Stress in the Age of Technology: Evaluating the Validity and Accessibility of Cardiac Data from Industry-Leading Smartwatches

---

A

Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Master of Science

by

Aaron Moore

December 2021

APPROVAL SHEET

This

Thesis

is submitted in partial fulfillment of the requirements
for the degree of

Master of Science

Author: Aaron Moore

This Thesis has been read and approved by the examing committee:

Advisor: Sara Riggs

Advisor:

Committee Member: Leidy Klotz

Committee Member: Michael Fontaine

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

December 2021

Objective: To test the validity and data accessibility of the health detection features on two leading smartwatches in the industry, the Apple Watch Series 6 and Fitbit Sense, and to determine if these devices can be used to accurately measure stress responses in the body.

Background: Stress causes the release of hormones in the human body that increase heart rate, blood pressure, and energy supplies. Over time, the constant release of stress hormones can lead to many medical, psychological, and behavioral health problems. Health problems related to or caused by stress have become increasingly prevalent in the modern world. Access to health information on mobile devices have allowed users to more easily keep track of their fitness, but there is still much to be learned when it comes to stress detection. The ability to detect stress responses and access the measured data needs to be further examined.

Method: Sixteen participants performed modules of a multi-tasking program while smartwatches measured cardiac information. Multi-tasking modules varied in difficulty and included low, medium, and high workload conditions.

Results: Workload has an effect on the amount of stress responses produced by the participants, as the testing conditions showed an increase in stress responses compared to the baseline. The presence of detectable stress responses aligned with the participant's perception of how hard they were working and how stressed they were feeling. The Apple Watch Series 6 was unable to detect significant differences in stress responses as compared to the baseline in this study, and did not align with the measurements of the official ECG device used.

Conclusion: The accessibility of health-related data for in depth analysis on current smartwatches is very limited, and is highly susceptible to variability when measurements are being taken. Data collection limitations in current smartwatches make it difficult to capture peak stress responses during varying workload conditions, and likely was the cause of insignificance. Stress detection in smartwatches must be investigated further.

Application: The findings of this research provide insight on the validity and accessibility of health metrics gathered from current smartwatches and how stress responses can be evaluated using these devices.

Keywords: smartwatches, stress detection, heart rate, heart rate variability, workload

---

INTRODUCTION

Stress, in its simplest form, is a nonspecific response produced by the body to address a demand (Selye, 1984). All living organisms have evolved to undergo various physiological responses to these demands. In vertebrates, stress responses are mediated by the release of hormones like adrenaline and cortisol, which increase heart rate, blood pressure, and available energy supplies (Taborsky et al, 2021). Long before modern technological advancements made humans the dominant living organism on Earth with little threat to survival, humans relied on stress responses to maximize cognitive and physical performance in life or death situations (Selye, 1984). During times when the body is presented with an acute physical stressor, like running from a wild animal, the release of stress hormones temporarily turns off bodily functions that are not essential for survival (Stanford University, 2007). In small doses over spaced intervals, the release of such stress hormones is harmless, but an increase in the frequency or intensity of stressors presented to an individual can lead to conditions like chronic stress or post-traumatic stress disorder (PTSD) (Nidiffer & Leach, 2010). Now that humans are to the point where it is not necessary to run from wild animals on a daily basis, stressors are instead produced from things like work deadlines, exercising, relationship troubles, and money management. These modern day stressors are perceived as real threats in the mind, so the body is unable to discern between these perceptions and an actual external threat. As a result, adrenaline and cortisol are once again released in response (Selye,

1956). The constant release of stress hormones over time (chronic stress) can lead to various medical, psychological, and behavioral health problems like cardiovascular diseases, cancer, insomnia, alcohol and drug abuse, violence, and family conflict (Quick, Horn, & Quick, 1987). Due to the growing presence of stress-related long term health conditions, it is increasingly important to be able to detect and manage daily stress levels.

