select individualized time-series features that best describe the stress response for a given person and to use these features within the future model. The evaluated parameters included comparisons of machine learning classifiers, stress detection performance for two stressful tasks (simple and complex), window size of the interval method, and features selected by the wrapper. The subjective stress measures showed that both the VR-ISS and N-back successfully affected participants' stress into three distinct levels. The results showed a possible data imbalance and, therefore, the F1-score was used for classifier comparison. However, accuracy was compared between other studies due to limited statistical reporting. Results of cross-validation and holdout were promising for both tasks, with ABayes F1-scores ranging from 74-85% and 74-87%, respectively. SVM had higher F1-scores compared to the other classifiers during cross-validation; however, it is suspected that the results were biased and overestimated.

When comparing both tasks, the classifier performance was slightly better for the less-complex laboratory task of N-back. All window sizes had consistent classifier performance. In contrast, the features selected for each window varied, with the 10-20 sec windows having selected SBP mean and frequency features more than the 30-40 second windows, suggesting that physiological time-scale may influence feature classification performance. Overall, a personalized stress detection system using ABayes and time-series interval methods was comparable to other classifiers and had good performance at classifying three stress levels.

When comparing the classifiers, SVM resulted in generally higher F1-score for the cross-validation with Approximate Bayes, Decision Tree, and Random Forest showing higher scores for specific windows during the holdout validation. For cross-validation, the SVM had higher performance in all windows and tasks on with an F1-score range of 83-92% and 81-92% for 10-fold and 2-fold, respectively. In contrast, ABayes showed an F1-score range of 74-85% and 74-83% for 10-fold and 2-fold, respectively. Although ABayes was slightly lower than SVM, its performance was consistent and similar to Random Forest. The Decision tree had the lowest performance compared to the other classifiers, at one point showing a F1-score of 53% for a 2-Fold 40-second window during the N-back. The holdout showed similar performance among classifiers, within a few percentage points of each other, except for the SVM during 40-sec VR-ISS (89%) and the ABayes during 20-sec N-back (87%) which showed large performance differences.

The advantages of using ABayes in comparison to other classifiers is the direct and transparent connection to probability modeling. This allows for conditional probability that may better represent the dataset as a whole, since ABayes approximates the optimal solution by minimizing the probability of misclassification relative to the estimated density. Results show

that ABayes had slightly higher F1-scores for the VR-ISS and slightly lower for N-back compared to Decision Tree and Random Forest. This suggests that ABayes has very similar performance to traditional classifiers, when compared with equal parameters and biases. The outcome of any other classifiers (aside from SVM, see below) outperforming ABayes is largely due to circumstances being unfavorable to the ABayes direct method of approximation as opposed to one of the various indirect methods of classification/approximation. A possible limitation is the ABayes will only be optimal in the case that the features follow Gaussian distributions (Mariooryad & Busso, 2015). Further, the distributions obtained by some features, like phasic and tonic EDA, can on occasion be non-Gaussian. SVM and Decision Trees offer a slight advantage, as they are non-parametric and do not require a Gaussian distribution; however, have the tradeoff of a less consistent multi-class conditional probability estimates (Malley et al., 2012). Results suggest the assumption of Gaussian distribution did not adversely affect ABayes, but rather it was effected by wrapper biases.

When considering the relatively large F1-score of SVM compared to other classifiers, SVM may have been biased and overestimated the performance. The SVM had higher accuracy in all windows for cross-validation, being 8-14% better than the other classifiers, whereas the holdout only averaged 2-8% better performance for SVM. There are two possible explanations for these differences, both attributed to the SFS wrapper configuration: wrapper bias and data leakage.

Both the cross-validation and holdout may be affected by wrapper bias. Wrappers are widely used to select relevant features and enhance classification performance. Past research indicates that accuracy varies with the chosen wrapper classifiers, but SVM is well-suited to be wrapped because of its consistency in attaining smallest feature subsets (Bajer, Dudjak, & Zorić,

2020). However, wrappers only use a single classifier, which creates biases based how the classifier favors features; hence, the biases of chosen classifier used within a wrapper may influence the final accuracy (Chrysostomou, Chen, & Liu, 2008). Typically, the classifier chosen for the wrapper and test validation are the same in order to reduce the effects of bias, acting as a tuning method to optimize learner performance and yield better results (Binder et al., 2020). However, this research shows that a SVM within the wrapper can create bias by priming the selected features to have higher performance with SVM than with other classifiers during test validation (Samala et al., 2020). Based upon the consistent range of accuracy for the other classifiers, a 2-8% increase may have been caused by the wrapper; this demonstrates that wrapper based feature selection can result in even better accuracy if wrapper and classifier are the same.

Data leakage may have been responsible for SVM overestimation during cross-validation, possibly exacerbated by SVM wrapper bias. More specifically, data leakage may explain why the cross-validation shows 8-14% better performance for SVM than other classifiers, rather than 2-8% difference as seen in the holdout. Data leakage can occur in cross-validation when a feature selection is guided by the performance of the validation set (Samala et al., 2020). This can happen by repeatedly evaluating a trained model in order to make model decisions, such that knowledge of the entire dataset is gained (i.e., both training and testing partitions) by the model or transferred by the human developer (Petrick et al., 2013). In this research, cross-validation uses the individual's entire dataset in the SFS wrapper to assess classification performance of the features subset as features were sequentially added. This then guides the selection of the next best feature as SFS iterates. Since the wrapper is using a cost function to minimize the error that each new features introduces based on a SVM model, the

repeated testing on the entire dataset causes information leakage which overestimates SVM performance (as opposed to other classifiers) when tested in the final cross-validation. Since the holdout is isolated before feature selection happens, it limits the effects of data leakage on holdout test performance (Umar, Zhanfang, & Liu, 2020). Therefore, the holdout may be a fairer comparison of how SVM may perform when deployed in the wild, and the SVM cross-validation results should be interpreted as a different approach than the other classifiers.

Despite the ensuing biases, the wrapper method was necessary to select features that could be used to personalize the model, and subsequently deployed to test real-time data in future research. Without a wrapper, features would have to be predetermined and would generalize to a broad population when the system is deployed. However, they would neglect important individual differences in the physiological stress response. The simplest way to implement an optimized pipeline is to have the classifier that is within the wrapper match the classifier that is used during the validation techniques. Other potential solutions are to use wrapper-based decision trees to combine multiple classifiers to select mutually agreed relevant features (Chrysostomou, Chen, & Liu, 2008). It is likely that changing the wrapper classifier would increase performance for ABayes, Decision Tree, and Random Forest, similar to the SVM results.

In comparing the tasks, the F1-score was slightly higher for the N-back than the VR-ISS. This is expected, since the N-back is a more controlled laboratory task. The VR-ISS task is more complex and more closely matches the dynamic task demands faced in training of a real-world task. The highest N-back F1-score was for SVM with a 40-second window during 10-fold validation, resulting in 92% compared to 91% for VR-ISS with same parameters. ABayes with the same parameters resulted in 82% and 80% for N-back and VR-ISS, respectively. The slightly

higher accuracy was expected because the N-back is a more sustained and controlled stressor. Further, the N-back task is well validated at eliciting different levels of mental workload, specifically different levels of working memory, and physiological stress indices (Herff et al., 2014; Fallahi et al., 2016). In contrast, the VR-ISS was a more complex task involving a variety of stressors including, noise, task load, decreased visibility, and simulated physical threat. The accuracy for both tasks was relatively close, suggesting the stress detection may be robust in translation to other complex training tasks.

In comparison to other research on multiclass stress detection, ABayes for VR-ISS had lightly lower 10-fold accuracy of 79% (Table 7). However, directly comparing research is difficult due to varying factors in the pipeline development or differences in datasets. Notably, many of these studies used generalized classifiers and predetermined features. Since physiological stress activation can vary between individuals, some studies found individual stress detection models had higher classification accuracies than general models (Can et al., 2019b). Further, the SFS wrapper used within this study's pipeline gives an added advantage (SVM accuracy of 88%) as supported by a similar wrapper feature selection process that resulted in high accuracy (Pourmohammadi & Maleki, 2020).

Analysis of the window size for feature selection offers insight into prominent time-series patterns. Both feature extraction packages used by the pipeline contain measures of mean, variance, linearity, stationarity, frequency, and entropy. The features selected were similar for the VR-ISS window sizes of 10, 20 and 40 seconds; however, the features differed greatly for 30 seconds. For the 10, 20, and 40 sec windows, the most prominent features were the mean, median, and power spectral density for the signals SBP, RMSSD, and EDA tonic. These selected features fit the physiological narrative. SBP and EDA mean and median have been shown to be

elevated during acute stressors, while RMSSD decreases due to vagal withdrawal (Fredrikson, et al., 1989; Shaffer, McCraty, & Zerr, 2014). The power spectral density coefficients overlap the very-low and low frequency range (0.01-0.05 Hz). For SBP, this frequency range is associated with sympathetic activation of vascular tone (Langager et al., 2007). Similarly, the EDA low frequency range is associated with sympatric activation from stress (Posada-Quintero et al., 2016). Although, it is surprising that HRV frequency-domain features were not selected, considering that the HRV sympathetic measures are correlated to EDA (Posada-Quintero et al., 2016). For the 30-sec window, the most prominent features selected included DBP median, EDA tonic power spectral density, respiration median.

Table 7: Stress detection and classifiers for three or more levels stress (10-fold cross-validation accuracy)

Reference	Levels	Classifier	Subjects	Sensors	Generalized/ Individualized	Accuracy
This study	3	ABayes (40-sec window)	27 (VR-ISS)	ECG, EDA, RSP, NIBP	Ind.	79%
This study	3	SVM (40-sec window)	27 (VR-ISS)	ECG, EDA, RSP, NIBP	Ind.	88%
Tartarisco et al. (2015)	4	SVM	20	ECG	Gen.	86.3%
Plarre et al. (2011)	3	Ada Boost	21	ECG, RSP	Ind.*	90.2%
Boateng & Kotz (2016)	3	SVM	4	ECG	Gen.	89.2%
Bichindaritz et al. (2017)	3	Multi. Perceptron	17	ECG, EMG, RSP, EDA	Gen.	80.6%
Can et al. (2019b)	3	Random Forest	21	ECG, EDA	Ind.	97.2%
Pourmohammadi & Maleki (2020)	3	SVM	34	ECG, EMG	Gen.	97.6%
Šalkevicius et al. (2019)	4	SVM	30	BVP,EDA, ST	Gen.	86.3%
Keshan, Parimi, & Bichindaritz (2015)	3	Decision Tree	17	ECG,EMG,EDA, RSP	Gen.	70.2%
Arsalan et al. (2019)	3	Multi. Perceptron	28	EEG	Gen.	60.71%

* Generalized model, but calibrated to the individual using subjective stress scores.

As physiological systems act at different time-scales, the reliance on features changed as the analysis timeframe increased. The 30-40 second window had decreased selection of SBP, RMSSD, EDA phasic, and any power spectral density coefficient features compared to the 10-20 second window. For the VR-ISS, the 30-40 second window showed reliance on DBP, RMSSD, and EDA. The N-back showed reliance in the 30-40 second windows on Heart Rate skew, Heart Rate Partial Autocorrelation (one lag), and RMSSD Autocorrelation. This is somewhat expected as high frequency may be indicative of stress in shorter time intervals, but become diluted over larger windows. The selection of HR Partial Autocorrelation and RMSSD Autocorrelation for the N-back suggests a difference in stationarity, such as stochastic trends and systematic patterns that are unpredictable including points of abrupt change in mean level, frequency or amplitude. This suggests that HR has varying stochastic patterns at different stressor levels of N-back. As the features selected are relatively different for the windows and tasks, this suggest the generalized stress detection systems may not be robust for changing stressful scenarios due to each task evoking a different physiological stress response.

This study is subject to a number of limitations. One potential limitation is our wrapper overfitting the models due to highly correlated variables. The SFS wrapper fit the model by selecting a combination of features that resulted in the highest accuracy. However, the wrapper was not accounting for correlation between features. If all the features are selected from the same sensors, it not only neglects important physiological responses in other bodily systems but also increases the reliance on that one sensor working correctly (Chen et al., 2017). Lastly, some estimation error could have been caused by underspecification, which occurs when the training process has multiple predictors (e.g., feature structures) that appear equal, but have divergent performance when deployed (D'Amour et al., 2020). In this study, the SFS wrapper selected the

feature subset with the highest wrapper accuracy, but chose the subset with the least amount of features if multiple subsets had equal maximal performance. These subsets may have had different performance during cross-validation or holdout, and further research is needed to evaluate performance when deployed.

Acknowledgements

This work was funded by the National Aeronautics and Space Administration [grant number 80NSSC18K1572]. The authors thank Robin Gillund, Silvia Verhofste, Matthew Kreul, and Kelly Thompson for their laboratory assistance with research participants. For their help developing the VR-ISS and fire equipment models, the authors thank Pete Evans, Grant Leacox, Peter Carlson, and Robert Slezak.

Conclusion

A physiological-based stress detection system for classifying multiple levels of acute stress was developed with a novel classifier, ABayes, using a personalized time-series interval approach. The approach was evaluated against common machine learning classification systems in their ability to classify stress for two different tasks, VR-ISS and N-back. The current findings suggest that three levels of stress can be classified by means of the ABayes approach, providing promising accuracy when compared to past research on multi-class stress detection. Stress was accurately predicted for both the simplified lab task, N-back, and the more complex VR spaceflight emergency fire. Analysis on the window sizes gave insight into which sensors/features were useful for varying time-intervals. Further, our results demonstrated both the potential advantages and biases associated with wrapper feature selection methods, which need to be carefully considered when developing future systems. Future work will further

investigate these personalized stress detection systems with the aim of implementing real-time stress monitoring.

References

Alexandratos, V., Bulut, M., & Jasinschi, R. (2014, May). Mobile real-time arousal detection. *In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4394-4398). IEEE. https://doi.org/10.1109/ICASSP.2014.6854432

Akmandor, A. O., & Jha, N. K. (2017). Keep the stress away with SoDA: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems, 3*(4), 269-282. https://doi.org/10.1109/TMSCS.2017.2703613

Arsalan, A., Majid, M., Butt, A. R., & Anwar, S. M. (2019). Classification of perceived mental stress using a commercially available EEG headband. *IEEE journal of biomedical and health informatics, 23*(6), 2257-2264. https://doi.org/10.1109/JBHI.2019.2926407

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery, 31*(3), 606-660. https://doi.org/10.1007/s10618-016-0483-9

Bajer, D., Dudjak, M., & Zorić, B. (2020, October). Wrapper-based feature selection: how important is the wrapped classifier?. In *2020 International Conference on Smart Systems and Technologies (SST)* (pp. 97-105). IEEE. https://doi.org/10.1109/SST49455.2020.9264072

Barshi, I., & Dempsey, D. L. National Aeronautics and Space Administration (NASA). (2016). Risk of Performance Errors Due to Training Deficiencies: Evidence Report. (No. JSC-CN-35755). Houston, TX: NASA Johnson Space Center.

Betti, S., Lova, R.M., Rovini, E., Acerbi, G., Santarelli, L., Cabiati, M., Del Ry, S. & Cavallo, F. (2017). Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers. *IEEE Transactions on Biomedical Engineering, 65*(8), 1748-1758. https://doi.org/10.1109/TBME.2017.2764507

Bichindaritz, I., Breen, C., Cole, E., Keshan, N., & Parimi, P. (2017, June). Feature selection and machine learning based multilevel stress detection from ECG Signals. In *International Conference on Innovation in Medicine and Healthcare* (pp. 202-213). Springer, Cham. https://doi.org/10.1007/978-3-319-59397-5_22

Binder, M., Moosbauer, J., Thomas, J., & Bischl, B. (2020, June). Multi-objective hyperparameter tuning and feature selection using filter ensembles. *In Proceedings of the 2020 Genetic and Evolutionary Computation Conference* (pp. 471-479). https://doi.org/10.1145/3377930.3389815

Boateng, G., & Kotz, D. (2016, November). StressAware: An app for real-time stress monitoring on the amulet wearable platform. In *2016 IEEE MIT Undergraduate Research Technology Conference (URTC)* (pp. 1-4). IEEE. https://doi.org/10.1109/URTC.2016.8284068

Bong, S. Z., Murugappan, M., & Yaacob, S. (2013). Methods and approaches on inferring human emotional stress changes through physiological signals: A review. *International Journal of Medical Engineering and Informatics, 5*(2), 152-162. https://doi.org/10.1504/IJMEI.2013.053332

Bowers, S. L., Bilbo, S. D., Dhabhar, F. S., & Nelson, R. J. (2008). Stressor-specific alterations in corticosterone and immune responses in mice. *Brain, behavior, and immunity, 22*(1), 105-113. https://doi.org/10.1016/j.bbi.2007.07.012

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.

Campbell, E., Phinyomark, A., & Scheme, E. (2019). Feature extraction and selection for pain recognition using peripheral physiological signals. *Frontiers in neuroscience, 13*, 437. https://doi.org/10.3389/fnins.2019.00437

Can, Y. S., Arnrich, B., & Ersoy, C. (2019a). Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics, 92*, 103139. https://doi.org/10.1016/j.jbi.2019.103139

Can, Y. S., Chalabianloo, N., Ekiz, D., & Ersoy, C. (2019b). Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors, 19*(8), 1849. https://doi.org/10.3390/s19081849

Carpenter, R. (2016). A review of instruments on cognitive appraisal of stress. *Archives of Psychiatric Nursing, 30*(2), 271-279. https://psycnet.apa.org/doi/10.1016/j.apnu.2015.07.002

Chen, L. L., Zhao, Y., Ye, P. F., Zhang, J., & Zou, J. Z. (2017). Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications, 85*, 279-291. http://dx.doi.org/10.1016/j.eswa.2017.01.040

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing, 307*, 72-77. https://doi.org/10.1016/j.neucom.2018.03.067

Chrysostomou, K., Chen, S. Y., & Liu, X. (2008). Combining multiple classifiers for wrapper feature selection. *International Journal of Data Mining, Modelling and Management, 1*(1), 91-102. http://dx.doi.org/10.1504/IJDMMM.2008.022539

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillside, NJ: Lawrence Earlbaum Associates.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., … & Sculley, D. (2020). Underspecification Presents Challenges for Credibility in Modern Machine Learning. *ArXiv, abs/2011.03395.*

Dorneich, M. C., Whitlow, S. D., Mathan, S., Ververs, P. M., Erdogmus, D., Adami, A., Pavel, M. & Lan, T. (2007). Supporting real-time cognitive state classification on a mobile individual. *Journal of Cognitive Engineering and Decision Making, 1*(3), 240-270. https://doi.org/10.1518%2F155534307X255618

Driskell, J. E., Salas, E., Johnston, J. H., & Wollert, T. N. (2008). Stress exposure training: An event-based approach. Performance under stress, 271–286. London: Ashgate.

Fallahi, M., Heidarimoghadam, R., Motamedzade, M., & Farhadian, M. (2016). Psycho physiological and subjective responses to mental workload levels during n-back task. *J Ergonomics, 6*(181), 1-7. https://doi.org/10.4172/2165-7556.1000181

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. A handbook of process tracing methods for decision research, 163-184.

Finseth, T., Dorneich, M. C., Keren, N., Franke, W. D., & Vardeman, S. (2020, October). Designing Training Scenarios for Stressful Spaceflight Emergency Procedures. *In 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC) (pp. 1-10).* IEEE. https://doi.org/10.1109/DASC50938.2020.9256403

Finseth, T., Keren, N., Franke, W.D., Dorneich, M.C., Anderson, C.C, & Shelly, M. (2018). Evaluating the Effectiveness of Graduated Stress Exposure in Virtual Spaceflight Hazard Training, *Journal of Cognitive Engineering and Decision Making. 12*(4), pp. 248-268. https://doi.org/10.1177%2F1555343418775561

Fredrikson, M., Blumenthal, J. A., Evans, D. D., Sherwood, A., & Light, K. C. (1989). Cardiovascular responses in the laboratory and in the natural environment: Is blood pressure reactivity to laboratory-induced mental stress related to ambulatory blood pressure during everyday life?. *Journal of Psychosomatic Research, 33*(6), 753-762. https://doi.org/10.1016/0022-3999(89)90091-3

Fulcher, B. D. (2018). Feature-based time-series analysis. In Feature engineering for machine learning and data analytics (pp. 87-116). CRC Press.

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2019). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing.* https://doi.org/10.1109/TAFFC.2019.2927337

Gjoreski, M., Luštrek, M., Gams, M., & Gjoreski, H. (2017). Monitoring stress with a wrist device using context. *Journal of biomedical informatics*, *73*, 159-170. https://doi.org/10.1016/j.jbi.2017.08.006

Greene, S., Thapliyal, H., & Caban-Holt, A. (2016). A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine*, *5*(4), 44-56. http://dx.doi.org/10.1109/MCE.2016.2590178

He, H., & Garcia, E.A. (2009). *Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21,* 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Healey, J., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, *6*(2), 156-166. https://doi.org/10.1109/TITS.2005.848368

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. *Frontiers in human neuroscience, 7*, 935. https://doi.org/10.3389/fnhum.2013.00935

Hidasi, B., & Gáspár-Papanek, C. (2011, September). ShiftTree: an interpretable model-based approach for time series classification. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 48-64). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23783-6_4

Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M. & Kumar, S. (2015). September. cStress: towards a gold standard for continuous stress assessment in the mobile environment. *In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 493-504. ACM. https://doi.org/10.1145/2750858.2807526

HTC Corporation: (2016). Vive - Home. https://www.htcvive.com/ca/

Jones, D., & Dechmerowski, S. (2016, July). Measuring Stress in an Augmented Training Environment: Approaches and Applications. *In International Conference on Augmented Cognition* (pp. 23-33). Springer, Cham. https://doi.org/10.1007/978-3-319-39952-2_3

Karthikeyan, P., Murugappan, M., & Yaacob, S. (2013). Detection of human stress using short-term ECG and HRV signals. *Journal of Mechanics in Medicine and Biology, 13*(02), 1350038. https://doi.org/10.1142/S0219519413500383

Keshan, N., Parimi, P. V., & Bichindaritz, I. (2015, October). Machine learning for stress detection from ECG signals in automobile drivers. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2661-2669). IEEE. https://doi.org/10.1109/BigData.2015.7364066

Langager, A. M., Hammerberg, B. E., Rotella, D. L., & Stauss, H. M. (2007). Very low-frequency blood pressure variability depends on voltage-gated L-type Ca2+ channels in conscious rats. *American Journal of Physiology-Heart and Circulatory Physiology, 292*(3), H1321-H1327. https://doi.org/10.1152/ajpheart.00874.2006

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., & Jones, N. S. (2019). catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery, 33*(6), 1821-1852. https://doi.org/10.1007/s10618-019-00647-x

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., & Ziegler, A. (2012). Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of information in medicine, 51*(1), 74. https://dx.doi.org/10.3414%2FME00-01-0052

Mariooryad, S., & Busso, C. (2015). The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier. *IEEE Transactions on Affective Computing*, *8*(1), 119-130. https://doi.org/10.1109/TAFFC.2015.2508454

Martinez, R., Irigoyen, E., Arruti, A., Martin, J. I., & Muguerza, J. (2017). A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine. *Computer methods and programs in biomedicine, 148*, 81-90. https://doi.org/10.1016/j.cmpb.2017.06.010

McGee, S. (2002). Simplifying likelihood ratios. *Journal of general internal medicine, 17*(8), 647-650. https://doi.org/10.1046/j.1525-1497.2002.10750.x

Molkkari, M., Angelotti, G., Emig, T., & Räsänen, E. (2020). Dynamical heart beat correlations during running. *Scientific reports, 10*(1), 1-9. https://doi.org/10.1038/s41598-020-70358-7

National Aeronautics and Space Administration (NASA). (2013). International Space Station, Emergency Procedures 1a: Depress, Fire, Equipment Retrieval (No. JSC-48566). Houston, TX: NASA Johnson Space Center.