When a person experiences stress, several physiological responses produced by the body can be used as measurable indicators of their stress levels. These include changes in heart rate, heart rate variability, cortisol levels, and blood pressure (Childs, White, & de Wit, 2014). As previously mentioned, the majority of daily stressors are rooted in the workplace (Wainwright & Calnan, 2002). Career fields that are considered to be highly stressful may involve daily threats of life and death (e.g., police officers and warfighters) or severe consequences in the event of a mistake (e.g., miners, airline pilots, surgeons) (Cranwell-Ward & Abbey, 2005). There have been several studies (e.g., Johnson et al., 2005, Robertson & Ruiz, 2010) that leverage the physiological responses to stress to measure daily stress levels in high intensity career fields, such as when Seoane et al., (2014) measured the mental stress of combatants in real time. Even in a job that may subjectively be considered as less stressful, the human body still experiences the highest number of physiological stress responses during hours of work. Stress response rates increase as more work-related tasks are piled on (Okada et al., 2013). Many studies have shown that of the acute stress responses, cardiac activity (e.g., heart rate, heart rate variability) is one of the best indicators of stress levels in humans (Okada et al., 2013; Seoane et al., 2014; Schwerdtfeger & Friedrich-Mai, 2009) because it is linked to the cortical regions that are involved in stressful situation appraisal (Kim et al., 2018). When measuring stress responses, the majority of these studies utilize equipment that is not practical for use by the average person such as chest-mounted, three-electrode electrocardiograms (ECGs) or a sensorized glove apparatus. This equipment is great for understanding human stress responses in a controlled research environment, but due to limited accessibility the majority of people are unable to learn anything about their day to day stress levels without visiting a doctor.

With the advancement of modern technology over the past decade, people are increasingly integrating their smartphones into their daily lives for communication and entertainment purposes which has been shown to have a negative impact on overall health, depending on the extent of use (Samaha & Hawi, 2016). There has also been an increase in health-consciousness among smart device users, which is enhanced by the use of smartwatches or wearable fitness trackers (Reeder & David, 2016). Currently, the leading smartwatches in the industry, such as the Apple Watch Series 6 or Fitbit Sense, come equipped with health and activity tracking features like global-positioning systems (GPS), altimeter, blood oxygen sensor, electrical heart sensor, optical heart sensor, accelerometer, and gyroscope (Apple Inc., 2020; Fitbit LLC, 2020). In addition to all of the health and activity features, the watches are equipped with the same

or similar capabilities as a smartphone, so the popularity of smartwatches among users is steadily increasing (Chuah et al., 2016).

This work aims to gain a better understanding of the capabilities of smartwatches in relation to the detection and measurement of acute physiological stress responses, as compared to an official Food and Drug Administration (FDA) approved mobile ECG device. The goal of this research is to ultimately test the validity and data accessibility of the health detection features on two of the leading smartwatches in the industry, the Apple Watch Series 6 and Fitbit Sense, and to determine if these devices can be used to accurately measure stress responses in the body. Based on what is known about the relationship between stress levels, workload, and cardiac responses (Okada et al., 2013; Cranwell-Ward & Abbey, 2005; Wainwright & Calnan, 2002; Seoane et al., 2014), the expected results of this study are that (a) as user workload increases, user stress will increase, (b) an increase in perceived user stress will result in a measurable increase in physiological stress responses, and (c) the data streams produced from the smartwatch ECGs will align with the heart rate data streams from the official FDA-approved ECG device.

METHODS

Participants

Sixteen (n = 16) University of Virginia undergraduate and graduate students participated in this study (8 males, 8 females; M = 23, SD = 2.66). It was determined that an even number of participants were needed based on the nature of how tasks were presented during the data collection phase so that proper task-participant counter-balancing could be achieved. Participants were compensated with a $10 gift card following completion of the experiment..

Experimental Setup

The participants' task in this experimental setup was to complete two 3-minute long evaluation trials using the Multi-Attribute Task Battery II (MATB-II) program developed by National Aeronautics and Space Administration (NASA). This computer-based multitasking program is designed to evaluate operator performance and workload (Santiago-Espada et al., 2011). This program was selected because it simulates a high-stress career field (pilot) and allows programmers to manipulate the number of tasks the participant is presented within each trial (Cranwell-Ward & Abbey, 2005). Because stress typically increases as workload increases (Okada et al., 2013), the proper manipulation and implementation of the MATB-II program should elicit measurable stress responses in the user.

MATB-II Task Overview

The MATB-II program consists of four main tasks that the user is required to monitor: System Monitoring (SYSM), Tracking (TRCK), Communications (COMM), and Resource Management (RMAN) (see Figure 1). All of these tasks are concurrent during the trials and are controlled through the use of a wireless mouse, keyboard, and attached joystick.