Novak, V., Novak, P., de Champlain, J., Le Blanc, A. R., Martin, R., & Nadeau, R. (1993). Influence of respiration on heart rate and blood pressure fluctuations. *Journal of Applied Physiology*, *74*(2), 617-626. https://doi.org/10.1152/jappl.1993.74.2.617

Petrick, N., Sahiner, B., Armato III, S.G., Bert, A., Correale, L., Delsanto, S., Freedman, M.T., Fryd, D., Gur, D., Hadjiiski, L., Huo, Z., Jiang, Y., Morra, L., Paquerault, S. Raykar, V., Samuelson, F., Summers, R.M., Tourassi, G., Yoshida, H., … & Chan, H. (2013). Evaluation of computer-aided detection and diagnosis systems. *Medical physics*, *40*(8), 087001. https://doi.org/10.1118/1.4816310

Plarre, K., Raij, A., Hossain, S.M., Ali, A.A., Nakajima, M., Al'Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M. and Siewiorek, D., 2011, April. Continuous inference of psychological stress from sensory measurements collected in the natural environment. *In Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks* (pp. 97-108). IEEE.

Posada-Quintero, H. F., & Chon, K. H. (2020). Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors, 20*(2), 479. https://doi.org/10.3390/s20020479

Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, A. D., Aljama-Corrales, T., Charleston-Villalobos, S., & Chon, K. H. (2016). Power spectral density analysis of electrodermal activity for sympathetic function assessment. *Annals of biomedical engineering, 44(*10), 3124-3135. https://doi.org/10.1007/s10439-016-1606-6

Pourmohammadi, S., & Maleki, A. (2020). Stress detection using ECG and EMG signals: A comprehensive study. *Computer Methods and Programs in Biomedicine*, 105482. https://doi.org/10.1016/j.cmpb.2020.105482

Reimer, U., Laurenzi, E., Maier, E., & Ulmer, T. (2017, January). Mobile Stress Recognition and Relaxation Support with SmartCoping: User-Adaptive Interpretation of Physiological Stress Parameters. *In Proceedings of the 50th Hawaii International Conference on System Sciences.* http://dx.doi.org/10.24251/HICSS.2017.435

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, *3*(22), 41-46.

Saeed, A., Trajanovski, S., Keulen, M. V., & Erp, J. V. (2017). Deep Physiological Arousal Detection in a Driving Simulator Using Wearable Sensors. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 486–493. https://doi.org/10.1109/ICDMW.2017.69

Šalkevicius, J., Damaševičius, R., Maskeliunas, R., & Laukienė, I. (2019). Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics, 8*(9), 1039. https://doi.org/10.3390/electronics8091039

Samala, R. K., Chan, H. P., Hadjiiski, L., & Koneru, S. (2020, March). Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In *Medical Imaging 2020: Computer-Aided Diagnosis* (Vol. 11314, p. 1131416). International Society for Optics and Photonics. https://doi.org/10.1117/12.2549313

Sedghamiz, H. (2018). BioSigKit: A Matlab Toolbox and Interface for Analysis of BioSignals. *Journal of Open Source Software, 3*(30), 671. https://doi.org/10.21105/joss.00671

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in public health, 5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Shaffer, F., McCraty, R., & Zerr, C. L. (2014). A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01040

Sharma, N., & Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine*, *108*(3), 1287-1301. https://doi.org/10.1016/j.cmpb.2012.07.003

Shields, G. S., Sazma, M. A., & Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neuroscience & Biobehavioral Reviews*, *68*, 651-668. https://doi.org/10.1016/j.neubiorev.2016.06.038

Singh, R. R., Conjeti, S., & Banerjee, R. (2013). A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, *8*(6), 740-754. https://doi.org/10.1016/j.bspc.2013.06.014

Singh, R. R., Conjeti, S., & Banerjee, R. (2014). Assessment of driver stress from physiological signals collected under real-time semi-urban driving scenarios. *International Journal of Computational Intelligence Systems*, *7*(5), 909-923. https://doi.org/10.1080/18756891.2013.864478

Smets, E., De Raedt, W., & Van Hoof, C. (2019). Into the wild: the challenges of physiological stress detection in laboratory and ambulatory settings. *IEEE journal of biomedical and health informatics*, *23*(2), 463-473. https://doi.org/10.1109/JBHI.2018.2883751

Sun, F. T., Kuo, C., Cheng, H. T., Buthpitiya, S., Collins, P., & Griss, M. (2010, October). Activity-aware mental stress detection using physiological sensors. *In International conference on Mobile computing, applications, and services*, 282-301. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29336-8_16

Tartarisco, G., Carbonaro, N., Tonacci, A., Bernava, G.M., Arnao, A., Crifaci, G., Cipresso, P., Riva, G., Gaggioli, A., De Rossi, D., & Tognetti, A. (2015). Neuro-fuzzy physiological computing to assess stress levels in virtual reality therapy. *Interacting with Computers*, *27*(5), 521-533. https://doi.org/10.1093/iwc/iwv010

Tong, S., & Koller, D. (1999). Bayes optimal hyperplanes → maximal margin hyperplanes. Submitted to *IJCAI*, 99.

Umar, M. A., Zhanfang, C., & Liu, Y. (2020, January). Network Intrusion Detection Using Wrapper-based Decision Tree for Feature Selection. In *Proceedings of the 2020 International Conference on Internet Computing for Science and Engineering* (pp. 5-13). https://doi.org/10.1145/3424311.3424330

Verleysen, M., & François, D. (2005, June). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks* (pp. 758-770). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11494669_93

Vollmer, M. (2019). HRVTool–an Open-Source Matlab Toolbox for Analyzing Heart Rate Variability. *In 2019 Computing in Cardiology (CinC)* (pp. Page-1). IEEE. https://doi.org/10.23919/CinC49843.2019.9005745

Zadrozny, B., & Elkan, C. (2002, July). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699). https://doi.org/10.1145/775047.775151

Zahabi, M., & Razak, A. M. A. (2020). Adaptive virtual reality-based training: a systematic literature review and framework. *Virtual Reality*, 1-28. https://doi.org/10.1007/s10055-020-00434-w

Zarei, A., & Asl, B. M. (2020). Automatic classification of apnea and normal subjects using new features extracted from HRV and ECG-derived respiration signals. *Biomedical Signal Processing and Control*, *59*, 101927. https://doi.org/10.1016/j.bspc.2020.101927

Zheng, B. S., Murugappan, M., & Yaacob, S. (2012, September). Human emotional stress assessment through Heart Rate Detection in a customized protocol experiment. *In Industrial Electronics and Applications (ISIEA), 2012 IEEE Symposium on Industrial Electronics and Applications* (pp. 293-298). IEEE. https://doi.org/10.1109/ISIEA.2012.6496647

Zhu, R., Wang, Z., Ma, Z., Wang, G., & Xue, J. H. (2018). *LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test. Pattern Recognition Letters*, *116*, 36-42. https://doi.org/10.1016/j.patrec.2018.09.012

Appendix A. Extracted Features

Table 8: List of all features included in the feature extraction, grouped by signal and listed in alphabetical order.

#	Feature	Abbreviation	Reference/Package
Extracted from all Biopac signals (HR,RMSSD,pNN50,RSP,NIBP,SBP,DBP,EDA,EDAtonic,EDAphasic)			
1	Absolute Energy	abs_energy	Tsfresh
2	Absolute sum of changes	abs_sum_changes	Tsfresh
3	Augmented Dickey Fuller	adfpValue	Tsfresh
4	Autocorrelation	Agg_AutoC	Tsfresh
5	Autocorrelation, first 1/e crossing	CO_f1ecac	Catch22
6	Autocorrelation, fist minimum	CO_FirstMin_ac	Catch22
7	Automutual info, m=2, τ=5	CO_HistogramAMI_even_2_5	Catch22
8	Automutual, first minimum	IN_AutoMutualInfoStats_40_gaussian_fmmi	Catch22
9	Binned entropy	binned_entropy	Tsfresh
10	C3 non-linearity measure	c3	Tsfresh
11	Change in correlation length after iterative differencing	FC_LocalSimple_mean1_tauresrat	Catch22
12	Complexity-invariant distance	cid_ce	Tsfresh
13	Count above mean	GTmean	Tsfresh
14	Count below mean	LTmean	Tsfresh
15	Cross Correlation	CC	Campbell, Phinyomark, & Scheme, 2019
16	Exponential fit to successive distances in 2-d embedding space	CO_Embed2_Dist_tau_d_expfit_meandiff	Catch22

Table 16 Continued

#	Feature	Abbreviation	Reference/Package
17	FFT aggregated spectral variance	fft_agg_var	Tsfresh
18	FFT real coefficients	rfft_real	Tsfresh
19	First location of maximum	first_loc_max	Tsfresh
20	First location of minimum	first_loc_min	Tsfresh
21	Kurtosis	Kurt	Tsfresh
22	Last location of maximum	last_loc_max	Tsfresh
23	Last location of minimum	last_loc_min	Tsfresh
24	Linear trend slope	linear_slope	Tsfresh
25	Longest strike above mean	longest_strike_above	Tsfresh
26	Longest strike below mean	longest_strike_below	Tsfresh
27	Mean	mean	Tsfresh
28	Mean absolute change	mean_abs_change	Tsfresh
29	Mean change	mean_change	Tsfresh
30	Mean second derivative central	mean_2nd	Tsfresh
31	Median	median	Tsfresh
32	Mode of distribution (5,10-bin histo)	DN_HistogramMode_5,10	Catch22
33	Number of crossings mean	num_cross_mean	Tsfresh
34	Partial Autocorrelations (1 lag)	Agg_PAutoC	Tsfresh
35	Percent of reoccurring data points to all data points	per_reocurr_dtp	Tsfresh
36	Percent of reoccurring values to all values	per_reocurr_val	Tsfresh
37	Ratio value number to time series length	ratio_val_totime_series	Tsfresh
38	Rolling 3-sample mean forecasting error	FC_LocalSimple_mean3_stderr	Catch22
39	Skewness	skew	Tsfresh
40	Periodicity measure	PD_PeriodicityWang_th0_01	Catch22
41	Power spectrum – Fourier, centroid	SP_Summaries_welch_rect_centroid	Catch22
42	Power spectrum – Fourier, total power of lowest fifth frequency	SP_Summaries_welch_rect_area_5_1	Catch22
43	Proportion of slower timescale fluctuations that scale with DFA (50% sampling)	SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1	Catch22
44	Proportion of slower timescale fluctuations that scale with linearly rescaled range fits	SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1	Catch22
4546	Shannon entropy of two successive letters in equiprobable 3-letter symbolization	SB_MotifThree_quantile_hh	Catch22
47	Standard deviation	stddev	Tsfresh
48	Standard deviation of successive differences	SDSD	Campbell, Phinyomark, & Scheme, 2019
49	Successive differences exceeding 0.04σ	MD_hrv_classic_pnn40	Catch22
50	Successive differences longest period of decreases	SB_BinaryStats_diff_longstretch0	Catch22
51	Sum of reoccurring data points	sum_reoccuring_dpt	Tsfresh
52	Sum of reoccurring values	sum_reoccuring_val	Tsfresh
53	Sum of squares	SS	Tsfresh
54	Sum values	sum_val	Tsfresh
55	Time intervals between events above mean	DN_OutlierInclude_p_001_mdrmd	Catch22
56	Time intervals between events below mean	DN_OutlierInclude_n_001_mdrmd	Catch22
57	Time reversal asymmetry statistic	time_reversal	Tsfresh
58	Time-reversibility statistic	CO_trev_1_num	Catch22
59	Trace of covariance of transition matrix between symbols in 3-letter alphabet	SB_TransitionMatrix_3ac_sumdiagcov	Catch22
60	Variance	variance	Tsfresh

Table 16 Continued

#	Feature	Abbreviation	Reference/Package
Extracted from all ECG Biopac signal			
61	Baseline width of the RR interval histogram	TINN	HRVTool
62	Heart rate	HR	HRVTool
63	High Frequency power	HF	HRVTool
64	Low Frequency power	LF	HRVTool
65	Peak Freq. of Low Freq. Band	pLF	HRVTool
66	Peak Freq. of High Freq. Band	pHF	HRVTool
67	Percent of R peaks in ECG that differ more than 50 millisecond	pNN50	HRVTool
68	Percent of R peaks in ECG that differ more than 20 millisecond	pNN20	HRVTool
69	Poincaré plot standard deviation perpendicular the line of identity	SD1	HRVTool
70	Poincaré plot standard deviation along the line of identity	SD2	HRVTool
71	Ratio of SD1-to-SD2	SD1SD2ratio	HRVTool
72	Ratio of Low-High Frequency power	LFHFratio	HRVTool
73	Root Mean Square of Successive Difference of RR interval	RMSSD	HRVTool
74	Standard deviation of successive differences	SDSD	HRVTool
75	Standard deviation of NN intervals	SDNN	HRVTool
76	Triangular index from the interval histogram	TRI	HRVTool
77	Very Low Frequency power	VLF	HRVTool

Appendix B. Approval for Research (IRB)

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 11/16/2018

To: Tor Finseth Michael Dorneich, Ph.D.

From: Office for Responsible Research

Title: **Developing and testing a stress gauge using virtual reality**

IRB ID: **18-432**

Submission Type: Initial Submission **Review Type:** Full Committee

Approval Date: **11/15/2018** **Date for Continuing Review:** **11/14/2020**

CHAPTER 7. AN APPROACH TO ADAPTIVE TRAINING FOR STRESS INOCULATION

Tor Finseth[1], Michael C. Dorneich[2], Nir Keren[3], Warren D. Franke[4], Stephen Vardeman[2], Jonathan Segal[5], Andrew Deick[5], Elizabeth Cavanah[2], & Kelly Thompson[1]

Portions of this chapter were submitted as part of an abridged version, with preliminary results, to *65th Human Factors and Ergonomics Society Annual Meeting* (Finseth et al., 2021b).

Statement of Authorship

As the lead author, I developed the theoretical background, programmed the *adaptive* system, designed and conducted the experiment, performed the analysis, and was the primary writer for the manuscript. Dr. Dorneich provided detailed supervision at all stages of the project. Dr. Keren, Dr. Franke, and Dr. Vardeman helped guide development of the system, create training practices, select the statistical analysis, and interpret the results. Jonathan Segal, Andrew Deick, Elizabeth Cavanah, and Kelly Thompson drafted sections of the introduction and methods under my supervision. The manuscript was reviewed with input from all authors.

Abstract

Astronauts operate in an environment with multiple hazards that can develop into life-threatening emergency situations. Managing stress in emergencies may require cognitive resources and diminish performance. Stress training aims to maintain performance under stress by methodically increasing stressor levels to build resilience. An adaptive virtual reality (VR) training system was developed with real-time stress detection by using machine learning on

[1] Dept. of Aerospace Engineering, Iowa State University, Ames, IA, 50011, USA
[2] Dept. of Industrial Manufacturing and Systems Engineering, Iowa State University, Ames, IA, 50011, USA
[3] Dept. of Agricultural and Biosystems Engineering , Iowa State University, Ames, IA, 50011, USA
[4] Dept. of Kinesiology, Iowa State University, Ames, IA, 50011, USA
[5] Dept. of Computer Science, Iowa State University, Ames, IA, 50011, USA

psychophysiological responses. Using a VR simulation of a spaceflight emergency fire, predictions of the individual's stress levels were used to trigger adaptations of the environmental stressors (e.g., smoke, alarms, flashing lights), with the goal of maintaining an optimal level of stress during training. Sixty-five healthy subjects underwent task training over eight trials with adaptive training (adaptive, N=23); results were compared to trials with predetermined gradual increases in stressors (graduated, N=22), and trials with constant low-level stressors (skill-only, N=20). Psychological responses were measured with subjective stress, task engagement, distress, worry, anxiety, and workload scales. Physiological responses were measured through heart rate, heart rate variability, blood pressure, electrodermal activity, and task performance. The change from before and after receiving training was analyzed and compared between training conditions. The adaptive condition showed a significant decrease in heart rate and a decreasing trend in LF/HF ratio, but no changes in the other training conditions. The distress showed a decreasing trend for the graduated and adaptive conditions. The task engagement showed a significant increase for adaptive and a significant decrease for the graduated condition, resulting in the two conditions being significantly different. All training conditions showed a significant decrease in worry and anxiety and a significant increase for the heart rate variability metrics of RMSSD and pNN50. Results suggests that all training conditions lowered stress, but the preponderance of trial effects for the adaptive condition suggest it is more successful decreasing stress over multiple trials. Task performance in the form of number of contaminate readings improved for the skill-only and adaptive condition, but together with psychological and physiological findings, suggests performance changes are due to repetitive task-training in the *skill-only* and development of emotional regulation strategies in the *adaptive* condition. Results suggest that

training with the adaptive stress system can prepare individuals for responding to stressors better than the *skill-only* and *graduated* training.

Introduction

As the length of spaceflights increases, astronauts are more likely to encounter emergencies. They will have to respond to these high-stress, life-threatening situations efficiently and quickly. Acute stress can have detrimental effects on attention, memory, perceptual-motor performance, judgment, and decision-making (Driskell et al., 2008). These stressful events may lead to increased risk of harm, injury, mortality, or mission failure. NASA seeks developing countermeasures to prevent the consequences due to acute stress (NASA, 2021). Some researchers have also suggested the need for advanced training systems in preparation for future deep space missions, including stress-inducing simulation training (Russi-Vigoya et al., 2020).

Traditional training practices focus on performance outcomes, but often use training environments that poorly replicate the stress felt during operations. Supplementing current training with training that focuses on an individual's acute stress and coping skills may prevent adverse behavior and performance degradation during actual emergencies. Building resilience for potentially hazardous spaceflight conditions may prevent freezing behavior and apportion cognitive resources for the task.

Advances in virtual reality (VR) technology allow simulating stressors that approximate those of the real task environment. It is possible to categorize stress levels by using machine learning algorithms with physiological sensors, which then facilitates personalizing an optimal environment for training (Jones & Dechmerowski, 2016). A VR-based training system in conjunction with an adaptive process can adapt stressor levels based on the crewmember's stress response and may increase resilience to stressors. However, there is a need for further evidence

that VR-based stress training programs can enhance performance along with preventing stress (Pallavicini et al., 2016).

For this study, a real-time adaptive system was developed and evaluated for stress training. An experiment was conducted to compare the effectiveness of training with adaptations in comparison to training with *graduated* or *skill-only* pedagogies. A simulated emergency in the form of a fire aboard a VR International Space Station (VR-ISS) was used. Stress responses and task performance were measured across training trials.

Real-time Adaptive System

An adaptive system was designed to assess stress level via physiological signals in real-time, and then adapt the simulated environment to optimize training to user capabilities (Figure 1). The system relied on five components: physiological sensors, context assessment & classification, training module, adaptation manager, and VR simulation. Rules and triggers were developed to trigger stressor level adaptations.

System Description

To adapt the system effectively to meet user needs, a personalized stress profile was established through continuous measurement of stress levels. Physiological sensors collected data in real-time, pre-processed to remove artifacts, and used to derive features that commonly measure autonomic system activation (Giannakakis et al., 2019).

The context assessment component evaluates the state of the system, simulated environment, and user stress state. To develop a real-time measure of stress, data from physiological sensors were continuously streamed (125Hz), buffered in memory, used to derive hundreds of features (e.g., mean, variance, linearity, stationarity, entropy, frequency features), and then reduced to a small subset of features that were most discriminant of the individual's

stress profile. The selected subset of features were then used by a novel machine learning algorithm Approximate Bayes to predict the individual's stress level continuously for every 30-second window.

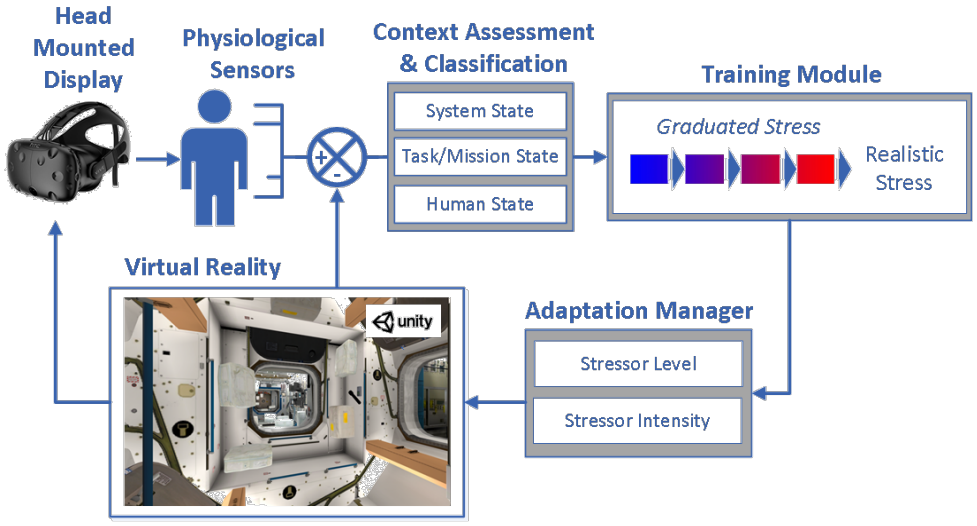

Figure 1: Conceptual diagram of the Adaptive Stress Training System.

The training module determined the target level of stress by following the SET protocol to introduce stressors over multiple user trials. Three levels of stressors were induced through a combination of environment changes (smoke, alarm noise, and interior lights) to create low, medium, and high stress levels (Figure 2). This manipulation of stress was validated in a previous study using subjective stress and workload scales (Finseth et al., 2020).

Figure 2: VR-ISS low, medium, high stressor levels.

The adaptation manager adapted the levels of the stressors in the VR environment based on a comparison of the target level of stress from the training module to the measured level of stress from the context assessment module. The adaptation manager sent adaptations to the VR simulation to implement in the VR-ISS (e.g., changes in smoke, alarm noise, lights).

For the purpose of stress training, changes in measured stress triggered adaptations to the simulated environment. This type of change is known as implicit feedback, in which the user observes feedback in the form of changes to the environment (e.g., visual or auditory stressors; Gaume et al., 2016). Instead of explicitly showing the sensor signals like biofeedback systems to the user, the physiological state is assessed and used to implicitly modify the environment.

System Architecture

A machine learning pipeline used for offline validation (Finesth et al., 2021) was modified to collect and classify stress level in real-time (Figure 3). The top three components (data collection, feature extraction, and feature selection) were used on a training dataset. This dataset includes the participant's physiological signals in each of the three stress classes, all collected prior to using the real-time adaptive system to predict stress. The features selected from the training dataset are then passed into a newly developed real-time classification, adaptive manager, and VR components (bottom three components).

Real-time Classification Component

A parallel-processing framework composed of four workers is used to manage the simultaneous network configuration, data streaming, classification, and adaptations. Each worker is a MATLAB computational engine that can allocate computations to physical processor cores. This parallel framework is necessary to avoid losing data, because functions executed in series would require the data streaming to temporally stop in order to execute classification.

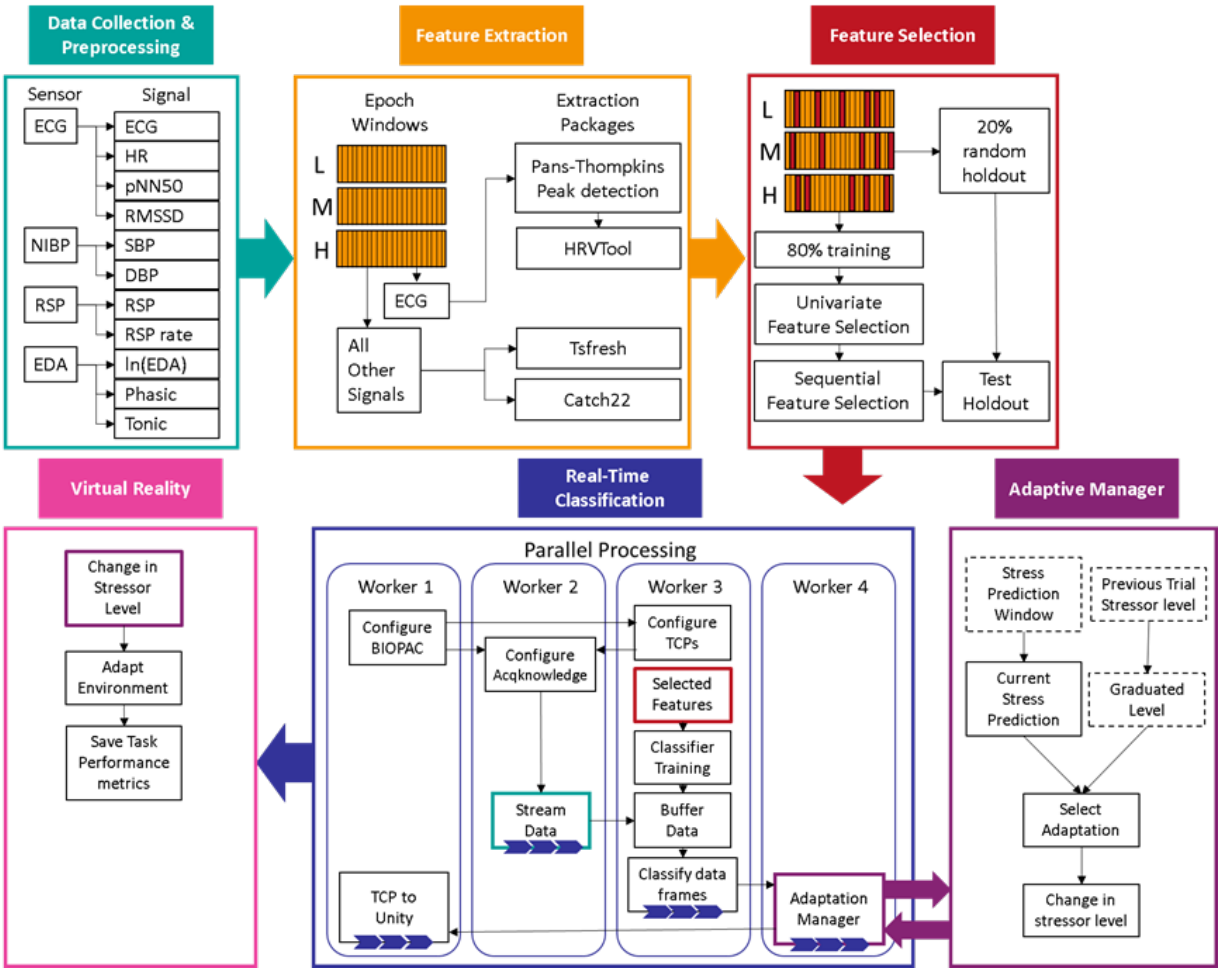

Figure 3: Adaptive VR stress training system architecture. Blue arrows represent continuous data streams. Dashed outlined boxes represent variable inputs and colored boxes represent inputs from other components.

The real-time classification component beings by configuring a Biopac MP150 and Biopac Acqknowledge software client for transferring physiological sensor signals via TCP connection to the MATLAB server. The selected features from the training data are loaded and used to train the Approximate Bayes (ABayes) classification algorithm (Finseth et al., 2021). Data streamed from the Biopac physiological sensors is stored in a buffer and extracted every 30 seconds, resulting in consecutive 30-second windows. Each data window is then classified with ABayes resulting in a stress label and conditional probabilities for each class, which is sent to the Adaptive Manager.

Adaptive Manager Component

The adaptation strategies used by the adaptation manager relied on a set of rules and triggering mechanisms. Triggers involve contextual knowledge that is use to initial changes in the adaptation manager, such as temporarily changing the automation behavior (Dorneich et al., 2016). Triggering adaptations at incorrect times can make the training ineffective, or even worsen the individual's skills (Jones & Dechmerowski, 2016).

The rules and triggering mechanisms for the adaptive training system are illustrated in Figure 4. After a window is classified, the stress label and class conditional probability is sent to the adaptation manger. Adaptations were designed to trigger only after the first minute and at 30-second increments thereafter; this was intended to allow adequate time for physiological responses to stabilize and avoid rapid adaptation switching. Adaptations were not triggered until the stress had a conditional probability greater than 70% in two consecutive windows (60-seconds) to prevent triggering too frequently or with too much uncertainty. Training info about the target level is sent to the manager, with the target acting as the maximum adaptation level. This rule acts as a safeguard to prevent overwhelming users by specifying the graduated stress trial levels (Low, Medium, Medium, High, High) as the maximum achievable level, thereby preventing a disconnected sensor or faulty classification model from adapting to high stressors in every trial and resulting in negative training effects. If the current stressor levels are below the target levels, then the stress labels tell the manager how to adapt. The manager is allowed to increase stressors, but not decrease. This allow participants to experience higher VR stressors when they are ready, but instead of mitigating the stressors like a traditional biofeedback system, gives the individual time to build resilience while facing adversity.

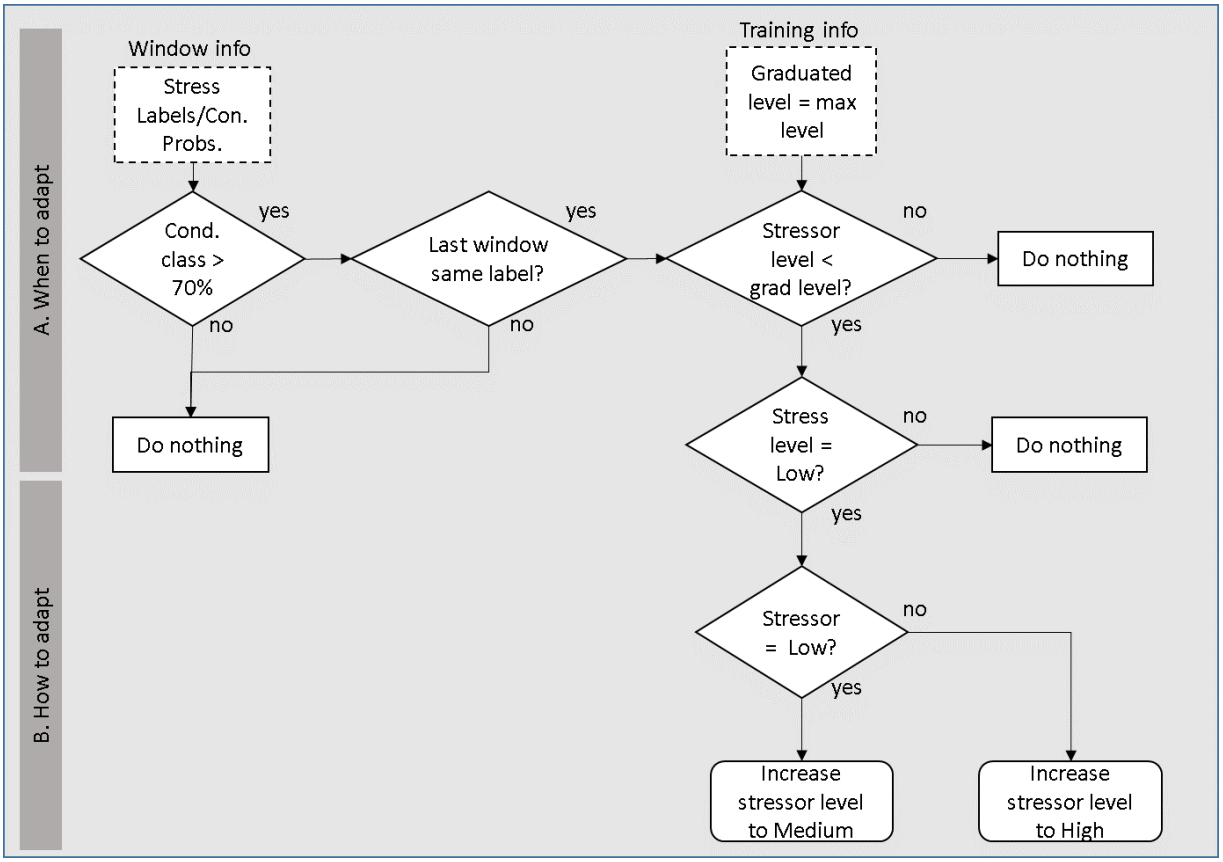

Figure 4: Decision flowchart of the rules for the adaptation manager.

Virtual Reality Component

After the adaptation has been selected by the adaptation manager, it is sent via TCP connection to another computer that is simultaneously running the VR-ISS simulation in the Unity 3D game engine (5.4.0f3, Unity Technologies, 2014). The VR-ISS then changes the environment based on the adaptation and records task performance metrics of the individual.

Methods

Participants

Sixty-five healthy subjects with academic background similar to astronauts (N= 36 males and 29 females, mean age of 28.7 years, *SD*=4.6) participated in the study. To recruit an astronaut-like population, the inclusion criteria were age range of 25-60 years old, holding (or pursuing) advanced Science, Technology, Engineering, and Mathematics (STEM) related

expertise (i.e., graduate degree or relevant industry experience). The demographic included 49% Non-Hispanic White, 36% Asian, 9% Middle Eastern or Arab, 5% Black or African American, 1% other. The protocol was approved by the Institutional Review Board (see Appendix).

Task Environment and Materials

Participants conducted a simplified ISS emergency fire response procedure in the VR-ISS with the goal of locating the source of the smoke. The simulation followed the NASA ISS emergency fire procedures, which contain instructions for crew responsibilities of measuring air contaminant levels, deciding on donning protective equipment, and identifying fireport sampling (NASA, 2013). The VR simulation consisted of three modules of the ISS. An HTC VIVE PRO head-mounted display and positioned-tracked controllers allowed for implementing zero-gravity-like locomotion. Users were asked to use contaminant level readings mimicking a NASA Compound Specific Combustion Analyzer (CSA-CP) to locate the fire's origin, then don oxygen-masks, and use fire extinguishers to facilitate extinguishing the fire (see Finseth et al., 2020). Fire location was randomized within the trials.

Experimental Design and Independent Variables

The experimental design consisted of 3 (training condition: *skill-only*, *graduated, adaptive*) x 2 (criterion trials) mixed within-between subject procedure. The training condition had three levels: (1) constant low level of stress (*skill-only*; N=20), (2) gradually increase stress levels at a fixed rate (*graduated*; N=22), and (3) real-time *adaptive* training (*adaptive*; N=23). Trial had two levels: (1) trial 3 criterion, and (2) trial 8 criterion (see Figure 5). The criterion are high-stressor simulations which were used to evaluate the within-subject change in stress response from before and after the training conditions (trial 8 - trial 3).

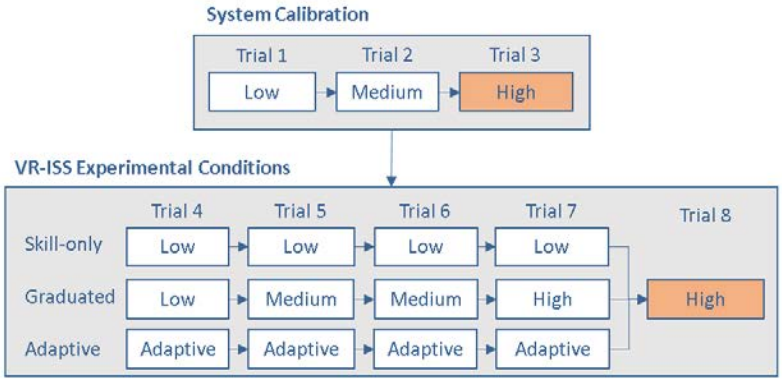

Figure 5: Trial order and experimental conditions with stressor levels. Trial 3 and trial 8 were used as criterion.

The *skill-only* condition is only exposed to a low-stressor level that is constant in the experimental trials. This low-stress training is not considered an effective form of training. The mere acquaintance with a stressor does not adequately acquaint the individual with future stressful situations (i.e., criterion) and has only been found to produce sub-par performance results compared to other training practices (Friedland & Keinan, 1992).

The *graduated* condition involves gradually increasing the stressor intensity until the highest level is achieved in the last trial (i.e., criterion). This condition is an extension of the graduated condition used in Finseth et al. (2018) which established that prior exposure to a mild-stressor lead to a reduced stress response in a subsequent high-stressor trial. Graduated exposure has been shown to be effective as long as the skill-acquisition happens prior to any stress exposure (Keinan & Friedland, 1996).

The *adaptive* condition involves closed-loop automation of gradually increasing stressors, such that automation optimally selects how to change the virtual environment based on the user's stress levels. Similar theories like the Maximal Adaptability Model (Szalma, 2008), Zone of Proximal Development (ZPD; Vygotsky, 1987; Murray & Arroyo, 2002), Yerkes-Dodson model (1908), and model of flow (Bian et al., 2019), each support the idea that there exists a 'zone of

optimal arousal' in which an individual can maximize learning and performance while avoiding being overwhelmed or underwhelmed. The *adaptive* condition tests these theories by providing physiological feedback through the simulated environment.

Dependent Variables

The study used both psychological and physiological indices of stress as follows: perceived subjective stress state and psychophysiological biomarkers of the stress response. The dependent variables are summarized in Table .

Table 1: Description of dependent variable metrics and frequencies

Dependent Variable	Sensor/Metric	Features/Type	Description, Association	Frequency
Physiological stress response	Electrocardiogram (ECG) and Heart Rate Variability (HRV)	HR	Heart rate	Before trial, throughout trial
		RMSSD	The root mean of the sum of the squares of differences between beat intervals	
		pNN50	Proportion of the successive normal to normal beat intervals that differ more than 50 ms	
		LF/HF ratio	Sympathovagal balance	
Physiological stress response	Continuous Non-Invasive Blood Pressure	SBP	Systolic Blood Pressure	Before trial, throughout trial
		DBP	Diastolic Blood Pressure	
Physiological stress response	Electrodermal Activity (EDA)	EDA	Tonic component (0 - 0.16 Hz)	Before trial, throughout trial
Psychological stress response	Post Stress Task Reaction Measure (PSTRM)	Likert scale		After trial
Psychological stress response	Short State Stress Questionnaire (SSSQ)	Likert scale		Before, after trial
Workload	NASA TLX	Likert scale		After trial
Anxiety	State-trait anxiety inventory (STAI)	Likert scale		After trial
Task Performance		Avg. distance from fire	Average distance away from the fire source.	Throughout trial
		Number of CSA-CP readings	Number of times atmospheric readings were recorded using the CSA-CP	
Coping	Coping Inventory for Stressful Situations (CISS)	Likert scale		Before experiment

The physiological stress response was measured using an electrocardiogram (ECG) sensor implemented in a Bioshirt and recorded with MP150 (Biopac System Inc., Santa Barbara, CA) equipped with an ECG100C module. The ECG signal was used to derive heart rate and a time-domain heart rate variability (HRV) metric of root mean squared of successive differences (RMSSD). Time domain analysis of the ECG was performed to quantify the amount of variance in the inter-beat-interval through the root-mean-square difference of successive normal R-wave intervals (RMSSD) and the proportion of the number of pairs of successive normal R-waves that differ by more than 50 ms (pNN50). Arousal and relaxation were indicated by changes in RMSSD and pNN50, where relaxation was marked by increasing values (vagal activation) and arousal as indicated by decreasing values (vagal inhibition). Spectral frequency analysis was conducted on HRV in the low-frequency (LF; 0.04–0.15 Hz) and high-frequency (HF; 0.15–0.4 Hz) bands. The sympathovagal (LF/HF) ratio was calculated to indicate relative activation/inhibition of the sympathetic and parasympathetic nervous systems. Systolic and Diastolic blood pressure was measured with an oscillometric non-invasive blood pressure cuff (CNAP Monitor 500, CNSystems Medizintechnik AG) which was placed on the participant's dominant upper arm and restrained with an arm sling at mid-abdomen. Participants were asked to remain seated for the entire experiment to reduce the influence of orthostatic changes. The electrodermal activity (EDA) signal was corrected with an IIR low pass 2^{nd} order Butterworth filter fixed at 5 Hz. The EDA tonic component, also called skin conductance level, was extracted by low pass filtering with a cut-off frequency of 0.16 Hz and used as an indicator of sympathetic activity (Braithwaite, Watson& Jones, 2013). Data were recorded in 30 second epochs for statistical analysis.

The psychological stress response was measured using a Post Task Stress Reaction Measure (PSTRM) and the Short State Stress Questionnaire (SSSQ). The PSTRM was used by Healey and Picard (2005). The PSTRM asks participants to rate the ground truth simulations on a scale of “1” to “9” where a rating of “1” was used to represent experiencing “no stress”, “5” was used to represent “medium stress”, and “9” was used to represent experiencing “high stress.” The PSTRM is intended to measure the immediate retrospective stress after completing a trial.

The SSSQ was administered before and after each laboratory visit to assess multiple dimensions of the subjective response to stressful environments. The SSSQ evaluates three state factors: task engagement, distress, and worry (Helton, 2004). Engagement refers to qualities of energetic arousal, motivation, and concentration. Distress is characterized by feelings of tense arousal, hedonic tone, and confidence-control. Worry relates to self-focus, self-esteem, and cognitive interference (Matthews et al., 1999). The three factor SSSQ scale scores for pre- and post-trial were calculated for each participant. The factor scores from both pre- and post-trial were standardized against normative means and standard deviation values from a large sample of British participants obtained by Matthews et al. (2002) and using the method of Helton and colleagues (Helton, 2004). Average difference scores for each state measure were then calculated for each training condition, and then used to calculate the absolute difference during each trial, resulting in a z-score.

Subjective workload of the task during exposure was measured with the NASA Taskload Index (TLX; Hart & Staveland, 1988). The NASA TLX measures six dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration level. The NASA TLX was administered after the completion of a trial. Participant scores on the six

numerical rating scales were computed in the 0 to 100 range and as an unweighted participant mean for each of the six-dimensional subscales (Nygren, 1991).

To measure the anxiety in response during the experiment, State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1983) Form Y-1 was given after the completion of a trial. The STAI-Y1 is a 20-item self-report scale for the assessment of state anxiety in adults. Based on the answers, the STAI score can be interpreted in the ranges: no stress<30, low<40, medium 40-55, high>55 (Tartarisco et al., 2015).

Task performance was measured by the participants' distance-from-fire (meters) and number of times contaminant (CSA-CP) readings were taken. A smaller distance-from-fire value and a larger number of readings suggested better performance. Data were recorded in 10 second epochs.

The Coping Inventory for Stressful Situations (CISS; Endler & Parker, 1990) scale was used to assess the degree to which an individual uses coping strategies in stressful situations, including three different coping types: task-oriented (the predisposition to deal with the problem at hand), emotion-oriented (concentration on resultant emotions such as becoming angry or sad), and avoidance-oriented (attempting to avoid the problem). A higher total score indicates greater coping abilities.

Procedure

The experiment was conducted in a single laboratory visit that lasted approximately 2-3 hours; the visit had two series of trials; (1) system calibration trials; and (2) VR-ISS experimental condition trials (Figure 6). Before participants arrived, they completed informed consent, a demographics questionnaire along with CISS, and watched a 10-minute video documenting the 1997 fire aboard the *Mir* space station for context on the severity of a fire in

space. During the visit, participants were acclimated to VR and instrumented with physiological sensors. Participants were given a 20-minute guided VR tutorial on the VR-ISS to practice conducting a fire procedure, which was rehearsed a second time without guidance. Participants indicated they fully understood how to complete the task before proceeding. Next, a 5-minute physiological baseline was taken.

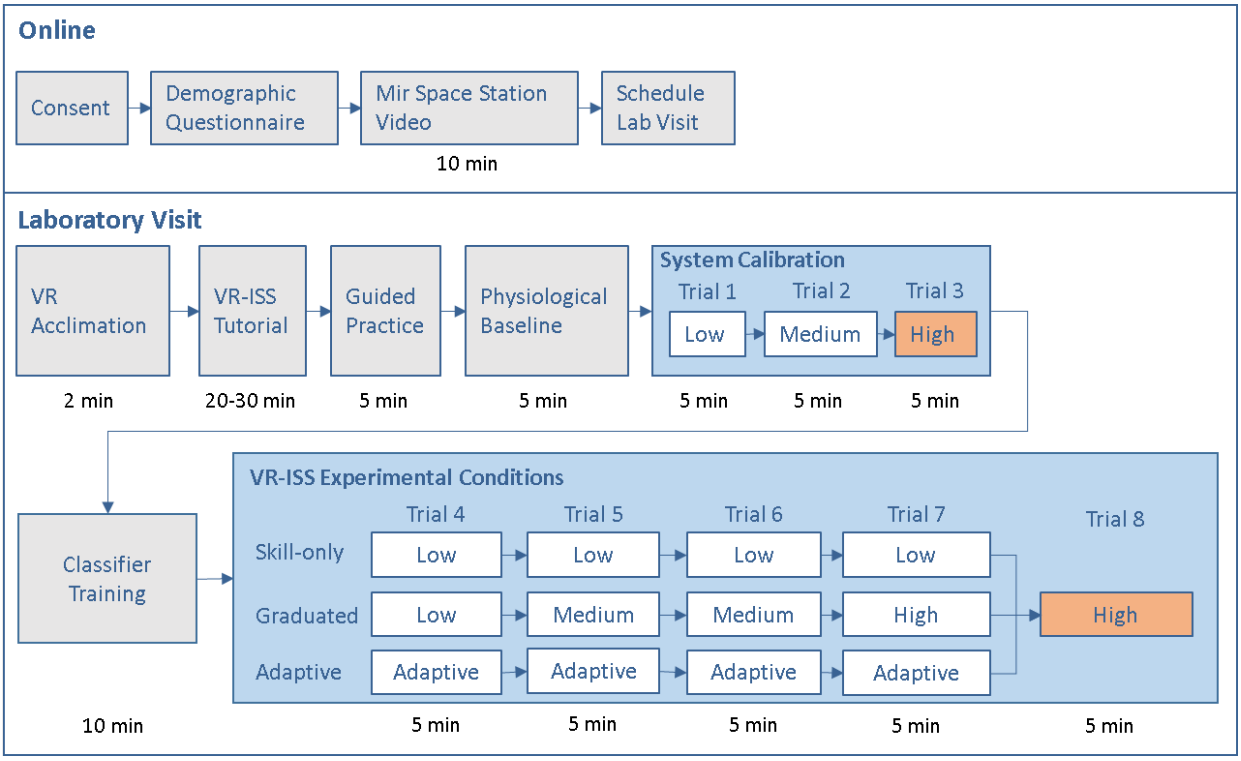

Figure 6: Experiment procedure. The two series of trials are highlighted in blue, and the trial 3 and trial 8 criterion highlighted in orange.

To calibrate the stress detection system to the participant, physiological sensor data were collected during three different trials, each with a different stressor level (i.e., low, medium, high). Each trial was a 5-minute VR-ISS emergency procedure. The recorded data were used to train a machine learning classifier to recognize the participant's stress levels. Participants were given clear indication that trial 3 was the highest intensity stressor to which they would be exposed (see Keinan & Friedland, 1996). Participants were then assigned one of three

conditions: *skill-only*, *graduated*, or *adaptive*. Participants were blind to the training condition assigned. The participants then completed five separate 5-minute trials. After each trial, participants were asked to complete the PSTRM, NASA-TLX, SSSQ, and STAI questionnaires.