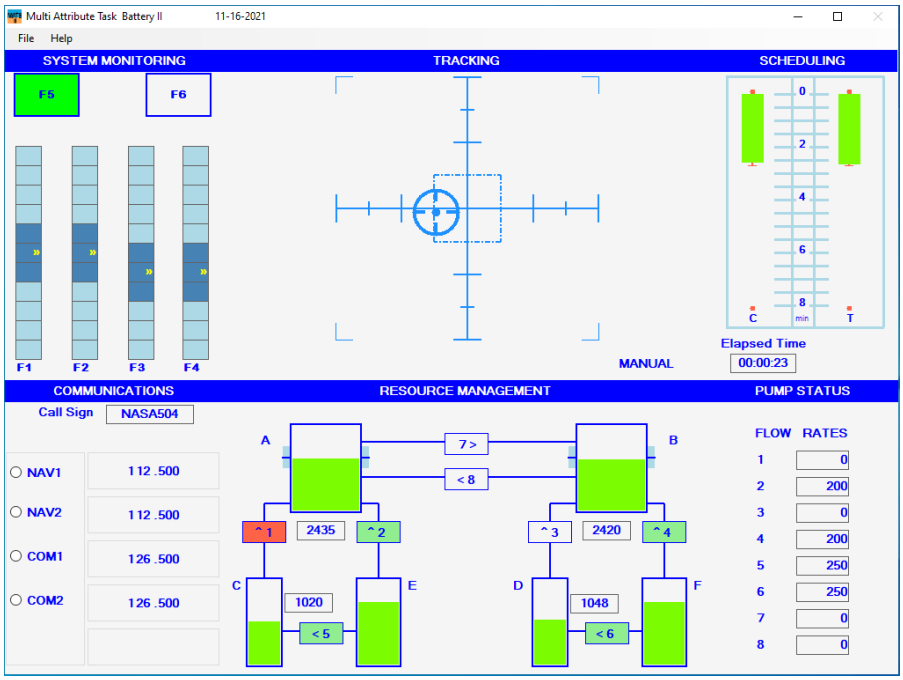

Figure 1. Screen-capture of the MATB-II program during an assessment.

The goal of the SYSM task is to ensure the darker blue panels within the light blue vertical gauges (labeled F1-F4) remain centered on the gauge. The normal state of this component is when the dark blue panels are centered on the gauge, and the failed state of this component is when the dark blue panels are at the top or bottom of the gauge. Additionally, the SYSM section has two buttons, labeled F5 and F6, that require attention. The F5 button must always remain green (failed state: button turns white) and the F6 button must always remain white (failed state: button turns red). All of these components are controlled through the use of the corresponding keys on the keyboard labeled F1-F6.

The TRCK task is meant to simulate flying an aircraft, or controlling an aircraft system during flight. The goal of this task is to keep the circle-shaped reticle as close as possible to the center of the crosshairs

through the use of a joystick connected to the computer. Left untouched, the circle-shaped reticle will randomly move away from the center of the crosshairs. The sensitivity of joystick inputs and reticle deviation intensity can be separated into three categories: Low, Medium, and High. The RMAN task represents fuel management while onboard an aircraft in flight. The goal of this task is to maintain ±250 units from the starting volumes in tanks alphabetically labeled A-D. Tanks A and B start at 2500 units and tanks C and D start at 1000 units. Tank fuel levels are controlled by operating fuel pumps numbered 1-8 (Pump States: Open = Green, Closed = White, Failed = Red). Tanks A and B are set to continually drain and there is no combination of pump flow that results in a steady state tank volume, so the user must constantly monitor this section throughout the duration of the trials.

The Communications task represents pilot interaction with aircraft controller requests (Gutzwiller, Wickens & Clegg, 2014). The user is called to action through audio rather than a visual queue in this section. The audio comes in the generic form of "*[Call sign]*, please change your *[radio]* to *[frequency]*." The goal is to change the radio and frequency using a mouse as quickly and accurately as possible if the audio message is referring to the user's assigned call sign (NASA504).

Tutorials and Trials

During the trials, participants wore an Apple Watch Series 6 on their left wrist and a Fitbit Sense on their right wrist. Both watches were positioned so that the optical and electrical heart sensors were flush against their skin and 1 cm above the end of the ulna (Figure 2).

Figure 2. Depiction of correct smartwatch placement on wrists.

Participants sat in a standard desk chair adjusted to their comfort level in front of a HP Z230 workstation with a 28" monitor, keyboard, wireless mouse, and Logitech joystick to monitor MATB-II tasks during each trial (Figure 3). An AliveCor Kardia ECG device was also placed within reach of the participants for official ECG measurements between trials.