Statistical Analysis

Artifacts were removed from the physiological signals if they displayed non-physiologic deflections (i.e., signal dropping to zero, spikes greater than six standard deviations from the total sample mean, or prolonged flat lines). The data were standardized for each participant using that participant's baseline. A one-sample t-test was used to analyze the change between trials (trial 8 – trial 3) for each training condition, since each training group was comprised of separate participants, and was hypothesized to have different trial effects. Therefore, separate t-test were determined to be more appropriate than an omnibus test. An analysis-of-variance (ANOVA) was used to calculate the training condition effect, as well as the coping as a continuous predictor variable (covariate) for task engagement. Multiple comparisons between conditions were adjusted with Bonferroni correction to control for type I errors (Bland & Altman, 1995). Results were considered significant for $p \leq 0.05$ and were considered marginal for $0.05 < p \leq 0.1$ (Gelman, 2013). Effect size calculations followed Cohen's *d* guidelines reported as small for $0.2 < |d| < 0.5$, medium for $0.5 < |d| < 0.8$, and large for $|d| > 0.8$ (Cohen, 1988).

Results

To verify that the stress manipulation resulted in differing stress levels used to train the system's classification model, subjective stress results are present on trials 1-3. Psychological response, physiological response, and task performance are reported for the training conditions.

Stress Manipulation

The subjective stress measured by the PSTRM between trials 1-3 are illustrated in Figure 7. The main effect of trial on subjective stress was significant, $F(2,122) = 78.4$, $p < .001$, $d = 2.27$. Pairwise comparisons showed that subjective stress was significantly higher during trial 3 (M=5.84, SD=1.17) compared to trial 2 (M=4.67, SD=0.99), p>.001, d=1.08, significantly higher for trial 3 compared to trial 1 (M=3.69, SD=0.89), p>.001, d=2.07, and significantly higher for trial 2 compared to trial 1, p>.001, d=1.04. The main effect of training condition on subjective stress was not significant, $F(2,61) = 0.357$, $p = .702$. The interaction effect (trial*condition) was not significant, $F(4,122) = 1.32$, $p = .268$.

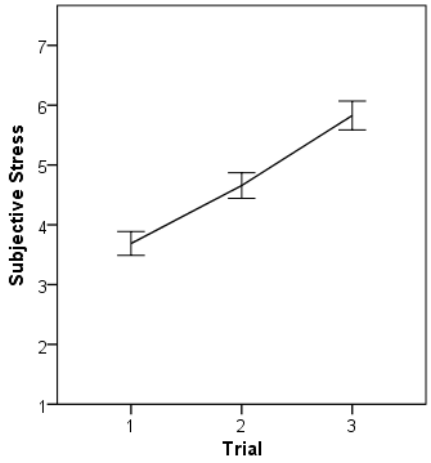

Figure 7: Subjective stress between trials 1-3. Errors bars in standard error.

Psychological Response

The subjective stress measured by the PSTRM for training conditions in trial 3 and trial 8 are illustrated in Figure 8. The change in subjective stress between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=-0.45, SD=1.76), $t(19) = -1.14$, $p = .27$, and *graduated* (M=-0.19, SD=1.94), $t(20) = -0.45$, $p = .66$, but the change was marginally less than zero for *adaptive* (M=-0.45, SD=1.18), $t(21) = -1.8$, $p = .086$, d=0.38. The effect of training

condition on subjective stress was not significant, $F(2,60) = 0.18$, $p = .84$. Pairwise comparisons between training conditions were not significant.

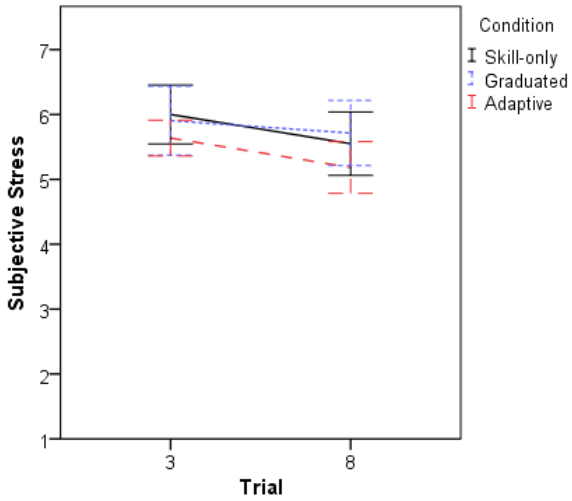

Figure 8: Subjective stress for training conditions between trial 3 and trial 8. Errors bars in standard error.

The task engagement measured by the SSSQ for training conditions in trial 3 and trial 8 are illustrated in Figure 9. The change in task engagement between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=0.05, SD=0.66), $t(19) = 0.72$, $p = .72$, marginally less than zero for *graduated* (M=-0.21, SD=0.53), $t(20) = -1.79$, $p = .089$, d= 0.39, and significantly greater than zero for *adaptive* (M=0.25, SD=0.46), $t(21) = 2.51$, $p = .021$, d= 0.53. The effect of training condition on task engagement was significant, $F(2,60) = 3.65$, p = .032, d= 0.85. Pairwise comparisons between training conditions indicated that *adaptive* had a significantly greater change than *graduated* (p=.027), but was not significantly different than control (p=.79), nor was *graduated* significantly different than control (p=.41).

The distress measured by the SSSQ for conditions in trial 3 and trial 8 are illustrated in Figure 10. The change in distress between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=-0.33, SD=0.89), $t(19) = 0.12$, $p = .12$, marginally less than zero for *graduated* (M=-0.25, SD=0.65), $t(20) = -1.79$, $p = .089$, d=0.39, and marginally less than zero for

adaptive (*M*=-0.23, *SD*=0.64), *t*(21) = -1.73, *p* = .098, d=0.37. The effect of training condition on distress was not significant, *F*(2,60) = 0.091, *p* = .91. Pairwise comparisons between training conditions were not significant.

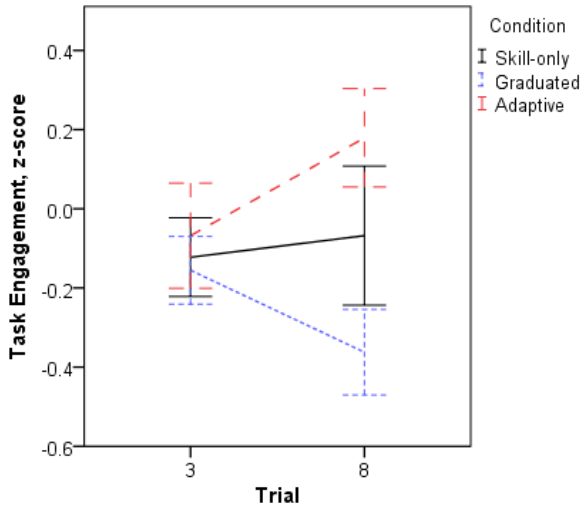

Figure 9: Task engagement for conditions between trial 3 and trial 8. Errors bars in standard error.

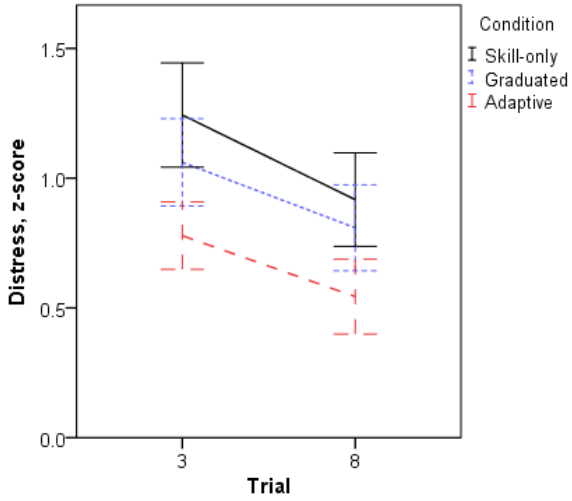

Figure 10: Distress for training conditions between trial 3 and trial 8. Errors bars in standard error.

The worry measured by the SSSQ for training conditions in trial 3 and trial 8 are illustrated in Figure 11. The change in worry between trials (trial 8 – trial 3) was significantly less than zero for *skill-only* (*M*=-0.28, *SD*=0.32), *t*(19) = -3.94, *p* = .001, d=0.88, *graduated* (*M*=-

0.23, *SD*=0.42), *t*(20) = -2.5, *p* = .021, d=0.55, and *adaptive* (*M*=-0.21, *SD*=0.46), *t*(21) = -2.19, *p* = .04, d=0.47. The effect of training condition on distress was not significant, *F*(2,60) = 0.146, *p* = .87. Pairwise comparisons between training conditions were not significant.

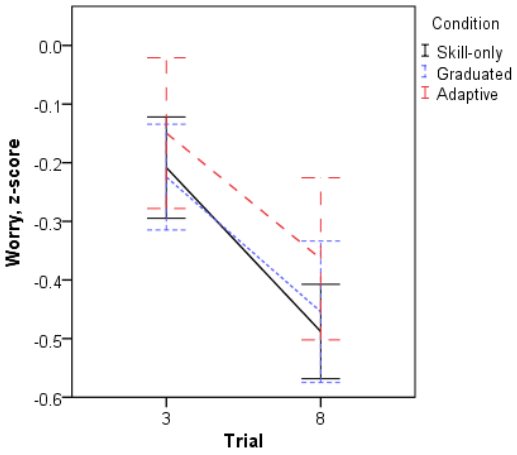

Figure 11: Worry for training conditions between trial 3 and trial 8. Errors bars in standard error.

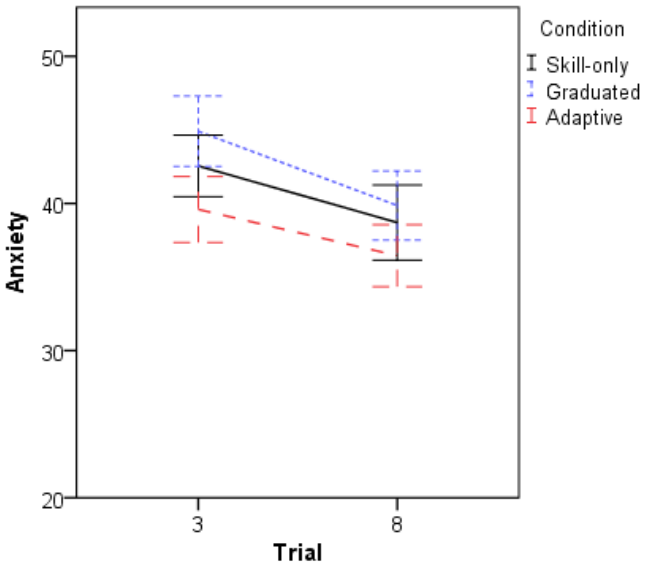

Figure 12: Anxiety for training conditions between trial 3 and trial 8. Errors bars in standard error.

The anxiety measured by the STAI for training conditions in trial 3 and trial 8 are illustrated in Figure 12. The change in anxiety between trials (trial 8 – trial 3) was marginally less than zero for *skill-only* (*M*=-3.85, *SD*=9.75), *t*(19) = -1.77, *p* = .094, *d*=0.40, while being

significantly less than zero for *graduated* (*M*=-5.05, *SD*=7.43), *t*(20) = -3.11, *p* = .005, *d*=0.68, and *adaptive* (*M*=-3.14, *SD*=5.98), *t*(21) = -2.46, *p* = .023, *d*=0.53. The effect of training condition on anxiety was not significant, *F*(2,60) = 0.33, *p* = .72. Pairwise comparisons between training conditions were not significant.

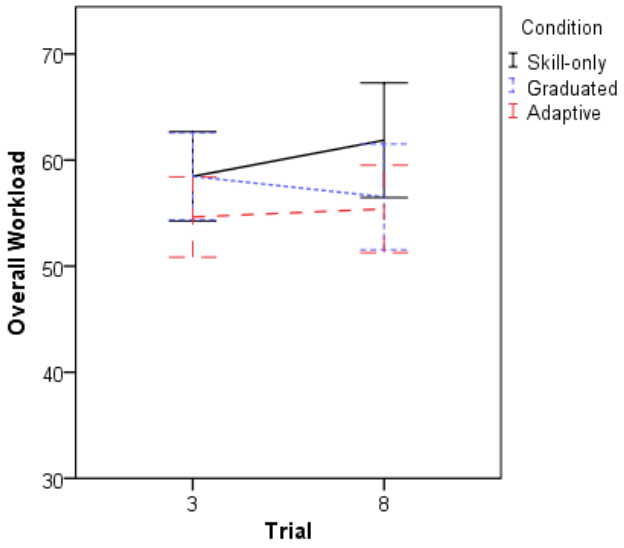

Figure 13: Workload for training conditions between trial 3 and trial 8. Errors bars in standard error.

The workload measured by the NASA-TLX for training conditions in trial 3 and trial 8 are illustrated in Figure 13. The change in workload between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (*M*=3.4, *SD*=13), *t*(19) = 1.17, *p* = .26, *graduated* (*M*=-1.93, *SD*=20.2), *t*(20) = -0.44, *p* = .67, and *adaptive* (*M*=0.76, *SD*=19.9), *t*(22) = 0.18, *p* = .86. The effect of training condition on workload was not significant, *F*(2,60) = 0.45, *p* = .64. Pairwise comparisons between training conditions were not significant.

Physiological Response

The heart rate for training conditions in trial 3 and trial 8 are illustrated in Figure 14. The change in heart rate between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (*M*=-0.86, *SD*=3.27), *t*(15) = -1.05, *p* = .31, or *graduated* (*M*=-0.84, *SD*=2.71), *t*(17) =

-1.29, p = .21, but significantly less than zero for *adaptive* (M=-1.52, SD=1.66), t(18) = -4, p = .001, d=0.93. The effect of training condition on heart rate was not significant, F(2,50) = 0.419, p = .66. Pairwise comparisons between training conditions were not significant.

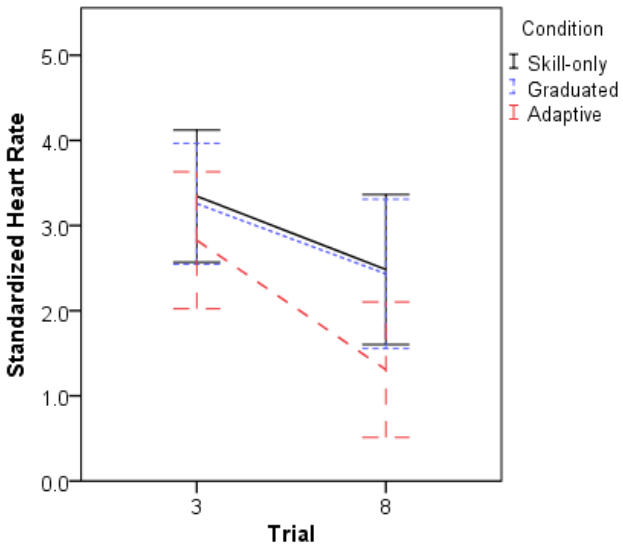

Figure 14: Standardized heart rate between trial 3 and trial 8. Errors bars standard error.

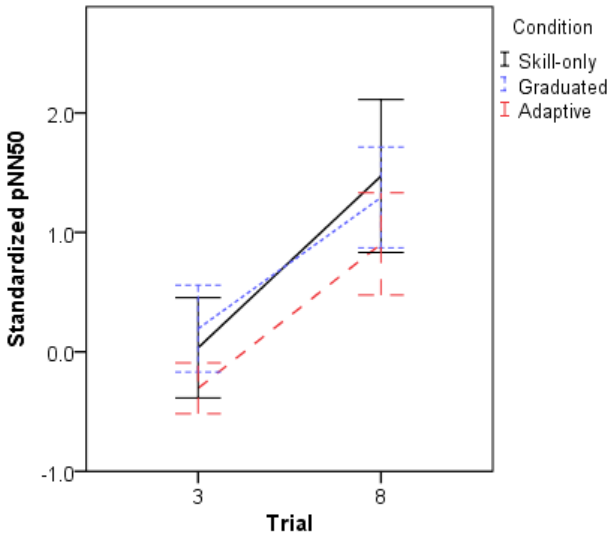

Figure 15: Standardized pNN50 between trial 3 and trial 8. Errors bars standard error.

The pNN50 for training conditions in trial 3 and trial 8 are illustrated in Figure 15. The change in pNN50 between trials (trial 8 – trial 3) was significantly greater than zero for *skill-only* (M=1.44, SD=1.53), t(14) = 3.64, p = .003, d=0.94, *graduated* (M=1.10, SD=2.15), t(19) = 2.28,

p = .034, d=0.5, and *adaptive* (*M*=1.21, *SD*=2.23), *t*(19) = 2.42, *p* = .026, d=0.54. The effect of training condition on pNN50 was not significant, *F*(2,52) = 0.13, *p* = .89. Pairwise comparisons between training conditions were not significant.

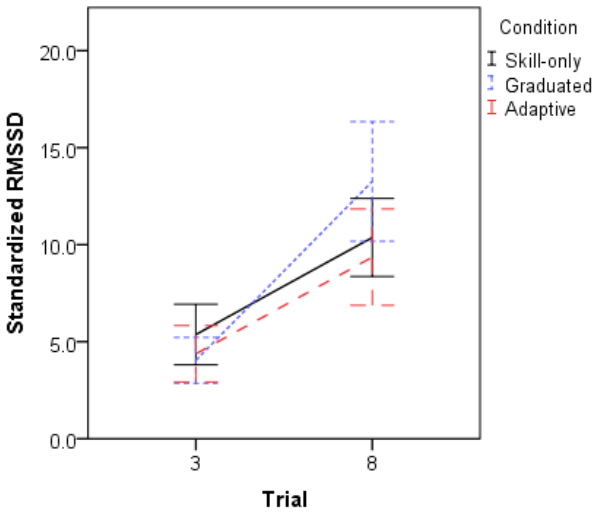

Figure 16: Standardized RMSSD between trial 3 and trial 8. Errors bars standard error.

The RMSSD for training conditions in trial 3 and trial 8 are illustrated in Figure 16. The change in RMSSD between trials (trial 8 – trial 3) was significantly greater than zero for *skill-only* (*M*=5.3, *SD*=6.2), *t*(17) = 3.63, *p* = .002, d=0.85, *graduated* (*M*=9.22, *SD*=15.41), *t*(19) = 2.68, *p* = .015, d=0.6, and marginally greater than zero for *adaptive* (*M*=5.11, *SD*=12.61), *t*(20) = 1.86, *p* = .078, d=0.4. The effect of training condition on RMSSD was not significant, *F*(2,56) = 0.72, *p* = .49. Pairwise comparisons between training conditions were not significant.

The LF/HF ratio for training conditions in trial 3 and trial 8 are illustrated in Figure 17. The change in LF/HF between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (*M*=-0.14, *SD*=0.72), *t*(19) = -0.85, *p* = .41, or *graduated* (*M*=-0.084, *SD*=0.41), *t*(20) = -0.93, *p* = .36, but marginally less than zero for *adaptive* (*M*=-0.28, *SD*=0.61), *t*(19) = -2.04, *p* = .056, d=0.46. The effect of training condition on LF/HF was not significant, *F*(2,58) = 0.59, *p* = .56. Pairwise comparisons between training conditions were not significant.

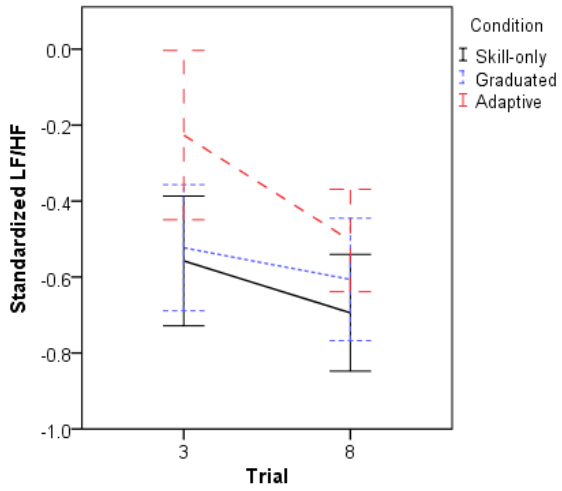

Figure 17: Standardized LF/HF ratio between trial 3 and trial 8. Errors bars standard error.

The SBP for training conditions in trial 3 and trial 8 are illustrated in Figure 18. The change in SBP between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=-1.43, SD=5.50), $t(13) = -0.98$, $p = .35$, *graduated* (M=-1.09, SD=4.41), $t(19) = -1.11$, $p = .28$, or *adaptive* (M=-1.01, SD=3.25), $t(19) = -1.39$, $p = .18$. The effect of training condition on LF/HF was not significant, $F(2,51) = 0.042$, $p = .96$. Pairwise comparisons between training conditions were not significant.

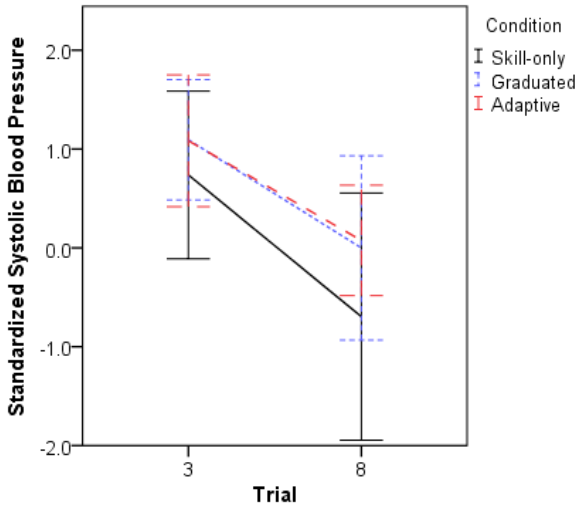

Figure 18: Standardized SBP between trial 3 and trial 8. Errors bars standard error.

The DBP for training conditions in trial 3 and trial 8 are illustrated in Figure 19. The change in DBP between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=-1.43, SD=5.50), $t(13) = -0.98$, $p = .35$, *graduated* (M=-1.09, SD=4.41), $t(19) = -1.11$, $p = .28$, or *adaptive* (M=-1.01, SD=3.25), $t(19) = -1.39$, $p = .18$. The effect of training condition on LF/HF was not significant, $F(2,51) = 0.042$, $p = .96$. Pairwise comparisons between training conditions were not significant.

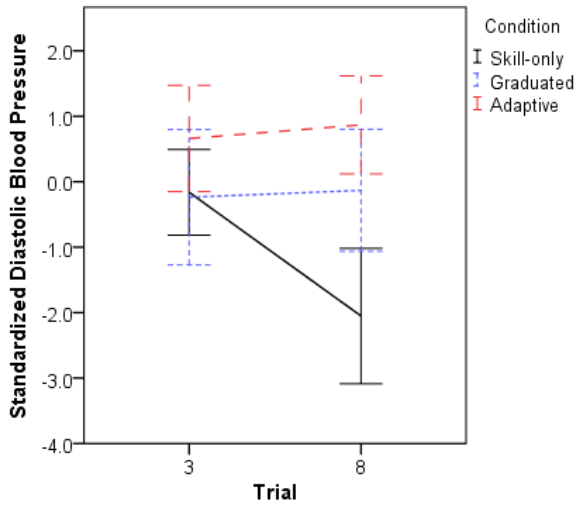

Figure 19: Standardized DBP between trial 3 and trial 8. Errors bars standard error.

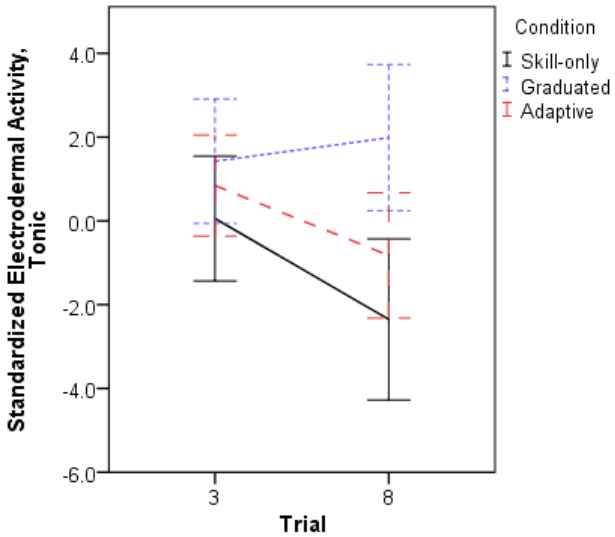

Figure 20: Standardized EDA tonic level between trial 3 and trial 8. Errors bars standard error.

The EDA tonic level for training conditions in trial 3 and trial 8 are illustrated in Figure 20. The change in EDA tonic between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=-2.41, SD=6.16), $t(11) = -1.36$, $p = .20$, *graduated* (M=0.56, SD=3.93), $t(15) = 0.57$, $p = .58$, or *adaptive* (M=-1.67, SD=7.31), $t(18) = -1.0$, $p = .33$. The effect of training condition on EDA tonic was not significant, $F(2,44) = 0.97$, $p = .39$. Pairwise comparisons between training conditions were not significant.

Task Performance

The distance-from-fire for the training conditions in trial 3 and trial 8 are illustrated in Figure 21. The change in distance-from-fire between trials (trial 8 – trial 3) was not significantly different than zero for *skill-only* (M=-0.75, SD=2.85), $t(18) = -1.14$, $p = .27$, *graduated* (M=0.32, SD=3.59), $t(20) = 0.41$, $p = .69$, or *adaptive* (M=-0.47, SD=3.9), $t(21) = -0.56$, $p = .58$. The effect of training condition on distance-from-fire was not significant, $F(2,59) = 0.51$, $p = .61$. Pairwise comparisons between training conditions were not significant.

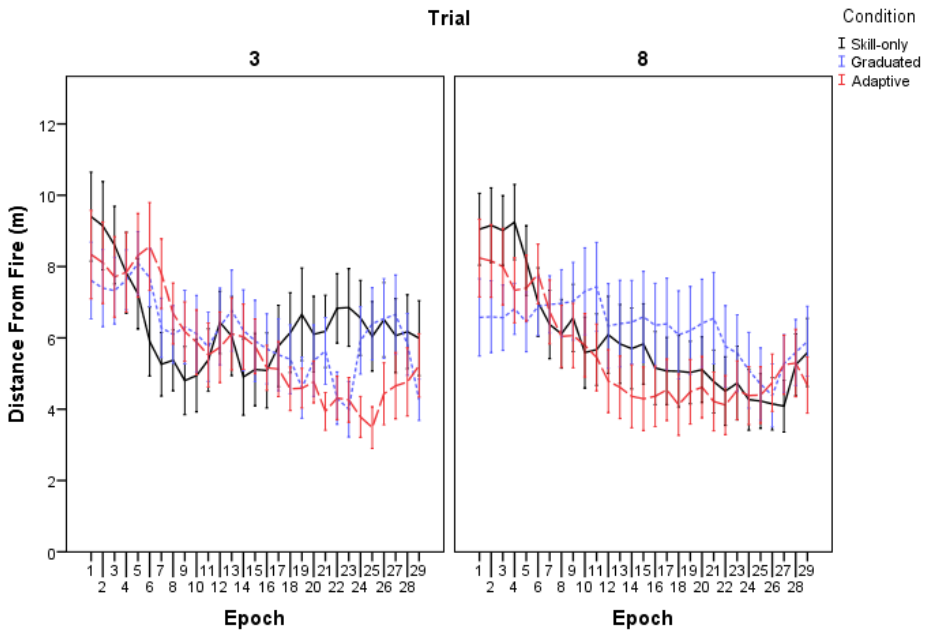

Figure 21: Distance from fire, recorded at 10-sec epochs, between trial 3 and trial 8. Errors bars standard error.

The number of contaminate (CSA-CP) readings for the training conditions in trial 3 and trial 8 are illustrated in Figure 22. The change in the number of contaminant readings between trials (trial 8 – trial 3) was significantly greater than zero for *skill-only* (*M*=9.53, *SD*=9.31), *t*(18) = 4.46, *p* = .27, d=1.02, and *adaptive* (*M*=4.95, *SD*=6.8), *t*(21) = 3.42, *p* = .003, d=0.73, but not significantly different for *graduated* (*M*=3, *SD*=13.6), *t*(20) = 1.01, *p* = .32. The effect of training condition on the change in the number of contaminant readings was not significant, *F*(2,59) = 2.1, *p* = .13. Pairwise comparisons between training conditions were not significant.

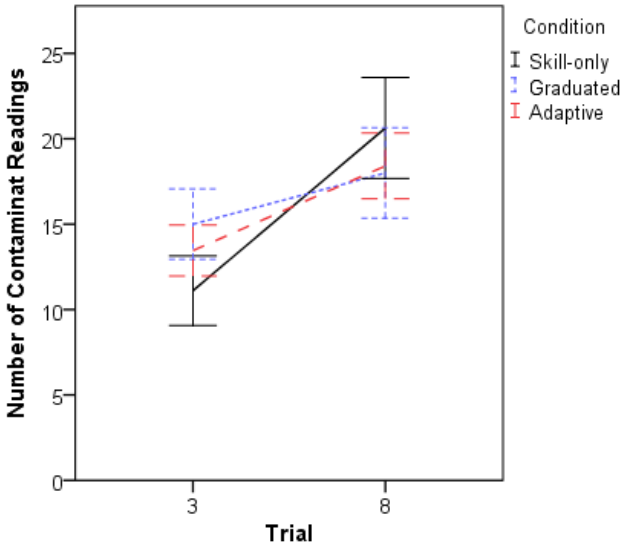

Figure 22: Number of contaminant (CSA-CP) readings between trial 3 and trial 8. Errors bars standard error.

Coping

To examine differences between training conditions and task-engagement, analyses were performed using coping as a continuous predictor variable. Task-oriented coping was not a significant predictor, *F*(1,57)= 0.26, *p*=.87, and the training condition main effect was not significant , *F*(2,57)= 1.87, *p*=.16. Emotion-oriented coping was not a significant predictor, *F*(1,57)= 0.59, *p*=.45, and the training condition main effect was not significant , *F*(2,57)= 2.08, *p*=.13. Avoidance-oriented coping was not a significant predictor, *F*(1,57)= 0.32, *p*=.57, while

the training condition main effect was marginal, $F(2,57)= 2.49$, $p=.092$, but the interaction effect was not significant, $F(2,57)= 0.93$, $p=.22$, $d=0.46$,even though the effect size was medium.

Adaptations

The stressor level adaptations compared to *skill-only* and *graduated* conditions are illustrated in Figure 23. The *adaptive* condition was designed such that the adaptations only triggered during trials 5, 6, and 7. Further, the adaptations during those trials adjusted the stressor level between the target level (*graduated*) and the minimum level (*skill-only*). In the *adaptive* training condition, 43% of participants had an adaptation in trail 5, 43% in trial 6, and 61% in trial 7. If the low, medium, and high stressor levels were assigned numbers 1-3, the average stressor level for trial 5 was 1.35 (*SD*= .48), for trial 6 was 1.47 (*SD*= .50), and for trial 7 was 2.0 (*SD*= .75).

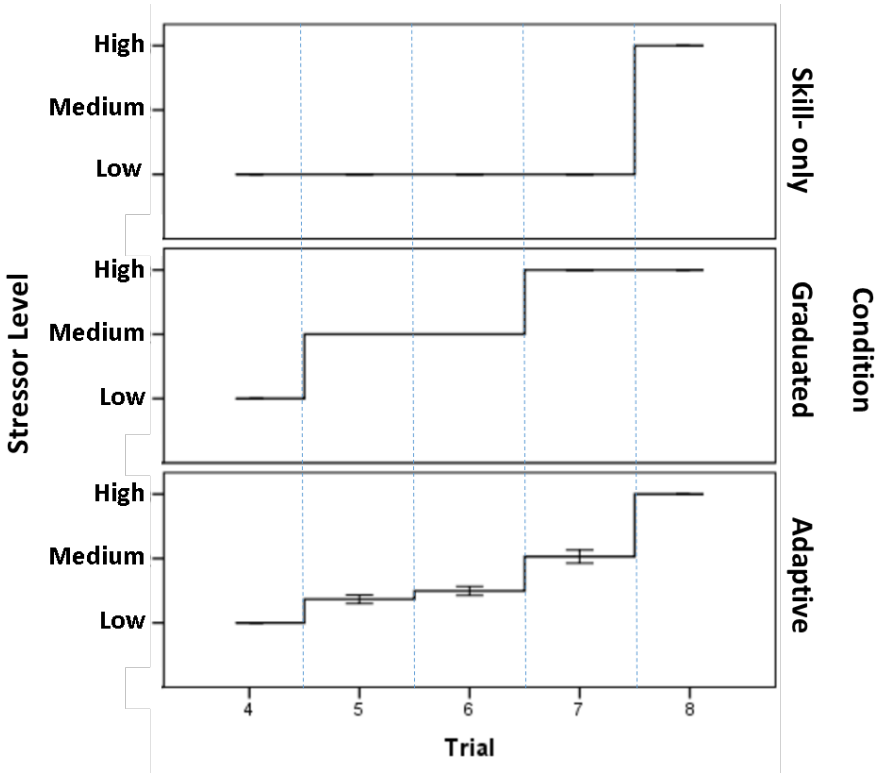

Figure 23: Stressor Level adaptation compare to the skill-only and graduated conditions. Error bars in standard error.

The adaptive participants showed different adaptation profiles, illustrated in Figure 24. Three participants did not have any adaptations, thus resembling *skill-only* participants. Five participants had adaptations in both trial 5 and 7, thus resembling the *graduated* participants.

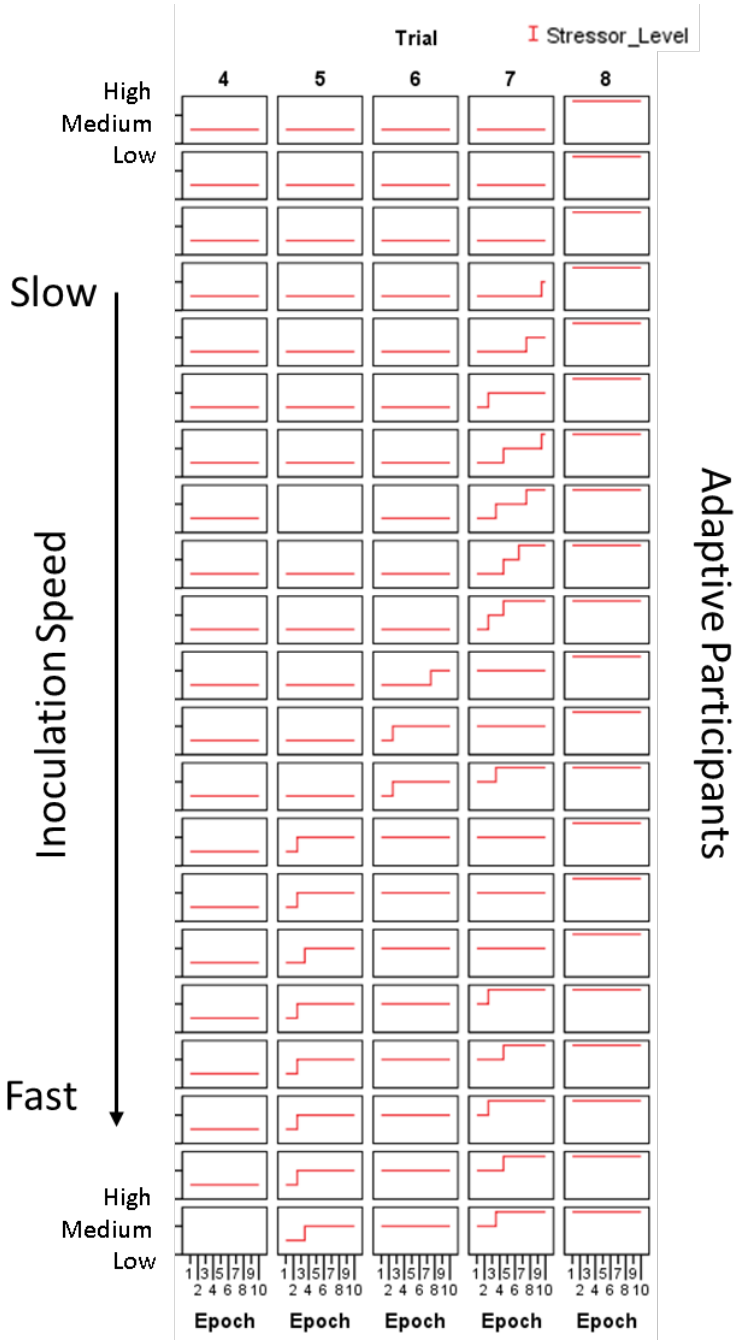

Figure 24: Stressor Level adaptation profiles, with participants in order by the speed of inoculation.

Summary of Results

A summary of the results are listed in Table 2 for quick reference.

Table 2: Inferential statistics summary for the experimental conditions. P-value (effect size).

	Change between trial, one-sample t-test			Condition effect, univariate ANOVA with multiple comparison			
DV	**Skill-only**	**Graduated**	**Adaptive**	**Condition Effect**	**Skill-only vs. Graduated**	**Graduated vs. Adaptive**	**Adaptive vs. Skill-only**
Psychological Response							
Subjective Stress	.27	.66	.086 (0.38)	.82			
Task Engagement	.72	.089 (0.39)	.021 (0.53)	.032 (0.85)	.41	.027 (0.38)	.79
Distress	.12	.089 (0.39)	.098 (0.37)	.89			
Worry	.001 (0.88)	.021 (0.55)	.04 (0.47)	.75			
Anxiety	.094 (0.40)	.005 (0.68)	.023 (0.53)	.77			
Workload	.26	.67	.86	.65			
Physiological Response							
Heart Rate	.43	.16	.001 (0.93)	.82			
pNN50	.003 (0.94)	.034 (0.5)	.026 (0.54)	.89			
RMSSD	.003 (0.85)	.015 (0.60)	.078 (0.40)	.53			
LF/HF	.41	.36	.056 (0.46)	.66			
SBP	.35	.28	.18	.80			
DBP	.19	.93	.85	.24			
EDA tonic	.20	.58	.33	.30			
Task Performance							
Distance from fire	.27	.69	.58	.61			
Contaminant Readings	<.001 (1.02)	.32	.003 (0.73)	.13			

Discussion

This study evaluated whether an adaptive system could be more effective in training than either skill-only training or graduated exposure. The results show all conditions reduced stress, but through the preponderance of psychological and physiological trial effects, the adaptive reduced stress significantly across more measures. Task performance was significantly higher for the *skill-only* and *adaptive*, but not the *graduated* training condition.

In a physiological context, the *adaptive* system's sympathetic (LF/HF) withdrawal and decrease in heart rate is more compelling evidence of stress inoculation than the parasympathetic changes (e.g., RMSSD, pNN50). This is supported by research showing that an induced acute mental stress can cause an asymmetrical shift of the autonomic system balance during which, the sympathetic activation overcomes the parasympathetic, and in turn reinforcing higher LF/HF values and higher stress (Visnovcova et al., 2014). All three experimental conditions increase in RMSSD and pNN50 across trials. Specifically, this suggests participants had increased vagal control (i.e., parasympathetic activation), which helped relax individuals. The LF/HF ratio and heart rate showed that only the *adaptive* condition significantly reduced stress between trial 3 and 8. As the LF/HF ratio represents sympathovagal balance, and there was no difference found in parasympathetic magnitude (RMSSD, pNN50) between conditions, this suggests that the sympathetic stress response was lower for the *adaptive* condition between trials. These measures together suggest the *adaptive* condition had the largest inoculation, although the *skill-only* and *graduated* training did show modest benefits.

The psychological response showed more supporting results that favor the *adaptive* condition. The subjective stress showed a marginal decrease across trials for the *adaptive* condition, along with the distress which also showed a marginal decrease for both the *adaptive* the *graduated* condition. Further, anxiety was decreased for the *graduated* and *adaptive* conditions, but only marginally for the skill only. Attentional Control Theory (Eysenck & Derakshan, 2011) proposes that anxiety interferes with executive control, specifically by allocating cognitive resources to manage task-irrelevant stimuli resulting in poorer-working memory (Matthews et al. 2013). This suggests that the *adaptive* condition's decrease in subjective stress, distress, and increased task engagement was able to provide more executive

control and cognitive resources for the development of emotional regulation strategies to counteract stress. In contrast, the *graduated* condition had a marginal decrease in distress, but also had a marginal decrease in task engagement. Decreasing distress may suggest more control over reorienting to spatial cues, whereas the decrease in task engagement may suggest individuals may be disengaging from the simulation to adaptively reduce stress or recover from fatigue, which is a form of avoidance coping (Matthews et al. 2013). Matthews et al. (2013) also showed that task engagement only has a medium correlation to task performance, which is possibly why contaminate readings did not improve across trials. However, avoidance coping was not found to be a moderator for the task engagement (although the analysis was likely statistically underpowered, even though the interaction effect was medium). Accounting for the marginal decrease in distress, decrease in task engagement, and lack of change in sympathetic measures, these results suggest that participants in the *graduated* condition may have been overwhelmed by the stressor level increases, whereas participants in the *adaptive* condition were able to see beneficial results in stress reduction.

The *graduated* condition task performance did not improve. A possible explanation is that the *graduated* condition represents a fixed schedule of stressor increases, of which the participants may not have been ready to handle. These changes in stressors may have been poorly timed and may account for the lack of change between trial 3 and 8 for this condition. The objective of the *adaptive* system is to personalize the training by timing changes based on the current stress levels for the participant, and thus improve stress inoculation.

The results suggest that the *skill-only* condition was less likely to inoculation stress, with the change from before and after potentially reflecting the relief that the experimental trials were almost finished. This idea is supported by the marginal decrease in anxiety (with a small effect

size), decrease in worry across trials (with a large effect), increase in relaxation measure of pNN50 and RMSSD, and lack of other findings suggests that the *skill-only* condition may have found relief that they would not have to face another intense stressor trial. The increase in task performance may be a result of repetitive task-training. The goal of some repetitive task-training techniques are to train a skill with a no-stress or a low-level stress beyond the initial level of proficiency in order to build up automatic responses (Driskell et al., 2008). Repetitive task training may explain why task performance improved, but *skill-only* did not see effects from inoculation during a high-stressor level.

In the context of training with stress, the development of emotional regulation strategies may be influenced by the availability of cognitive resources and capability to exert self-control. Self-control requires mutually exclusive beneficial options, with the individual intent on furthering a more valued goal over another: one that provides instant gratification, while the other furthers more long-term valued goals (Inzlicht & Schmeichel, 2012; Duckworth et al., 2016). Unlike situations with unescapable stressors, most stress exposure during training is either optional and/or temporary. Therefore, it becomes a self-control problem of either avoidance coping (i.e., disengagement or quitting from the VR training) to remove the stressor and provide instant gratification, or continuing stress exposure to build self-regulation for the long-term prospects of resilience. However, research has shown that self-control depletes cognitive resources and requires more resources to initiate in subsequent situations (Inzlicht & Schmeichel, 2012). Particularly, stronger impulses may overbear attempts at self-control. The magnitude of stressor used in training should allow for some stress, but not enough that the option to avoid and disengage becomes the dominant impulse. Because the alternative to avoid or disengage from training is so easy in VR, it is unlikely that sensitization to stress would occur. Negative

affect/emotions would only reinforce the self-control selection instant gratification to remove the stressor (i.e., stop training altogether). Consequently, this intention of self-control may determine whether the individual may become more resilient to stress during training with repeated exposure.

The adaptation strategies, rules, and triggers employed by the *adaptive* manage showed positive results. The *adaptive* system was able to trigger adaptations for trials 4-7 and provide a middle ground between the control and *graduated* condition. Further, most participants only showed an adaptation triggering once per trial allowing time for the physiological response to stabilize. In reviewing the adaptation profiles, some participant displayed signs of slower inoculation whereas other displayed sign of faster inoculation. It is recommended that future studies with a larger sample compare the faster and slower inoculators to understand further adaptation strategies that may help these conditions. Lastly, this study only evaluated simple environmental adaptations with constrained rules; it may be beneficial to evaluate rules that allow more diverse adaptations (e.g., increasing and decreasing stressor levels, adapting task procedures or information delivery) and triggering.

Conclusion

Stress training has shown benefits for building resilience to high-stress situations. Preliminary results suggest that preparing individuals for future stressors with an adaptive system can increase task performance. Further research is needed to validate these findings. Astronauts may someday use adaptive systems to train for emergency fires. These systems may also be used for other high-stress occupations such as military, police, firefighters, and aircraft pilots.

Acknowledgments

This work was funded by the National Aeronautics and Space Administration (grant number 80NSSC18K1572). For their help developing the VR-ISS software, the authors thank Jacob Liebman, Peter Carlson, and Robert Slezak.

References

Bian, D., Wade, J., Swanson, A., Weitlauf, A., Warren, Z., & Sarkar, N. (2019). Design of a Physiology-based Adaptive Virtual Reality Driving Platform for Individuals with ASD. *ACM Transactions on Accessible Computing (TACCESS), 12*(1), 2. https://doi.org/10.1145/3301498

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *Bmj, 310*(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

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.

Dorneich, M. C., Rogers, W., Whitlow, S. D., & DeMers, R. (2016). Human performance risks and benefits of adaptive systems on the flight deck. *The International Journal of Aviation Psychology, 26*(1-2), 15-35. https://doi.org/10.1080/10508414.2016.1226834

Driskell, J. E., Salas, E., Johnston, J. H., & Wollert, T. N. (2008). Stress exposure training: An event-based approach. *Performance under stress*, 271–286. London: Ashgate. https://doi.org/10.1037/10278-007

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2016). Situational strategies for self-control. *Perspectives on Psychological Science, 11*(1), 35-55. https://doi.org/10.1177%2F1745691615623247

Endler, N. S., & Parker, J. (1990). *Coping inventory for stressful situations.* Multi-Health systems Incorporated.

Eysenck, M. W., & Derakshan, N. (2011). New perspectives in attentional control theory. *Personality and Individual Differences, 50*(7), 955-960. https://doi.org/10.1016/j.paid.2010.08.019

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2). https://doi.org/10.1037/1528-3542.7.2.336

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: a framework for researchers and system designers. *Human Factors, Vol. 54*, No.6, 1008-1024. https://doi.org/10.1177/0018720812443983

Finseth, T., Dorneich, M. C., Keren, N., Franke, W. D., & Vardeman, S. (2020, October). Designing Training Scenarios for Stressful Spaceflight Emergency Procedures. In *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)* (pp. 1-10). IEEE. https://doi.org/10.1109/dasc50938.2020.9256403

Finseth, T., Dorneich, M. C., Vardeman, S., Keren, N., & Franke, W. D. (2021). *Physiologically Based Stress Detection for Hazardous Operations Using Approximate Bayes Algorithm* [Manuscript submitted for publication]. Industrial Manufacturing and Systems Engineering, Iowa State University.

Finseth, T. T., Keren, N., Dorneich, M. C., Franke, W. D., Anderson, C. C., & Shelley, M. C. (2018). Evaluating the Effectiveness of Graduated Stress Exposure in Virtual Spaceflight Hazard Training. *Journal of Cognitive Engineering and Decision Making*, *12*(4), 248-268. https://doi.org/10.1177/1555343418775561

Friedland, N., & Keinan, G. (1992). Training effective performance in stressful situations: Three approaches and implications for combat training. *Military Psychology*, *4*(3), 157-174. https://doi.org/10.1207/s15327876mp0403_3

Gaume, A., Vialatte, A., Mora-Sánchez, A., Ramdani, C., & Vialatte, F. (2016). A psychoengineering paradigm for the neurocognitive mechanisms of biofeedback and neurofeedback. *Neuroscience & Biobehavioral Reviews*, *68*, 891-910. https://doi.org/10.1016/j.neubiorev.2016.06.012

Gelman, A. (2013). P values and statistical practice. *Epidemiology* (Cambridge, Mass.), *24*(1), 69–72. https://doi.org/10.1097/ede.0b013e31827886f7

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2019). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing.* https://doi.org/10.1109/taffc.2019.2927337

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in psychology, 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Healey, J., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, *6*(2), 156-166.