Figure 3. Experimental setup for all three workload conditions (Low, Medium, High).

For each individual task within the MATB-II program, participants were presented with a 1-minute long tutorial that introduced the task and expectations of the study. During this time, participants were encouraged to ask questions and review the tutorial as many times as they felt necessary until they fully understood the material being presented. Three separate full-length MATB-II trials, lasting 3 minutes each, of varying difficulty (low workload: Test 1, medium workload: Training, high workload: Test 2) were utilized during the data collection phase. At the end of each trial, participants were asked to fill out a NASA Task Load Index (NASA-TLX) survey to assess subjective participant workload.

Experimental Design

This study utilized the customizability of the MATB-II program to create three trial conditions (labeled Training, Test 1, and Test 2) that varied in difficulty by the workload required from the user. The variables altered to create the three conditions were task frequency, joystick sensitivity, and pump flow rates. The training condition was designed to be a middle ground between low workload and high workload, as this condition was to be used as a qualifier for the two testable sections of the experiment. For the two testable sections, Test 1 was designed to be the least demanding of the three conditions while Test 2 was designed to be the most demanding. The experiment was intentionally customized this way so that a measurable difference in resulting stress responses would be elicited from low workload to high workload conditions. A full breakdown of the three MATB-II workload conditions are shown in Table 1.

TABLE 1: Task-specific breakdown of varying workload conditions

	Workload Conditions		
	Low	Training	High
Communication Tasks	3	6	9
System Monitoring Tasks	6	12	22
Tracking Sensitivity	Low	Medium	High
Tracking Deviation	Low	Medium	High
Resource Management Flow Rates: volume/minute	Pump 1: 300	Pump 1: 600	Pump 1: 900
	Pump 2: 200	Pump 2: 400	Pump 2: 700
	Pump 3: 300	Pump 3: 600	Pump 3: 900
	Pump 4: 200	Pump 4: 400	Pump 4: 700
	Pump 5: 250	Pump 5: 500	Pump 5: 750
	Pump 6: 250	Pump 6: 500	Pump 6: 750
	Pump 7: 200	Pump 7: 400	Pump 7: 700
	Pump 8: 200	Pump 8: 400	Pump 8: 700
	Tank A: -300	Tank A: -600	Tank A: -900
	Tank B: -300	Tank B: -600	Tank B: -900

The way in which each workload condition was scored also added to the varying difficulty between the tests. This was achieved by altering the data recording intervals of the MATB-II sections that require continuous monitoring (TRCK and RMAN). Test 1 was the lowest rated workload condition, so its recording intervals for the continuous tasks were the most spread out. As a result, participant scores were not as greatly affected by mistakes. This allows the participants more time to address tasks between each recording interval, resulting in the ability to apply attention elsewhere in the MATB-II program. Conversely, Test 2 was the highest rated workload condition, so the continuous task recording intervals were the shortest. This limits the amount of time participants have to address other tasks within the MATB-II program, adding to the difficulty level of the high workload condition. Each COMM task results in two scoring opportunities as the participant is asked to change both radio and frequency, so points are awarded for the accuracy of both of those components. Each SYSM task is only worth one point, but the participant will lose points if a working component is attempted to be fixed. The TRCK task was scored based on the root-mean-square error (RMSE) of the location of the circle-shaped reticle relative to the center of the crosshairs for each recording interval. If the RMSE indicates that the reticle was outside of the dotted square surrounding the center of the crosshairs, the participant does not receive points for that recording interval. The RMAN task is scored by logging the volumes of tanks A-D at time of each recording interval, creating four scoring opportunities per recording interval. If tank volumes are outside

the allowable volume range, then no points are awarded for that recording interval. A full breakdown of the scored events for each workload condition are shown in Table 2. Scoring each trial provides insight on how well participants are objectively performing under each workload condition.

TABLE 2: Workload Condition Scoring Breakdown

	Workload Conditions		
Scoring Opportunities	Low	Training	High
Communication	6	12	18
System Monitoring	6	12	22
Tracking	12	18	36
Resource Management	36	48	72
Perfect Score	60	90	148

Low workload and high workload conditions were presented to participants in a predetermined order so that proper task-participant counter-balancing could be achieved. Half of the participants completed low workload followed by high workload conditions, while the other half of the participants completed high workload followed by low workload conditions. In an effort to increase stress responses, participants were informed after completing the training condition that 3 participants who scored the highest over the following two testable trials would receive an extra $20 in reward money. The incentivization of a reward has been shown to increase cardiac activity under conditions of unclear difficulty (Richter & Gendolla, 2009).