Helton, W. (2004). Validation of a Short Stress State Questionnaire. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(11), 1238–1242. Los Angeles, CA: SAGE Publications. https://doi.org/10.1177/154193120404801107

Inzlicht, M., & Schmeichel, B. J. (2012). What is ego depletion? Toward a mechanistic revision of the resource model of self-control. *Perspectives on Psychological Science, 7*(5), 450-463. https://doi.org/10.1177%2F1745691612454134

Johnston, J. H., & Cannon-Bowers, J. A. (1996). Training for stress exposure. In J. E. Driskell & E. Salas (Eds.), Series in applied psychology. *Stress and human performance* (p. 223–256). Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203772904-15

Jones, D., & Dechmerowski, S. (2016). Measuring Stress in an Augmented Training Environment: Approaches and Applications. In *International Conference on Augmented Cognition* (pp. 23-33). Springer, Cham. https://doi.org/10.1007/978-3-319-39952-2_3

Keinan, G., & Friedland, N. (1996). Training effective performance under stress: Queries, dilemmas, and possible solutions. *Stress and human performance*, 257–277. Mahwah, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9780203772904-16

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping.* New York, NY: Springer-Verlag. https://doi.org/10.1017/s0141347300015019

Matthews, G., Campbell, S., Falconer, S., Joyner, L., Huggins, J., Gilliland, K., Grier, R., & Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion, 2*(4), 315. https://doi.org/10.1037/1528-3542.2.4.315

Matthews, G., Joyner, L., Gilliland, K., Campbell, S. E., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire: Towards a state big three. *Personality psychology in Europe, 7*, 335-350.

Matthews, G., Szalma, J., Panganiban, A. R., Neubauer, C., & Warm, J. S. (2013). Profiling task stress with the dundee stress state questionnaire. *Psychology of stress: New research, 1*, 49-90.

Meichenbaum, D. (1985). *Stress Inoculation Training.* New York: Pergamon.

Murray, T., & Arroyo, I. (2002, June). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In *International conference on intelligent tutoring systems* (pp. 749-758). Springer, Berlin, Heidelberg.

National Aeronautics and Space Administration (NASA). (2013). International Space Station, emergency procedures 1a (No. JSC-48566). Houston, TX: NASA Johnson Space Center.

National Aeronautics and Space Administration (NASA). (2021). Human Research Roadmap: Human Factors and Behavioral Performance: CBS-BMed1. Retrieved from https://humanresearchroadmap.nasa.gov

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human factors, 33*(1), 17-33. https://doi.org/10.1177/001872089103300102

Pallavicini, F., Argenton, L., Toniazzi, N., Aceti, L., & Mantovani, F. (2016). Virtual reality applications for stress management training in the military. *Aerospace medicine and human performance, 87*(12), 1021-1030. https://doi.org/10.3357/amhp.4596.2016

Russi-Vigoya, M. N., Dempsey, D., Munson, B., Vera, A., Adelstein, B., Wu, S. C., & Holden, K. (2020, July). Supporting Astronaut Autonomous Operations in Future Deep Space Missions. In *International Conference on Applied Human Factors and Ergonomics* (pp. 500-506). Springer, Cham. https://doi.org/10.1007/978-3-030-50943-9_63

Szalma, J. L. (2008). Individual differences in stress reaction. *Performance under stress,* 323-357. London: Ashgate.

Spielberger, C.D., Gorsuch, R.L., Lushene, R., Vagg, P.R. and Jacobs, G.A. (1983) *Manual for the State-Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Tartarisco, G., Carbonaro, N., Tonacci, A., Bernava, G.M., Arnao, A., Crifaci, G., Cipresso, P., Riva, G., Gaggioli, A., De Rossi, D., & Tognetti, A. (2015). Neuro-fuzzy physiological computing to assess stress levels in virtual reality therapy. *Interacting with Computers, 27*(5), pp.521-533. https://doi.org/10.1093/iwc/iwv010

Visnovcova, Z., Mestanik, M., Javorka, M., Mokra, D., Gala, M., Jurko, A., Calkovska, A., & Tonhajzerova, I. (2014). Complexity and time asymmetry of heart rate variability are altered in acute mental stress. *Physiological measurement, 35*(7), 1319. https://doi.org/10.1088/0967-3334/35/7/1319

Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (eds.)., The collected works of L. S.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology, 18*(5), 459-482.

Appendix. Approval for Research (IRB)

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 05/09/2019

To: Tor Finseth Michael Dorneich, Ph.D.

From: Office for Responsible Research

Title: **Astronaut training system for Virtual Reality Spaceflight Emergency**

IRB ID: **19-184**

Submission Type: Initial Submission **Review Type**: Full Committee

Approval Date: 05/09/2019 **Approval Expiration Date: 05/08/2020**

CHAPTER 8. LESSONS LEARNED & RECOMMENDATIONS

The lessons learned from the design, implementation, and evaluations have been listed, along with recommendations for researchers developing similar training or adaptive systems. The recommendations have been sorted into categories that reflect different areas of research: graduated stress exposure, stress manipulation during training or VR, stress detection with machine learning based on physiological signals, adaptation rules and triggering by the adaptive system, and skill acquisition prior to the stress exposure.

Lessons Learned

Graduated Stress Exposure

1. **Graduated exposure is a beneficial training component in Stress Inoculation Training (SIT), providing some amount of stress inoculation even without Phase 1 and 2.**

   Issue: Limited evidence exists for using graduated training (SIT phase 3) by itself to reduce stress and enhance training. Specifically, how the separate components of SIT contribute differentially to appraisal, biological arousal, and training effectiveness during occupational tasks (Saunders et al., 1996; Robson & Manacapilli, 2014).

   Lesson/Finding: Results from Study 1 (CHAPTER 4.) found that gradually increasing exposure can positively affect the responses to exposure to a more severe version of the stressor. The physiological and psychological measures indicated that a group receiving prior exposure had relatively less activation of the sympathetic nervous system, enhanced allostasis, and adaptability to a stress response in a subsequent trial. These results suggest that prior exposure to a low-level stressor attenuates the sympathovagal response to a more stressful condition of the same task in virtual reality.

Benefit: This research provides support that stress inoculation can develop passively, with the potential to build resilience.

Future Work: While the findings provided support that graduated training (SIT phase 3) by itself can reduce stress, future work could assess the impact of including SIT phase 1 of education and phase 2 of skill acquisition in conjunction with phase 3 of graduated training in order to provide maximum stress inoculation benefits. Further, the lessons taught in phase 1 and phase 2 may help participants distinguish positive vs. negative coping strategies and reduce the risk of inadvertently developing a maladaptive strategy and negative training effects.

Future work is needed on long-term retention and stress inoculation over longer durations (days/weeks). Future work is needed to test different methods to deliver stressors) to see if more pronounced improvements in stress response could be obtained.

2. **Graduated exposure is a useful method to help healthy individuals decrease future stress during tasks with highly-demanding procedures and high levels of stress.**

Issue: SIT has been shown to be successful for clinical therapy applications. However, research is needed on stress training for healthy people working in challenging environments (Rose et al., 2013).

Lesson/Finding: This research found that healthy participants who received prior exposure to a stressor decreased the stress response when exposure to a similar future stressor and enhanced their ability to remain calm in a spaceflight emergency fire simulation.

Benefit: These findings provide support that graduated exposure (phase 3 of SIT) may be a useful method for training individuals to manage stress in hazardous operations.

Future Work: More work is needed to understand how specific tasks influence the amount of inoculation obtainable. While stress exposure was found to be effective for a spaceflight emergency fire procedure, other tasks may be more or less effective at producing stress inoculation.

3. **Testing multiple trials and varying time between trials may help to optimize the stress inoculation for the specific task/stressors.**

Issue: There is limited evidence for how many trials are needed or amount of time between trials to result in stress inoculation. Saunders et al. (1996) found that 6-7 trials could help reduce performance anxiety, while 4-5 five trials helped reduce state anxiety. There is even less guidance on the amount of time between trials, but it is possible that too short of duration between trial would not be adequate for learning new skills/coping, but too long of a duration would result in participants forgetting the training. This may also vary depending on the specific task and complexity of skill/coping.

Lesson/Finding: This research found that graduated exposure worked for two trials for a spaceflight procedure (Study 1, CHAPTER 4.).. A short 5-10 minute duration between trials was sufficient to see an effect of stress inoculation. Similarly, the research also found that five trials of graduated exposure showed a stress reduction (Study 5, CHAPTER 7.).

Benefit: These findings provide support that physiological stress inoculation can happen in 2-5 trials for a specific task.

Future Work: Future work is needed on the resolution of stressor levels, number of sessions, session duration, and the duration between sessions. Further, work is needed on quantifying the positive benefits that can accrue from a range in the number training

sessions with graduated exposure to determine when the user reaches diminishing returns (Saunders et al., 1996).

4. **Fatigue can have detrimental effects on training outcomes; thus researchers should be mindful of the duration and timing of training intervals.**

   Issue: Training multiple trials of graduated stress exposure may take a long time, which can lead to the trainee becoming fatigued. Both task performance and emotional self-regulation take considerable cognitive resources, which may result in fatigue and performance decrements (Matthews et al., 2013).

   Lesson/Finding: This research found that the graduated training condition, where the training was not personalized to the current participant stress level but instead used a predetermined gradual increase in stressors, participants over five trials showed lower task engagement and no change in performance, which can be an indicator of fatigue (Study 5, CHAPTER 7. ; Matthews et al., 2013).

   Benefit: Including measure of fatigue in future experiments may help to identify if participants have enough energy to proceed into the next training trial.

   Future Work: More research is needed on how to provide stress training without fatiguing participants, as well as the duration of time required in-between training session to recover cognitive resources.

Stress Manipulation

5. **Environment-based stressor manipulations can manipulate the users' stress response while keeping the task constant.**

   Issue: Manipulating task difficulty to increase stress is a potentially confound when designing training simulations, because there is no clear indication if the simulation or the task load is primarily responsible for changes in the stress response. For example, if

procedures are allowed to change between graduated exposure trials, and the participant chooses a less complex procedure during a subsequently higher-stress trial, there would be uncertainty in whether a resulting reduction in stress is attributed to inoculation or lower procedure complexity.

Lesson/Finding: Stressors can be categorized as coming from the task and the environment (Study 2, CHAPTER 5.). To keep task difficultly the same between levels/groups, stressors within the training environment can be changed to produce different stimuli intensity. This results in isolating the source of the stress response while making the task constant between individuals/groups. However, keeping environment and task stressor strictly separate is a challenge. If the procedure steps diverge from other levels/groups, there is a potential for differing task load, which may alter the amount of stress the individual experiences.

Benefit: By controlling for the task difficulty, it is easier to quantify the independent effectiveness of graduated exposure.

Future Work: Future work should investigate how much a procedure can deviate within a specific task while not resulting in changes to the stress response. Training programs may be trying to adapt the training environment to best meet the needs of the user, such as providing stressors that are within the abilities of the individual. Isolating the stress response as a product of the environment stressor will help inform how to increase/decrease stressors, whereas too much procedure deviation could result in task load that significantly alters the anticipated amount of stress that as environmental stressor may elicit.

6. **Stressors should be identified through subject matter experts, operators, interviews, of documented case studies.**

Issue: It is often unclear which stressors should be selected for the training simulations and which have the largest effects on users doing tasks in real-life situations. Some stressors are specific to the event (e.g., smoke during emergency fire training) and therefore harder to identify as a potential influence on individuals' response.

Lesson/Finding: When designing the VR simulations for this research, the researchers found that the spaceflight stressors identified by system designers may be different from those identified from operators who have experienced the real-life situation (Study 2, CHAPTER 5.).

Benefit: If proper stressors are identified, they will be more relevant to the task environment, and may increase training transfer to real-life.

Future Work: More research is needed on how to identify and measure the effect that certain stressor has on an individual, such as a questionnaire to evaluate the task environment or a list of possible stressors and their characteristics. Simulation designers may also benefit from the creation of a database of stressors relevant to real-life procedures, how to manipulate them, their type, their impacts, and relevant domains.

7. **Experiments can use VR to elicit different stress states.**

Issue: Graduated training requires stressor levels to increase for each new trial. In order to conduct graduated stress exposure training in VR, it is important to establish that changing aspects of the VR can induce different levels of stress.

Lesson/Finding: This research found that stressors can be manipulated in the VR environment to evoke different levels of stress in participants for a spaceflight procedure (Study 2, CHAPTER 5.). Specifically, changes were found in subjective measures of distress, post task stress, workload, and relative stress between levels.

Benefit: These findings provide support that manipulation of stress can be achieved with VR. Successful manipulations then allow the simulation to be used for graduated stress exposure and other exposure programs.

Future Work: More research is needed on the amount of stress that can be induced in VR and the comparability to that of stress experienced in a similar real-world situation/event. Further, work is needed to evaluate if more than three levels can elicit a reliable stress response and if more levels may make participants more comfortable during level transitions. Lastly, more research is needed across domains and how stressor levels can be created to train during different tasks.

8. **Measures other than workload may be needed to assess taskload during gradually increasing stressor levels.**

Issue: Stress may not be completely separate from workload when doing task manipulation. It is possible that by increasing stress, the availability of cognitive resources was reduced for task processing, therefore increasing the amount of workload (or effort) required (Eysenck et al., 2007). Subjective workload may be influenced by stress, as it is a multi-dimensional construct and impacted by task engagement (Matthews et al., 2015).

Lesson/Finding: The results from testing graduated exposure (Study 1, CHAPTER 4.) found changes in workload, measured with NASA Task Load Index, between trials (but not groups) were associated with changing environment and not the taskload. Similarly, some task performance measures show changes between trials, but it is unclear if the repetitive training or the stress inoculation is responsible for the changes.

Benefit: Including more measures of task load that are not easily influenced by stress may help clarify the interaction between workload, stress, and task load.

Future Work: More research is needed to develop measurement instruments that can objectively be used to measure task load between different tasks and evaluated on their reliability to help identify the amount of mental workload generated from the task under stressful conditions.

9. **Stressors should be tested in isolation and in combination.**

Issue: Multiple stressors may be involved in the operational environment. However, when designing training simulations, it is unclear what the magnitude of each environmental stressors should be, or the combination of different stressors (e.g., noise, smoke). Further, there is some research that suggests physical and psychosocial stressors can have cumulative effects that are greater than the individual effects of the stressors alone (Abdelall et al., 2020).

Lesson/Finding: When designing the VR simulations for this research, the researchers found that the stressor combinations for the three levels where not linear combinations (Study 2, CHAPTER 5.). The magnitude of some of the stressors had to be adjusted in VR multiple times. In early stages of testing, pilot study results showed that the low and medium stressor levels had very similar stress responses across participants, so the smoke intensity for the medium stressor was increased and the alarm noise for the low stress simulation was decreased.

Benefit: Testing stressors separately and in combinations grants greater predictability of the effects of stressors and more precise manipulations of participant stress.

Future Work: More research is needed on which specific stressors are more likely to have cumulative effects. Further, how the stressor cumulative effects can vary between individuals.

Stress Detection

10. Use multiple classes for stress detection while minimizing error rate.

Issue: Stress is indicated by multiple continuous variables with many different states, which does not naturally map into two levels (i.e., stress or no-stress). By using binomial classification, the physiological states that are not extremes introduce error through the assumption that state can be classified as stressed vs. not stressed (e.g., classifying a low-level arousal with a binomial class of stressed vs. not stressed). However, using too many classes with the machine learning model will lower the accuracy and need more training data to ensure reliable stress predictions. A balance needs to be made between including multiple levels necessary for the application and minimizing the error rate.

Lesson/Finding: This research found that three levels of stress can be accurately detected, and that multiple stress classes offered finer classification model resolution (Study 3, CHAPTER 6.).

Benefit: Including multiple classes in a model may better represent the continuous scale of the stress response and allow for more measurement resolution. This enables an adaptive system to provide feedback or make changes that is more tailored to the user's current stress levels.

Future Work: Future work should evaluate the tradeoff between adding classes and classification error rate. Further, research should evaluate how many classes are necessary to get a reliable scale of how the person is responding to stressors.

11. Stress detection should be personalized (or calibrated) based on the individual's subjective or physiological stress.

Issue: Stress responses vary among individuals and depend on the individuals' appraisals of the stressor; therefore, individualized models may be more accurate than generalized

classifiers (Smets, Raedt, & Van Hoof, 2019; Gjoreski et al., 2017). Different physiological systems are activated across participants in response to the same stressor (Bowers et al., 2008).

Lesson/Finding: The findings from this research show that each person has their own set of physiological indicators that are discriminative for levels of stress (Study 3, CHAPTER 6.). A classifier is personalized when the model accounts for that person's response either through well-labeled stress states to train/calibrate the model (i.e., ground truth for supervised machine learning) or periodic calibration of the model.

Benefit: Implementing personalized models will make the detection more robust, useful for a wider audience, and will have higher accuracy in detecting stress.

Future Work: More work is needed on how to identify key features that characterize a personal stress response, how those features vary over time, and how to quickly collect ground truth using the stress response. Further, more research is needed on personalized models such as using data from an individual to create a classifier but then calibrating the model over time with that specific person's feedback.

12. Time-series data must use a method that takes into account the temporal correlations in the continuous data before classification.

Issue: Machine learning algorithms assume that the data are independently and identically distributed (Verleysen & François, 2005). However, the time series nature of physiological signals creates temporal correlations, which will result in severely biased classification accuracy if not addressed either by the data structure or the classification model.

Lesson/Finding: Several methods exist for time series classification, of which methods that calculate signal characteristic for time-intervals were found to be an effective way to

quantify discriminative properties (Bagnall et al., 2017). Results from the current research show that interval methods adjusted for the temporal correlations leading to less biased results than signals with data dependencies (Study 3, CHAPTER 6.).

Benefit: Time-series methods are more likely to capture signal patterns leading to higher accuracy.

Future Work: More research is needed on how interval methods can be calibrated for stress response variation over time.

13. **In order to have confidence in the conditional probabilities output by a classifier, researchers should consider the assumptions that each machine learning algorithms is making and how those assumptions may affect the conditional probabilities.**

Issue: Traditional machine learning algorithms make data-driven approximations of class probabilities. However, these approximations can have varying accuracy and use indirect methods to approximate conditional probabilities. For example, decision tree classifiers produce rectangles that partition the input space and calculate the approximate class probabilities based on the number of vectors located within each rectangle. Thus, the class probability is constant for each rectangle and always discontinuous at the rectangle boundaries, leading to a probability that is more defined by how the rectangles are positioned within the input-space rather than the vector distribution across the entire input-space. Similarly, Support Vector Machines (SVMs) create a hyper-planes intended to produce maximum separation between class vectors in the input space. Ad hoc "approximate class probabilities" are often created using normalized probability distributions based on the estimate hyperplane position—a practice that may not match empirical probability estimates (Zadrozny & Elkan, 2002). Therefore, hyperplanes and

rectangles can sometimes be positions in irrational ways, which can jeopardize the interpretability of how the stress predictions are generated.

Lesson/Finding: The ABayes method was developed to produce conditional probabilities of all three classes, providing an indication of the confidence in its classification. ABayes was comparable to other classifiers and had good performance at classifying three stress levels, but has more direct estimates of conditional probability which give more interpretability if the stress prediction is reliable.

Benefit: Increasing the interpretability of how the model creates probability distributions from the data allows the designer to assess the reliability when deployed. For this reason, classification models that have less-complex rules (e.g., ABayes, SVM) are favored for high-risk tasks (e.g., analysis of aircraft engine lifecycles) whereas more complex models (e.g., random forest, neural nets) that have more specification uncertainty.

Future Work: Future work should evaluate the difference in class conditional probabilities between ABayes and traditional classifiers.

Adaptive Stress Training System

14. **An adaptive system is an effective way to implement stress training over multiple sessions with the goal of reducing stress.**

    Issue: Training with constant low-levels of stress (or no stress) may reinforce skill development but lead to performance problems in a high stress environment. In contrast, gradually increasing stress may prepare for those stressors. However, graduated stress is applied incrementally but not explicitly tailored to the individual's needs (Friedland & Kenian, 1996).

    Lesson/Finding: The adaptive system developed in this work was able to reduce stress over multiple trials utilizing multiple physiological and psychological measures. No

stress exposure and fixed graduated stress exposure pedagogies both elicited some reduction in stress, but had a smaller effect than individuals who used the adaptive system where the graduated stress exposure was personalized to the participant's current level of stress (Study 4, CHAPTER 7.).

Benefit: Training that autonomously changes to meet the users' needs may be easier to implement and make faster assessments of the individuals current stress state than training practices that require direction from a trainer. This may lead to better physiological outcomes while requiring less training support and resources.

Future Work: More research is needed to develop the adaptive stress training to autonomously troubleshoot issues such as sensor disconnects and restarting simulations without direct supervision from a trainer.

15. Evaluate each component of a closed loop adaptive system before testing the full system.

Issue: A closed-loop adaptive system relies on multiple components to provide feedback and or adapt its behavior. However, debugging the system can be challenging when it is unclear which component is causing the problems. For example, a poorly functioning adaptive system may be due to either good adaptations that trigger at the wrong time, or the bad adaptations that are triggered appropriately (Ververs et al., 2005).

Lesson/Finding: Testing that the components have the desired effect is important before integration. This research tested that the training module was able to reduce stress over multiple trials (Study 1, CHAPTER 4.), tested the stress manipulations (Study 2, CHAPTER 5.), tested that the stress detection was able to accurate classify multiple levels of stress (Study 3, CHAPTER 6.), then integrated all the working components

together to test that the system could adapt the VR environmental stressors in real-time to reduce user stress (Study 4, CHAPTER 7.).

Benefit: This process mitigates risk of component error/failure and increases the likelihood that the system will work as intended.

Future Work: Future work needs to address the limitations of the individual components, such as the enhancing the stress detection accuracy, expanding the adaptations rules, and further replicating the emergency procedure task in a VR environment.

16. Training systems should avoid frequent and rapid adaptations.

Issue: Frequent adaptations can be volatile and not allow enough time for the user to adjust (Fuchs et al., 2008). However, how much time should be placed between adaptations is dependent on the type of task. For stress training, adjusting training quickly may be better for participant needs, but physiological system need time to respond to the changing stress levels.

Lesson/Finding: This research has found that 30 seconds was an appropriate window for adaptations, giving ample time for physiological systems to respond, while offering enough time for participants (Study 3, CHAPTER 6. ; Study 4, CHAPTER 7.).

Benefit: Longer durations between adaptations can allow users to transition between states. Further having consistency in the timing of adaptations creates more transparency in and allows users to interpret how the automation is behaving and a prediction of how it will behave in the future.

Future Work: More research is needed on different adaptation intervals within physiological adaptive systems.

17. Safeguards should be created in the adaptation rules to make the system more robust when there is unforeseeable system errors.

Issue: Adaptation rules and triggers may perform as intended, however other system components may have unforeseeable issues. If the adaptation rules have too much flexibility, system wide errors caused by signal noise and incomplete information could result in in adaptations that do not match the users current state, thus lead to negative training effects. For example, a disconnected heart rate sensor could lead to an erroneous stress prediction that would correctly trigger adaptations, but still result in the training overwhelming the user with stress.

Lesson/Finding: The adaptation rules or this adaptive system acted as a safeguard and resulted in stress inoculation (Study 4, CHAPTER 7.). The rules were constrained to only allow the maximum stressor level to be the graduated exposure level, while also not allowing decreasing stressor levels.

Benefit: Adaptation safeguards increased the reliability that the system behaves as intended by minimize the risk from adapting in the wrong conditions.

Future Work: More research is needed on how rules effect stress inoculation and what rules support psychological theories about building resilience.