This research complied with the American Psychological Association (APA) Code of Ethics and was approved by the University of Virginia's Institutional Review Board (IRB-SBS, Protocol 4428). Upon arrival, participants were asked to read and sign an informed consent form and fill out a demographic survey. The experimenter then explained the details of the study including the equipment and overview of the tasks required. Participants took a baseline ECG using the AliveCor Kardia device, and adjusted the Fitbit Sense to fit securely on their right wrist. The experimenter read the scripted instructions for each task in the MATB-II program as the participant completed the 1-minute long task tutorials. The participant then completed the full training module to become familiar with handling all main tasks at once. Participants were required to score at least 70% before moving onto the experimental portions of the study. The 70% threshold was determined during program pilot testing. Fourteen out of the 16 participants achieved or exceeded the 70% threshold score on the training module. The passing participants then

moved onto the testing portion of the experiment and fit the Apple Watch Series 6 to their left wrist. In both testing scenarios, the program was paused at the 1.5 minute mark, at which time the participants recorded an ECG using the Apple Watch Series 6. Following the Apple Watch Series 6 ECG, the program resumed until completion at the 3 minute mark, at which time the participants recorded an ECG using the AliveCor Kardia ECG. At the conclusion of the study, each participant completed a debriefing questionnaire. The study was approximately 1 hour long in duration from start to finish. A full breakdown of the flow of events involved in this study are as follows:

1. Participants sign informed consent form, complete demographic survey, and fit Fitbit to right wrist
2. Participants complete individual tutorials for each MATB-II task
3. Baseline ECG taken with AliveCor Kardia ECG Device
4. Participants complete 3-minute full training module
    - If 70% threshold is met, move on
5. Participants fit Apple Watch Series 6 to left wrist
6. Participants complete first 3-minute experimental portion (low or high workload condition)
    - At 1.5 minute mark, take ECG with Apple Watch Series 6
    - At 3 minute mark, take ECG with AliveCor Kardia ECG
    - Complete NASA-TLX questionnaire
7. Participants complete second 3-minute experimental portion (low or high workload condition)
    - After 1.5 minutes, take ECG with Apple Watch Series 6
    - After 3 minutes, take ECG with AliveCor Kardia ECG
    - Complete NASA-TLX questionnaire
8. Participants complete debriefing questionnaire

RESULTS

The dependent variable in this study was cardiac activity of the participants during each examination period. Cardiac activity was used to calculate differences in heart rate (HR) and heart rate variability (HRV) under varying participant workload conditions. The HR and HRV data was recorded using an ECG at the conclusion of the trials utilizing both the AliveCor Kardia ECG and Apple Watch Series 6 ECG features. The results were analyzed using paired samples t-tests to identify any differences between baseline cardiac measurements and examination period cardiac measurements. The two separate examination workload conditions were also directly compared through the use of t-tests. Due to a reduced

sample size, t-tests were conducted after a Shapiro-Wilk normality test was conducted to show that the distribution of the differences in the paired samples were not significantly different from a normal distribution. Subjective workload differences between participants were also calculated and analyzed using a NASA Task Load Index (NASA-TLX) questionnaire.

MATB-II Scoring

Participants were scored for each of the workload conditions so that any changes in performance could be measured between the trials, as outlined in the Methods section. During the low workload condition (Test 1), participants achieved an average score of 96.79%. Under the high workload condition (Test 2), the average score among participants dropped to 64.04%.

Heart Rate (HR)

AliveCor Kardia

A paired samples t-test between the baseline ECGs (M = 69.998, SD = 10.741) and low workload condition Kardia ECGs (M = 73.143, SD = 10.250) showed a significant increase in HR, $t(13) = 2.4089$, $p = 0.032$. A paired samples t-test between the baseline ECGs (M = 69.998, SD = 10.741) and high workload condition Kardia ECGs (M = 74.231, SD = 11.476) showed a significant increase in HR, $t(12) = 2.4847$, $p = 0.029$ (Figure 4).