Skill Acquisition and Task Training Prior To Stress Exposure

18. **Before graduated exposure training, provide participants with context of the environmental demand by the stressor and potential consequences, thus allowing the participant to mimic how they would respond in real-life situation.**

Issue: Researchers may not always be able to evaluate new approaches using astronaut participants. Often the approach is to use an “astronaut-like” population: 50/50 gender split, ages 25-55, and STEM trained. However, this population does not have the experiences of astronauts and are novices in spaceflight operations. While they can be trained on the tasks, novice users may not have the experience to understand the

consequences of the threat cues they may see. Thus, novice users may apply different coping and decision-making strategies than they would in a real high-stress scenario (Gamble et al., 2018). Participants can be unsure how to feel about the stress within the training, how to gauge the seriousness of a fire during spaceflight, and consequently, the purpose of training.

Lesson/Finding: During skill acquisition, before the stress exposure or use of the adaptive system, a video depicting the emotional reaction from people in real-life scenarios provides users with a mental model of how dangerous the worst outcome can be, and how they may interact when they encounter that same stressor later in the graduated exposure (Study 4, CHAPTER 7.).

Benefit: Providing context improves the participant's ability to appraise the situation. Users will have a better ability to replicate emotional responses and have a greater commitment to follow the procedures.

Future Work: More research is need on what types of information can allow participants to develop similar appraisals in training as in the real-life scenario.

19. Researchers should customize the training tutorials to deliver information based on the specific needs of the participant.

Issue: Participants may have different learning styles which influences how quickly they process and remember information, as well as what kinds of information they process. During skill acquisition (i.e., training tutorial, before the individual experienced the stress exposure), some participants grasp task very quickly, struggled with learning movement, or ask questions frequently during the narrated tutorial in VR.

Lesson/Finding: Participants have different learning styles, which may affect how they how well they acquire new skills, the rate at which information is learned, and their

overall task proficiency. Training needs to be customized not only during the stress exposure and use of adaptive system, but also in the training tutorials (Study 4, CHAPTER 7.).

Benefit: Help the training generalize across participants who may have different learning styles. By customizing how the task procedure is taught (e.g., information delivery, trainer guided compared to video or VR tutorial) participants may reach skill proficiency faster.

Future Work: Need a method to assess and compensate for different learning styles. This could include different ways to deliver information, pacing, adaptive tutoring, or separating skills into different components.

20. Researchers should ensure that participants have acquired and become proficient in relevant skills before any stress exposure occurs.

Issue: Stress exposure can interfere with learning the task and therefore negatively impact stress and performance. Separate phases allow the trainee to practice and maximize skill proficiency without being subject to any stress, and then subsequently allow the trainee to perform their new skills under the influence of stressors (Friedland & Keinan, 1992; Keinan & Friedland, 1996). SIT/SET establishes that skill training relevant to stress exposure is conducted in Phase 2. These skills can include cognitive control strategies, physiological control strategies, overlearning, mental practice, decision-making and team skills (Driskell et al., 2008). After successfully acquiring the skills needed to perform effectively under stress, skills can be practiced in Phase 3 with gradually increase sing stress.

Lesson/Finding: Stress can interfere with learning new skills. Some participants did not fully understand the procedure or how to move adequately before moving into the stress

exposure sessions, causing frustration and extra stress (Study 4, CHAPTER 7.). Participants have different proficiency levels, even after multiple practice sessions. Extra training may be necessary to solidify an understanding of the emergency procedure, especially if more procedure branches are practiced during training.

Benefit: Separating skill acquisition may help identify participants that require more skill training before stress exposure. Ensure participant learn the skills, minimize stress interference with learning new skills, and stress response is not confounded by frustration from skill level.

Future Work: Evaluate different methods to identify which participants require extra training. Researchers need to develop a measure of task proficiency that is specific to their training application, including ability to move and interact with the training simulation, before proceeding to graduated exposure. This may prevent negative training effects from occurring once stress exposure has begun.

Recommendations for Future R&D: Roadmap

The experimental findings from the research provide insight into future research and development (R&D) directions. Results demonstrate that the technology developed has advantages and disadvantages with regards to practical training application of NASA astronauts. In addition, the lessons learned provided earlier in this section include possible area so future work. Table 3 summarizes the challenges and potential research questions with regards to development, system implementation, and training by comparing the current maturity of the stress detection and adaptive stress training. Findings from this table were then used to generate recommendations for future R&D projects.

Table 3: Challenges for development, implementation, and training compared for Adaptive Stress Training and Stress Detection

#	Challenges	Category	Discussion /Research Question
1	Personalizing Training	Adaptive stress training, Stress detection	The adaptive stress training requires upfront setup to position physiological sensors and calibrate the personalized stress detection model, but the training runs autonomous once the setup is complete. To further increase the ease of use, more research is needed now how to personalize the stress detection by minimizing the amount of model information needed for multiple baseline trials at different stressor levels (i.e., stress labeled data for supervised machine learning). Tackling this challenge would make both the stress detection and adaptive stress training easier to use.
2	Sensitivity to sensor noise	Adaptive stress training, Stress detection	The SD used by the adaptive stress training relies on physiological sensor input that is prone to classifying signal artifacts, which can create lower the accuracy of classification. More research is needed on how to filter the noise and prevent classification on data that does not represent change in the stress response.
3	Ground-based or/ in-flight training	Adaptive stress training, Stress detection	Both the stress detection and adaptive stress training can be implemented on ground-based and in-flight in their current form. Since the stress detection is personalized, the stress detection would be calibrated for on-orbit physiological changes. However, more development is needed to increase the stress detection and adaptive stress training mobility such as portable wireless sensors and VR headsets.
4	Amount of effort needed to complete training	Adaptive stress training	Users need to be proficient in the task procedure before using the adaptive stress training system. In its current form, there are 5-minutes sessions that can be conducted in a series. However, users may become fatigued over multiple sessions. More research is needed on how many sessions can be performed while maintaining optimal effort before a long break is needed.
5	Amount of Stress Inoculation	Adaptive stress training	Research findings show that using the adaptive stress training results in stress inoculation over multiple training sessions. However, it is currently unknown how many sessions can be trained before diminishing returns. Further, it is unclear how long the stress inoculation can be retained over long periods of time.
6	Oversight needed	Adaptive stress training	The adaptive stress training requires minimal oversight to implement the training. However, the oversight can be further optimized by more development to reduce system errors and automate the technology troubleshooting (e.g., easy restart if a sensor disconnects or hand controller dies mid-training session).
7	Types of skills that can be trained	Adaptive stress training	Since the adaptive stress training is created to supplement task training, insight into skills for reducing stress and maintaining performance can be taken from both Stress

			Inoculation Training (SIT) and Stress Exposure Training (SET) frameworks. These skills would be learned prior to using the system and practiced in AST when proficiency is achieved. These skills include cognitive self-regulation strategies, coping, biofeedback, deep-breathing, or overlearning (see Driskell et al., 2008).
8	Individual and team training	Adaptive stress training	Training with adaptive stress training can be currently used with individuals. The technology for hosting multiple users in a VR training environment exists, but more development is needed to incorporate that technology in the adaptive stress training system. Further, research is needed on how to adapt the stressors for multiple users while still personalizing the training for the individuals stress response.
9	What are optimal training durations?	Adaptive stress training	More research is needed on how long training sessions should be or how much time is needed between sessions while still retaining developing skills.
10	Transfer of task skills and inoculation to new environment	Adaptive stress training	The SIT and SET frameworks promote the transfer of inoculation to situations with novel tasks or stressors. It is possible the type of training, duration, and task have effects on the transferability. More research is needed on the characteristic that effect training and stress inoculation transfer.
11	Types of procedures that can be trained	Adaptive stress training	The adaptive stress training manipulates environmental stressors in VR that coincide with an emergency fire procedure with the goal of inoculating to those stressors. However, some procedures may not have associated environment stressors and many need to adapt other types of information presented to the user (e.g., depressurization on the ISS). More research is need on what types of adaptations be support adaptive stress training when environmental stressors are not present.
12	Diversity of the user characteristics that can benefit from training	Adaptive stress training	adaptive stress training can benefit a broad population of personal characteristic because of the customized training and personalized stress detection. Current user limitations are primarily regarding understanding of the task procedure and user ability to become proficient in the task execution before using the adaptive stress training system.

1. Developing more reliable stress detection

The effectiveness of the adaptive stress training system primarily relies on the accuracy of the stress detection. The NASA HRP roadmap lists the gap *CBS-BMed2: We need to identify and validate measures to monitor behavioral health and performance during exploration class missions to determine acceptable thresholds for these measures.* Stress detection could be used

to help close this gap by developing ways to monitor stress of crew members during exploratory class missions as well as inform how the training system behaves. However, several challenges limit the reliability of the stress detection, this includes (1) physiological changes over the time of hours, days and week that may diverge from the detections model, (2) noise sensitivity of the sensors (e.g., movement, or disconnected sensors) that may lead to a wrong stress classification, and (3) confounding of cardiovascular activity or orthostatic changes. To address these challenges more research is needed on the preprocessing of sensor data and periodic model calibration to rectify the model for physiological changes. Developing stress detection capabilities may further increase the reliability real-time systems for use in ambient environments.

2. Team Adaptive Stress Training

Many exploratory class missions will include situations and tasks where teamwork is required. The NASA HRP roadmap details the gap of *Team Gap 5: We need to identify validated ground-based training methods that can be both preparatory and continuing to maintain team function in autonomous, long duration, and/or distance exploration missions.* Adaptive stress training may be able to help close this gap with further development on multiple user interaction within the system. This research project looked an individual task skill and ability of adaptive stress training to help inoculate individuals to stress. Findings who show that there are task and stress inoculation benefits for one individual. However, it is unclear if these benefits will transfer to adaptive stress training with multiple users working together to complete a procedure. More research is needed on how susceptible team skills are to break down under stress and how to train team skills such that they are resistant the deteriorating effects of stress.

3. Coping Strategies during Training

Stress Inoculation Training (SIT) and Stress Exposure Training (SET) have a three-phase framework with Phase 1 focused on education about the nature of the stress response, Phase 2 about developing skills, and Phase 3 where skills are practiced under gradually increasing stressor levels. Coping strategies are very successful at managing stress and have been promoted as a beneficial skill for Phase 2 and 3 (Meichenbaum, 2007). The NASA HRP roadmap details the gap of *CBS-BMed1: We need to identify and validate countermeasures that promote individual behavioral health and performance during exploration class missions.* The adaptive stress training is an environment based on the concepts of SIT and SET Phase 3. Therefore, the adaptive stress training may be able to help close the research gap with more research on teaching copings skills and having users practice them with an adaptive stress training system. This research may provide valuable insight into which coping strategies provide the most stress inoculation against different stressors and situations, while simultaneously help researchers identify ways to prevent the formation of maladaptive coping strategies (e.g., rumination) that could degrade task performance in an operations setting.

4. Mixed Reality Emergency Training Scenarios

The adaptive stress training allows users to practice an ISS emergency fire procedure in a fully simulated VR environment. Many other procedures exist that crewmembers will need to perform for exploratory class missions. The NASA HRP roadmap details the research gap of *CBS-BMed1: We need to identify and validate countermeasures that promote individual behavioral health and performance during exploration class missions.* The adaptive stress training system may be able to help close this gap with the addition of a mixed-realty environment that allows uses to practice on exploratory spacecraft on procedures which may be

executed in stressful conditions. If the procedure is intended to be performed with a spacecraft interface (e.g., avionics console), a mockup of the interface could be constructed and used in conjunction with augmented realty headsets to change environment stressors. The long-term vision would be to incorporate training *in-situ* using the actual vehicle interfaces in a training mode with mixed reality providing the environment augmentations. Practicing the procedures with an adaptive stress training simulated spacecraft may help crewmembers build stress inoculation, but more research needs to be done on which emergency procedures can use adaptive stress training most effectively.

5. Training Transfer – VR to analog to onboard training

The main goal of training is to help develop skills that can be transferred to the operational environment. Similarly, the goal of adaptive stress training is to help practice skills in gradually increasing stress conditions that can help build stress resilience for the real spaceflight environment. The NASA HRP roadmap details the research gap of *TRAIN-03: We need to develop guidelines for effective onboard training systems that provide training traditionally assumed for pre-flight. (Previous title: SHFE-TRAIN-03).* The adaptive stress training system can help close this gap with further research into the effectiveness of adaptive stress training as an onboard training system. The adaptive stress training would supplement any task training that is conducted in analog environments. Prior to being used on-board, research is also needed on the effect of training transfer from the adaptive stress training system to an analog environment (e.g., ISS mockup) and the spaceflight environment on-board a NASA spacecraft. A research approach investigating training transfer would provide support that using the adaptive stress training as an on-board training system would result in stress inoculation.

CHAPTER 9. CONCLUSION

Emergency training is an essential countermeasure tool to mitigate safety risks to vehicles, operators, and increase the probability of mission success. This dissertation presented research on the development and testing of an adaptive system for astronauts to prevent stress during spaceflight emergencies.

Study 1 (see CHAPTER 4.) addressed the first research question "*Can the combination of graduated stress exposure in an interactive 3D VR environment inoculate people against stress?*" The results from study 1 showed that prior exposure, as would be experienced during graduated stress exposure, enhanced relaxation behavior when confronted with a subsequent stressful condition in VR. These results support the prior studies that graduated stress exposure through the SIT and SET enhances coping ability to an acute stress. Specifically, the results are similar to findings on graduated VR training for polish military officers that had reduced stress at the conclusion of training (Ilnicki, Wiederhold, & Maciolek, 2011). These results motivated the development of a training module component of an adaptive system.

Study 2 (See CHAPTER 5.) addressed the research question, "*Can a spaceflight procedure simulated in VR be manipulated to evoke multiple stress levels?*" The results demonstrated that environmental stressor could be manipulated to elicit a multi-level subjective stress response. This was a key requirement for graduated training, because different levels of stressors must reliably increase stress exposure over time. The ability to manipulate stress in a predictable manner by changing stressors provides the foundation for adaptive training. The VR stimulation and stressor levels were then used in subsequent studies.

Study 3 (see CHAPTER 6.) addressed the research question, "*Can multiple stress levels be detected and identified from physiological measures taken during simulated tasks with ever-*

higher levels of complexity?" A physiologically based stress detection system was developed with a multi-class personalized classification model for time-series data. Specifically, a novel Approximate Bayes (ABayes) classifier was developed and evaluated in comparison of traditional machine learning classifiers for tasks of ever-higher complexity. The results of this study showed that the ABayes classifier had similar performance to other standard machine learning classifiers and that classification performance was consistent for tasks of varying complexity. The advantages of using ABayes in comparison to other classifiers is the direct and transparent connection to probability modeling. These results provided support for deploying the stress detection with an ABayes classifier in a real-time training system.

Study 4 (see CHAPTER 7.) addressed the research question, "*Can a real-time physiology-driven VR adaptive system enhance resilience to stress without degrading performance?*" The subcomponents developed in the prior studies were integrated into an adaptive VR stress training system. This system incorporated high-fidelity virtual environments, monitored physiological stress real-time, and adapted based on user stress levels. The results from study 4 demonstrated that the adaptive system can help inoculate individuals to higher levels of stress, help increase task performance, and may be more beneficial than skill-only and graduated training pedagogies that do not adjust the training to the user's perceived stress.

Lastly, a list of lessons-learned and recommendations was compiled for each of the studies (see CHAPTER 8.). Many of these lessons were generated from the task training, system development, and experimental methods. The recommendations may be useful for future researchers looking to develop a training system or generate research questions.

Contribution

The results of these studies provide contributions to the design of effective stress training systems. Stress training has a number of practical applications for spaceflight training. First, graduated stress exposure provides unique advantages in comparison to traditional skill training. By introducing individuals to stressors through multiple sessions of graduated stress exposure, users became familiar with the stress encountered in emergencies while promoting competency and control of their stress response.

Second, real-time physiological stress detection may have the capacity to account for individual differences in stress responses. In this work, a real-time stress detection system was developed that can classify multiple stress levels. The classification using the newly developed Approximate Bayes classifier gives class conditional probabilities and is more transparent in calculating probability densities than other standard machine learning classifiers. Further, a comparison of this work on time-series classification to past research on stress detection showed that a personalized interval method approach may be beneficial and could help other researcher create future systems for detecting stress using physiological signals.

Third, the classification and stress detection may help trainers monitor stress levels and align training to match the user's current state. By considering stress-performance relationships and theories on learning, the adaptive training system can consistently adapt to keep user in an optimal learning zone that challenges the user without under- or over-whelming them. When ground support and training resources are limited, stress can be monitored during long duration space flights to adjust training practices. A mobile training system can minimize setup and implementation difficulty and require minimal trainer/psychologist support.

Finally, the lessons learned and recommendations from this work may help develop and evaluate future adaptive VR stress training systems. Future development of this system could provide training support for performance under stressful conditions by helping U.S. astronaut crews develop the critical proficiencies of flexibility and adaptability. Further, this is the first step toward a mobile training system for use at home, during spaceflight, or on the Martian surface.

REFERENCES

Abrams, M. P., Carleton, R. N., Taylor, S., & Asmundson, G. J. (2009). Human tonic immobility: measurement and correlates. *Depression and anxiety*, *26*(6), 550–6. https://doi.org/10.1002/da.20462

Anglin, K. M., Anania, E., Disher, T. J., & Kring, J. P. (2017, March). Developing skills: A training method for long-duration exploration missions. In *2017 IEEE Aerospace Conference* (pp. 1-7). IEEE. https://doi.org/10.1109/AERO.2017.7943602

Bachman, K. R. O., Otto, C., & Leveton, L. National Aeronautics and Space Administration (NASA). (2012). Countermeasures to mitigate the negative impact of sensory deprivation and social isolation in long-duration space flight. (No. NASA/TM-2012-217365). Houston, TX: NASA Johnson Space Center.

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery, 31*(3), 606-660. https://doi.org/10.1007/s10618-016-0483-9

Balmain, C., & Fleming, M. (2009). A Methodology for Training International Space Station Crews to Respond to On-Orbit Emergencies (No. 2009-01-2446). SAE Technical Paper. https://doi.org/10.4271/2009-01-2446

Barshi, I., & Dempsey, D. L. National Aeronautics and Space Administration (NASA). (2016). Risk of Performance Errors Due to Training Deficiencies: Evidence Report. (No. JSC-CN-35755). Houston, TX: NASA Johnson Space Center.

Brammer, J. C., Van Peer, J. M., Michela, A., van Rooij, M. M., Oostenveld, R., Klumpers, F., ... & Roelofs, K. (2021). Breathing biofeedback for police officers in a stressful virtual environment: challenges and opportunities. *Frontiers in psychology*, *12*, 401. https://doi.org/10.3389/fpsyg.2021.586553

Buckey, J. C. (2006). Space physiology. Oxford University Press, USA.

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological psychology*, *42*(3), 249-268. https://doi.org/10.1016/0301-0511(95)05161-9

Campbell, J., & Ehlert, U. (2012). Acute psychosocial stress: does the emotional stress response correspond with physiological responses?. *Psychoneuroendocrinology*, *37*(8), 1111-1134. https://doi.org/10.1016/j.psyneuen.2011.12.010

Carpenter, R. (2016). A review of instruments on cognitive appraisal of stress. *Archives of Psychiatric Nursing*, *30*(2), 271-279. https://doi.org/10.1016/j.apnu.2015.07.002

Cater, J. P., & Huffman, S. D. (1995). Use of the Remote Access Virtual Environment Network (RAVEN) for Coordinated IVA—EVA Astronaut Training and Evaluation. *Presence: Teleoperators & Virtual Environments, 4*(2), 103-109. https://doi.org/10.1162/pres.1995.4.2.103

Crawford, C., Wallerstedt, D., & Khorsan, R. (2013). A systematic review of biopsychosocial training programs for the self-management of emotional stress: potential applications for the military. *Evidence-Based Complementary and Alternative Medicine.* https://doi.org/10.1155/2013/747694

Ćosić, K., Popović, S., Horvat, M., Kukolja, D., Dropuljić, B., Kostović, I., Judaš, M., Radoš, M., Radoš, M., Vasung, L., & Spajić, B.B., (2011). Virtual reality adaptive stimulation in stress resistance training. Proceedings *RTO-MP-HFM-205 on" Mental Health and Well-Being across the Military Spectrum.*

Ćosić, K., Popović, S., Kostovic, I., & Judas, M. (2010a). Virtual reality adaptive stimulation of limbic networks in the mental readiness training. *Studies in health technology and informatics, 154*, 14-19. https://doi.org/10.3233/978-1-60750-561-7-14

Ćosić, K., Popović, S., Kukolja, D., Horvat, M., & Dropuljić, B. (2010b). Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders. *CyberPsychology, Behavior, and Social Networking, 13*(1), 73-78. https://doi.org/10.1089/cyber.2009.0260

Delahaij, R., Gaillard, A.W.K., & Soeters, J.M.L.M. (2006). Stress training and the new military environment. In Human Dimensions in Military Operations – Military Leaders' Strategies for Addressing Stress and Psychological Support (pp. 17A-1 – 17A-10). Meeting Proceedings RTO-MP-HFM-134, Paper 17A. Neuilly-sur-Seine, France: RTO.

Driskell, J. E., & Johnston, J. H. (1998). Stress exposure training. *Making decisions under stress: Implications for individual and team training,* 191-217. https://psycnet.apa.org/doi/10.1037/10278-007

Driskell, J. E., Johnston, J. H., & Salas, E. (2001). Does stress training generalize to novel settings?. *Human Factors, 43*(1), 99-110. https://doi.org/10.1518%2F001872001775992471

Driskell, J. E., & Salas, E. (Eds.). (1996). *Stress and human performance*. Lawrence Erlbaum Associates.

Driskell, J. E., Salas, E., & Johnston, J. (1999). Does stress lead to a loss of team perspective?. *Group dynamics: Theory, research, and practice, 3*(4), 291. https://psycnet.apa.org/doi/10.1037/1089-2699.3.4.291

Driskell, J. E., Salas, E., Johnston, J. H., & Wollert, T. N. (2008). Stress exposure training: An event-based approach. *Performance under stress*, 271–286. London: Ashgate

Eichler, P., Seine, R., Khanina, E., & Schön, A. (2006). Astronaut training for the European ISS contributions Columbus module and ATV. *Acta Astronautica, 59*(12), 1146-1152. https://doi.org/10.1016/j.actaastro.2006.03.004

Ellis, A. (2006). System breakdown: The role of mental models and transactive memory in the relationship between acute stress and team performance. *Academy of Management Journal, 49*(3), 576–589. https://doi.org/10.5465/AMJ.2006.21794674

Ellis, B. J., Del Guidice, M., & Shirtcliff, E. A. (2013). Beyond allostatic load. In T. Beauchaine & S. Hinshaw (Eds.), *Child and Adolescent Psychopathology* (3rd ed., pp. 251–284). Hoboken, NJ: Wiley. https://doi.org/10.1017/S0954579413000849

Evetts, S. N. (2009). Overview of Bioastronautics. In *Safety Design for Space Systems* (pp. 105-161). https://doi.org/10.1016/B978-0-7506-8580-1.00003-8

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion, 7*(2), 336. https://psycnet.apa.org/doi/10.1037/1528-3542.7.2.336

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: a framework for researchers and system designers. *Human Factors, 54*(6), 1008-1024. https://doi.org/10.1177%2F0018720812443983

Finseth, T., Dorneich, M. C., Keren, N., Franke, W. D., & Vardeman, S. (2020). Designing Training Scenarios for Stressful Spaceflight Emergency Procedures. In *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)* (pp. 1-10). IEEE. https://doi.org/10.1109/DASC50938.2020.9256403.

Finseth T., Dorneich, M.C., Keren, N., Franke, W., Vardeman, S., Segal, J., Deick, A., Cavanah, E., & Thompson, K. (2021b). *An Approach to Adaptive Training for Stress Inoculation.* [Manuscript submitted for publication]. Department of Aerospace Engineering, Iowa State University.

Finseth, T., Dorneich, M. C., Vardeman, S., Keren, N., & Franke, W. D. (2021a). *Physiologically Based Stress Detection For Hazardous Operations Using Approximate Bayes Algorithm.* [Manuscript submitted for publication]. Department of Aerospace Engineering, Iowa State University.

Finseth, T. T., Keren, N., Dorneich, M. C., Franke, W. D., Anderson, C. C., & Shelley, M. C. (2018). Evaluating the Effectiveness of Graduated Stress Exposure in Virtual Spaceflight Hazard Training. *Journal of Cognitive Engineering and Decision Making, 12*(4), 248-268. https://doi.org/10.1177%2F1555343418775561

Fletcher, D., & Sarkar, M. (2013). Psychological resilience: A Review and Critique of Definitions, Concepts, and Theory. *European Psychologist, 18*(1), 12-23. https://doi.org/10.1027/1016-9040/a000124

Foley, F. W., Bedell, J. R., LaRocca, N. G., Scheinberg, L. C., & Reznikoff, M. (1987). Efficacy of stress-inoculation training in coping with multiple sclerosis. *Journal of consulting and clinical psychology*, *55*(6), 919–22. https://doi.org/10.1037/0022-006X.55.6.919

Folkman, S. (1984). Personal control and stress and coping processes: a theoretical analysis. *Journal of personality and social psychology*, *46*(4), 839. https://psycnet.apa.org/doi/10.1037/0022-3514.46.4.839

Friedland, N., & Keinan, G. (1992). Training effective performance in stressful situations: Three approaches and implications for combat training. *Military Psychology*, *4*(3), 157-174. https://doi.org/10.1207/s15327876mp0403_3

Fuchs, S., Hale, K. S., Berka, C., & Juhnke, J. (2008). A mitigation framework for enhancing situation awareness. *Augmented Cognition: A Practitioner's Guide*, 112-143.

Gaab, J., Blättler, N., Menzi, T., Pabst, B., Stoyer, S., & Ehlert, U. (2003). Randomized controlled evaluation of the effects of cognitive–behavioral stress management on cortisol responses to acute stress in healthy subjects. *Psychoneuroendocrinology*, *28*(6), 767-779. https://doi.org/10.1016/S0306-4530(02)00069-0

Gaillard, A. W. K. (2001). Stress, workload, and fatigue as three biobehavioral states: A general overview. *Stress, workload, and fatigue*, 623-640. Mahwah, NJ: Lawrence Erlbaum.

Gamble, K. R., Vettel, J. M., Patton, D. J., Eddy, M. D., Davis, F. C., Garcia, J. O., Spangler, D. P., Thayer, J. F., & Brooks, J. R. (2018). Different profiles of decision making and physiology under varying levels of stress in trained military personnel. *International Journal of Psychophysiology*, *131*, 73-80. https://doi.org/10.1016/j.ijpsycho.2018.03.017

Gancet, J., Chintamani, K., & Letier, P. (2012). Force feedback and immersive technologies suit (FITS): an advanced concept for facility-less astronaut training. *In International symposium on artificial intelligence, robotics and automation in space, Turin, Italy*.

Garcia, A. D., Schlueter, J., & Paddock, E. (2020). Training astronauts using hardware-in-the-loop simulations and virtual reality. In *AIAA Scitech 2020 Forum* (p. 0167). https://doi.org/10.2514/6.2020-0167

Gaume, A., Vialatte, A., Mora-Sánchez, A., Ramdani, C., & Vialatte, F. B. (2016). A psychoengineering paradigm for the neurocognitive mechanisms of biofeedback and neurofeedback. *Neuroscience and biobehavioral reviews*, *68*, 891–910. https://doi.org/10.1016/j.neubiorev.2016.06.012

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., & Tsiknakis, M. (2019). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2019.2927337

Grassi, A., Gaggioli, A., & Riva, G. (2011). New technologies to manage exam anxiety. *Stud Health Technol Inform 167*:57–61. https://doi.org/1010.3233/978-1-60750-766-6-57

Hadfield, C. (2016). *An astronaut's guide to life on earth* (1st ed., p. 320). Back Bay Books.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological psychology, 45*(1-3), 73-93. https://doi.org/10.1016/S0301-0511(96)05223-4

Homan, D., & Gott, C. (1996). An integrated EVA/RMS virtual reality simulation, including force feedback for astronaut training. *In Flight Simulation Technologies Conference* (p. 3498).

Hourani, L. L., Kizakevich, P. N., Hubal, R., Spira, J., Strange, L. B., Holiday, D. B., Bryant, S., & McLean, A. N.(2011). Predeployment stress inoculation training for primary prevention of combat-related stress disorders. *Journal of CyberTherapy & Rehabilitation 4*(1), 101-116.

Hysong, S. J., Galarza, L., & Holland, A. W. (2007). A Review of Training Methods and Instructional Techniques: Implications for Behavioral Skills Training in US Astronauts (DRAFT) (JSC-CN-10905). Houston, TX: National Aeronautics and Space Administration.

Ilnicki, Wiederhold, & Maciolek. (2011). Effectiveness evaluation for short-term group pre-deployment VR computer-assisted stress inoculation training provided to Polish ISAF soldiers. *Annual Review of Cybertherapy and Telemedicine 2012*, 113. https://doi.org/10.3233/978-1-61499-121-2-113

Jackson, S., Baity, M. R., Bobb, K., Swick, D., & Giorgio, J. (2019). Stress inoculation training outcomes among veterans with PTSD and TBI. *Psychological Trauma: Theory, Research, Practice, and Policy, 11*(8), 842–850. https://doi.org/10.1037/tra0000432

Johnston, J. H., & Cannon-Bowers, J. A. (1996). Training for stress exposure. In J. E. Driskell & E. Salas (Eds.), Series in applied psychology. *Stress and human performance* (p. 223–256). Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203772904-15

Jones, D., & Dechmerowski, S. (2016, July). Measuring Stress in an Augmented Training Environment: Approaches and Applications. In *International Conference on Augmented Cognition* (pp. 23-33). Springer, Cham. https://doi.org/10.1007/978-3-319-39952-2_3

Kaber, D. B., Wright, M. C., Prinzel III, L. J., & Clamann, M. P. (2005). Adaptive automation of human-machine system information-processing functions. *Human factors, 47*(4), 730-741. https://doi.org/10.1518%2F001872005775570989

Kalisch, R., Baker, D. G., Basten, U., Boks, M. P., Bonanno, G. A., Brummelman, E., Chmitorz, A., Fernàndez, G., Fiebach, C. J., Galatzer-Levy, I., Geuze, E., Groppa, S., Helmreich, I., Hendler, T., Hermans, E. J., Jovanovic, T., Kubiak, T., Lieb, K., Lutz, B., ... & Kleim,B. (2017). The resilience framework as a strategy to combat stress-related disorders. *Nature human behaviour, 1*(11), 784-790. https://doi.org/10.1038/s41562-017-0200-8

Kalisch, R., Müller, M. B., & Tüscher, O. (2015). A conceptual framework for the neurobiological study of resilience. *Behavioral and brain sciences, 38*. https://doi.org/10.1017/S0140525X1400082X

Keinan, G., & Friedland, N. (1996). Training effective performance under stress: Queries, dilemmas, and possible solutions. *Stress and human performance*, 257–277. Mahwah, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9780203772904-16

Kluge, M. G., Maltby, S., Walker, N., Bennett, N., Aidman, E., Nalivaiko, E., & Walker, F. R. (2021). Development of a modular stress management platform (Performance Edge VR) and a pilot efficacy trial of a bio-feedback enhanced training module for controlled breathing. *Plos one, 16*(2), e0245068. https://doi.org/10.1371/journal.pone.0245068

Kosinska., L., Ilnicki., S., Wiederhold., B.K., Maciolek., J., Szymanska., S., Zbyszeswki., M., Opaslko-Piotrikiewicz., E., Siatkowska., A., Borzetka., D., Ilinicki., P., Filarowska., M., Pleskack., K., & Murawski., P. (2013). VR stress inoculation training results for polish ISAF soldiers—a study of 4 cases. In: *NATO Advanced Study Institute. Invisible Wounds—New Tools to Enhance PTSD Diagnosis and Treatment, Ankara, Turkey*

Kramer, M. (2015, January 14). *Space Station Ammonia Leak Scare Likely a False Alarm, NASA Says*. SPACE. https://www.space.com/28262-space-station-ammonia-leak-false-alarm.html

Lamontagne, A. D., T. Keegel, A. M. Louie, A. Ostry, and P. A. Landsbergis (2007). A systematic review of the job-stress intervention evaluation literature, 1990–2005. *International Journal of Occupational and Environmental Health, 13*(3), 268–280. https://doi.org/10.1179/oeh.2007.13.3.268

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping.* New York: Springer-Verlag.

Leipold, B., & Greve, W. (2009). Resilience: A conceptual bridge between coping and development. *European Psychologist, 14*(1), 40-50. https://doi.org/10.1027/1016-9040.14.1.40

Linenger, J. M. (2000). Off the planet: Surviving Five Perilous Months Aboard the Space Station Mir (1st ed., p. 256). McGraw-Hill Education.

Lugrin, J. L., Latoschik, M. E., Habel, M., Roth, D., Seufert, C., & Grafe, S. (2016). Breaking bad behaviors: A new tool for learning classroom management using virtual reality. *Frontiers in ICT, 3*, 26. https://doi.org/10.3389/fict.2016.00026

Maciolek, J., Ilnicki, S., Wiederhold, B. K., Kosinska, L., Szymanska, S., Zbyszewski, M., ... & Murawski, P. (2013). Th e influence of pre-deployment VR computer-assisted stress inoculation training on the anxiety level in the Polish ISAF soldiers. *New tools to enhance posttraumatic stress disorder diagnosis and treatment, 108*, 161-1.

Malbos, E., Mestre, D. R., Note, I. D., & Gellato, C. (2008). Virtual reality and claustrophobia: multiple components therapy involving game editor virtual environments exposure. *CyberPsychology & Behavior, 11*(6), 695-697. https://doi.org/10.1089/cpb.2007.0246

Marciacq, J. B., & Bessone, L. (2009). Crew Training Safety: An Integrated Process. In G. Musgrave, A. Larsen, & T. Sgobba (Eds.), Safety design for space systems. (pp. 745-815). Butterworth-Heinemann. https://10.1016/B978-0-7506-8580-1.00025-7

Matthews, G., Reinerman-Jones, L., Wohleber, R., Lin J., Mercado J., & Abich J. (2015). Workload Is Multidimensional, Not Unitary: What Now? In: Schmorrow D.D., Fidopiastis C.M. (eds) *Foundations of Augmented Cognition. AC 2015. Lecture Notes in Computer Science, vol 9183.* Springer, Cham. https://doi.org/10.1007/978-3-319-20816-9_5

McEwen, B. S. (2004). Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. *Annals of the New York Academy of Sciences, 1032*(1), 1-7. https://doi.org/10.1196/annals.1314.001

McEwen, B. S. (2005). Stressed or stressed out: what is the difference?. *Journal of Psychiatry and Neuroscience, 30*(5), 315.

McEwen, B. S. (2013). The brain on stress: Toward an integrative approach to brain, body, and behavior. *Perspectives on Psychological Science, 8*, 673–675. https://doi.org/10.1177%2F1745691613506907

Meichenbaum, D. (1985). Stress Inoculation Training. New York: Pergamon Press.

Meichenbaum, D. (2007). Stress Inoculation training: a preventative and treatment approach. In R. W. P.M. Lehrer, *Principles and Practice of Stress Management,* 3rd Edition. Guilford Press.

Meichenbaum, D., & Cameron, R. (1989). Stress Inoculation Training. Stress Reduction and Prevention, 115–154. New York: Springer.

Meyerbröker, K., & Emmelkamp, P. M. (2010). Virtual reality exposure therapy in anxiety disorders: a systematic review of process-and-outcome studies. *Depression and anxiety, 27*(10), 933-944. https://doi.org/10.1002/da.20734

Nacke, L. E., Kalyn, M., Lough, C., & Mandryk, R. L. (2011, May). Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 103-112). ACM. https://doi.org/10.1145/1978942.1978958

Oberg. (1998). Shuttle-Mir's lessons for the international space station. https://doi.org/10.1109/6.681969

Olbrich, M., Graf, H., Keil, J., Gad, R., Bamfaste, S., & Nicolini, F. (2018, July). Virtual reality based space operations–a study of ESA's potential for VR based training and simulation. *In International Conference on Virtual, Augmented and Mixed Reality* (pp. 438-451). Springer, Cham. https://doi.org/10.1007/978-3-319-91581-4_33

Olff, M., Langeland, W., & Gersons, B. P. (2005). Effects of appraisal and coping on the neuroendocrine response to extreme stress. *Neuroscience & Biobehavioral Reviews, 29*(3), 457-467. https://doi.org/10.1016/j.neubiorev.2004.12.006

Orasanu, J., & Backer, P. (1996). Stress and military performance. *Stress and human performance*, 89–125. Mahwah, NJ: Lawrence Erlbaum.

Orasanu, J.M., Kraft, N., Tada, Y. (2006). *Augmented Cognition: Past, Present, and Future*, 2nd edn.

Ozhiganova, G. V. (2018). Self-regulation and self-regulatory capacities: components, levels, models. RUDN Journal of Psychology and Pedagogics, 15(3), 255-270. https://doi.org/10.22363/2313-1683-2018-15-3-255-270

Palinkas, L. A. (2007). Psychosocial issues in long-term space flight: overview. *Gravitational and Space Research, 14*(2).

Pallavicini, F., Argenton, L., Toniazzi, N., Aceti, L., & Mantovani, F. (2016). Virtual reality applications for stress management training in the military. *Aerospace medicine and human performance, 87*(12), 1021-1030. https://doi.org/10.3357/AMHP.4596.2016

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30*(3), 286-297. https://doi.org/10.1109/3468.844354

Parnandi, A., & Gutierrez-Osuna, R. (2015). A comparative study of game mechanics and control laws for an adaptive physiological game. *Journal on Multimodal User Interfaces, 9*(1), 31–42. Springer. https://doi.org/10.1007/s12193-014-0159-y

Parnandi, A., Son, Y., & Gutierrez-Osuna, R. (2013, September). A control-theoretic approach to adaptive physiological games. *In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 7-12). IEEE. https://doi.org/10.1109/ACII.2013.8

Parsons, T. D., & Reinebold, J. L. (2012). Adaptive virtual environments for neuropsychological assessment in serious games. *IEEE Transactions on Consumer Electronics, 58*(2). https://doi.org/10.1109/TCE.2012.6227413

Perna, F. M., Antoni, M. H., Baum, A., Gordon, P., & Schneiderman, N. (2003). Cognitive behavioral stress management effects on injury and illness among competitive athletes: a randomized clinical trial. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine, 25*(1), 66–73. https://doi.org/10.1207/S15324796ABM2501_09

Prachyabrued, M., Wattanadhirach, D., Dudrow, R. B., Krairojananan, N., & Fuengfoo, P. (2019, March). Toward virtual stress inoculation training of prehospital healthcare personnel: A stress-inducing environment design and investigation of an emotional connection factor. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 671-679). IEEE. https://doi.org/10.1109/VR.2019.8797705

Pritchett, A. R., Kim, S. Y., & Feigh, K. M. (2014). Measuring human-automation function allocation. *Journal of Cognitive Engineering and Decision Making, 8*(1), 52-77. https://doi.org/10.1177%2F1555343413490166

Popović, S., Horvat, M., Kukolja, D., Dropuljic, B., & Cosic, K. (2009). Stress inoculation training supported by physiology-driven adaptive virtual reality stimulation. *Annual review of cybertherapy and telemedicine*, 50-54. https://doi.org/10.3233/978-1-60750-017-9-50

Rachman, S. (Ed.). (1983). Fear and courage among military bomb-disposal operators. *Advances in Behavior Research and Therapy, 4*(3).

Regehr, C., Glancy, D., & Pitts, A. (2013). Interventions to reduce stress in university students: a review and meta-analysis. *Journal of affective disorders, 148*(1), 1–11. https://doi.org/10.1016/j.jad.2012.11.026

Riva, G., Grassi, A., Villani, D., Gaggioli, A., & Preziosa, A. (2007). Managing exam stress using UMTS phones: the advantage of portable audio/video support. *Stud Health Technol Inform 125*:406–408

Rizzo, A., Parsons, T.D., Lange, B., Kenny, P., Buckwalter, J.G., Rothbaum, B., Difede, J., Frazier, J., Newman, B., Williams, J., & Reger, G. (2011). Virtual reality goes to war: A brief review of the future of military behavioral healthcare. *Journal of clinical psychology in medical settings, 18*(2), pp.176-187. http://dx.doi.org/10.1007%2Fs10880-011-9247-2

Robson, S., & Manacapilli, T. (2014). *Enhancing Performance Under Stress: Stress Inoculation Training for Battlefield Airmen.* Santa Monica, CA: The Rand Corporation.

Rose, R. D., Buckey, J. C., Zbozinek, T. D., Motivala, S. J., Glenn, D. E., Cartreine, J. A., & Craske, M. G. (2013). A randomized controlled trial of a self-guided, multimedia, stress management and resilience training program. *Behaviour research and therapy, 51*(2), 106–12. https://doi.org/10.1016/j.brat.2012.11.003

Rothbaum, B. O., Hodges, L., Smith, S., Lee, J. H., & Price, L. (2000). A controlled study of virtual reality exposure therapy for the fear of flying. *Journal of consulting and Clinical Psychology, 68*(6), 1020. https://doi.org/10.1037/0022-006X.68.6.1020

Russo, S. J., Murrough, J. W., Han, M.-H., Charney, D. S., & Nestler, E. J. (2012). Neurobiology of resilience. *Nature Neuroscience, 15*(11), 1475–1484. https://doi.org/10.1038/nn.3234

Salamon, N., Grimm, J. M., Horack, J. M., & Newton, E. K. (2018). Application of virtual reality for crew mental health in extended-duration space missions. *Acta Astronautica, 146*, 117-122. https://doi.org/10.1016/j.actaastro.2018.02.034

Saunders, T., Driskell, J. E., Johnston, J., & Salas, E. (1996). The effect of stress inoculation training on anxiety and performance. *Journal of occupational health psychology, 1*(2), 170–186. https://doi.org/1010.1037/1076-8998.1.2.170

Schwartz, M. S., & Andrasik, F. (Eds.). (2017). Biofeedback: A practitioner's guide. Guilford Publications.

Serino, S., Triberti, S., Villani, D., Cipresso, P., Gaggioli, A., & Riva, G. (2014). Toward a validation of cyber-interventions for stress disorders based on stress inoculation training: a systematic review. *Virtual Reality, 18*(1), 73–87. https://doi.org/10.1007/s10055-013-0237-6

Sharma, N., & Gedeon, T. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine, 108*(3), 1287-1301. https://doi.org/10.1016/j.cmpb.2012.07.003

Smets, E., De Raedt, W., & Van Hoof, C. (2019). Into the wild: the challenges of physiological stress detection in laboratory and ambulatory settings. *IEEE journal of biomedical and health informatics, 23*(2), 463-473. https://doi.org/10.1109/JBHI.2018.2883751

Smith-Jentsch, K. A., & Sierra, M. J. (2016). Teamwork training needs analysis for long-duration exploration missions (p. 44). Technical Report No. JSC-CN-40388.

Staal, M. A. (2004). Stress, cognition, and human performance: A literature review and conceptual framework (NASA Tech. Memorandum 212824). Moffett Field, CA: NASA Ames Research Center.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. *Neuroscience and biobehavioral reviews, 36*(4), 1228–48. https://doi.org/10.1016/j.neubiorev.2012.02.003

Stetz, M. C., Kaloi-Chen, J. Y., Turner, D. D., Bouchard, S., Riva, G., & Wiederhold, B. K. (2011). The effectiveness of technology-enhanced relaxation techniques for military medical warriors. *Military medicine, 176*(9), 1065-1070. https://doi.org/10.7205/MILMED-D-10-00393

Stetz, M. C., Long, C. P., Schober Jr, W. V., Cardillo, C. G., & Wildzunas, R. M. (2007). Stress assessment and management while medics take care of the VR wounded. *Annual Review of CyberTherapy and Telemedicine, 5*, 165-171.

Streufert, S., & Streufert, S. C. (1981). *Stress and information search in complex decision making: Effects of load and time urgency (No. TR-4)*. Milton S Hershey Medical Center PA Dept of Behavioral Science.

Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Gotlib, I. H. (2010). Maladaptive coping, adaptive coping, and depressive symptoms: Variations across age and depressive state. *Behaviour research and therapy, 48*(6), 459-466. https://doi.org/10.1016/j.brat.2010.01.007

Thompson, M. M., & McCreary, D. R. (2006). Enhancing mental readiness in military personnel. Defence Research and Development Toronto (Canada).

Timmons, P. L., Oehlert, M. E., Sumerall, S. W., Timmons, C. W., & Borgers, S. B. (1997). Stress inoculation training for maladaptive anger: Comparison of group counseling versus computer guidance. *Computers in Human Behavior, 13*(1), 51-64. https://doi.org/10.1016/S0747-5632(96) 00029-5

Uhlig, T., Roshani, F. C., Amodio, C., Rovera, A., Zekusic, N., Helmholz, H., & Fairchild, M. (2016). ISS emergency scenarios and a virtual training simulator for Flight Controllers. *Acta Astronautica, 128*, 513-520. https://doi.org/10.1016/j.actaastro.2016.08.001

Verleysen, M., & François, D. (2005, June). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks* (pp. 758-770). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11494669_93

Ververs, P. M., Whitlow, S., Dorneich, M., & Mathan, S. (2005, July). Building Honeywell's adaptive system for the augmented cognition program. In *Foundations of augmented cognition. Proceedings of the 1st International Conference on Augmented Cognition* (pp. 460-8).

Villani, D., Grassi, A., Cognetta, C., Toniolo, D., Cipresso, P., & Riva, G. (2013). Self-help stress management training through mobile phones: an experience with oncology nurses. *Psychol Serv.* https://doi.org/10.1037/a0026459

Wachtel, P. L. (1968). Anxiety, attention, and coping with threat. Journal of Abnormal *Psychology, 73*(2), 137. https://psycnet.apa.org/doi/10.1037/h0020118

Wadsworth, M. E. (2015). Development of maladaptive coping: A functional adaptation to chronic, uncontrollable stress. *Child development perspectives, 9*(2), 96-100. https://doi.org/10.1111/cdep.12112

Wiederhold, B. K., & Wiederhold, M. D. (2006). From SIT to PTSD: Developing a continuum of care for the warfighter. *Annual Review of CyberTherapy and Telemedicine, 4,* 13-18.

Winslow, B., Carroll, M. B., Martin, J. W., Surpris, G., & Chadderdon, G. L. (2015). Identification of resilient individuals and those at risk for performance deficits under stress. *Frontiers in neuroscience, 9*, 328. https://doi.org/10.3389/fnins.2015.00328

Wu, D., & Parsons, T. D. (2011). Inductive transfer learning for handling individual differences in affective computing. *In Affective Computing and Intelligent Interaction* (pp. 142-151). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24571-8_15

Zadrozny, B., & Elkan, C. (2002, July). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694-699). https://doi.org/10.1145/775047.775151

Zahabi, M., & Razak, A. M. A. (2020). Adaptive virtual reality-based training: a systematic literature review and framework. *Virtual Reality*, 1-28. https://doi.org/10.1007/s10055-020-00434-w

Zbyszeswki, M., Ilnicki, S, Wiederhold, B.K., Maciolek, J, Kosinska, L, Szymanska, S, Opaslko-Piotrikiewicz, E., Siatkowska, A., Filarowska, M., Glibowska, A., Borzetka, D., Ilnicki, P., Filarowska, M., Pleskack, K., & Murawski, P. (2012). Personality and stress-coping factors in VR-computer-assisted stress inoculation training in the polish soldiers. In: *NATO Advanced Study Institute. Invisible Wounds—New Tools to Enhance PTSD Diagnosis and Treatment, Ankara, Turkey*