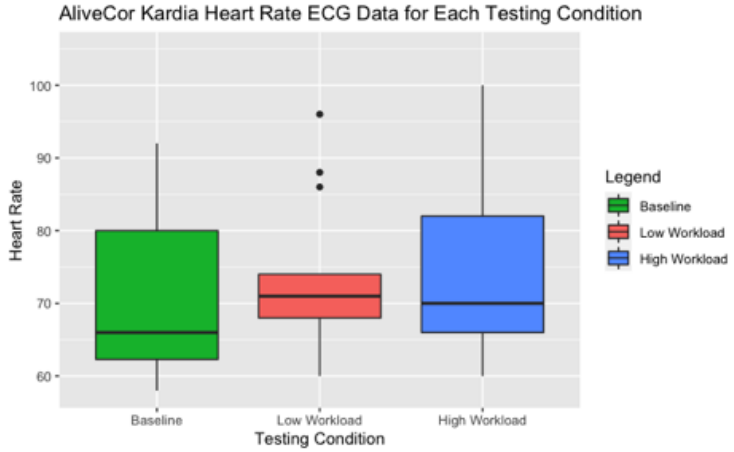

Figure 4. Differences in Heart Rate between baseline, low workload condition, and high workload condition ECGs using AliveCor Kardia ECG.

No significant differences were found between low workload (M = 73.143, SD = 10.250) and high workload (M = 74.231, SD = 11.476) condition HR measurements using the Kardia ECG device, $t(12) = 0.21707$, $p = 0.832$.

Apple Watch Series 6

No significant differences were found between baseline ECGs (M = 69.998, SD = 10.741) and low workload condition (M = 69.150, SD = 10.058) or high workload condition (M = 72.451, SD = 11.803) using the Apple Watch Series 6 ECG, $t(12) = 0.94836$, $p = 0.362$ and $t(13) = 1.8194$, $p = 0.092$, respectively. A paired samples t-test between the low workload condition Apple Watch Series 6 ECGs (M = 69.150, SD = 10.058) and high workload condition Apple Watch Series 6 ECGs (M = 72.451, SD = 11.803) showed a significant increase in HR, $t(13) = 2.6114$, $p = 0.022$ (Figure 5).

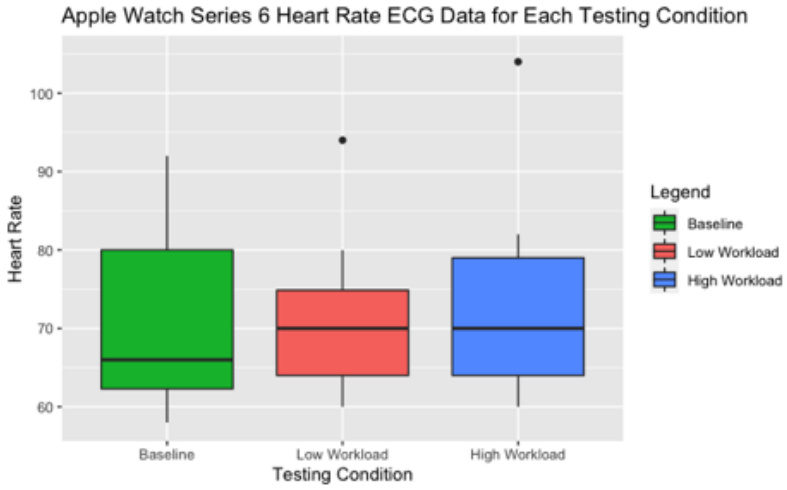

Figure 5. Differences in HR between baseline, low workload condition, and high workload condition ECGs using Apple Watch Series 6.

Fitbit Sense

The Fitbit Sense was worn by participants throughout the entire duration of the experiment and measured HR continuously on 5-second intervals. During the training module, participants showed an average HR increase of 13.500 beats per minute (bpm) and an overall average HR of 76.984 bpm. HR data gathered during the low workload condition showed an average HR increase of 18.000 bpm, and an overall average HR of 75.190 bpm. The high workload condition showed an average HR increase of 17.214 bpm, and an average HR of 76.645 bpm. Paired samples t-tests were run on the increases in HR over the three conditions (Training, Test1, and Test2) using the continuous HR data gathered from the Fitbit Sense. There were no significant differences found between low workload (M = 18.000, SD = 6.051) and high workload (M = 17.214, SD = 5.549) conditions, $t(13) = -0.45101$, $p = 0.659$. However, there were significant differences between the training module (M = 13.500, SD = 4.942) and the two varying workload conditions (Test1 and Test2), $t(13) = 3.698$, $p = 0.003$, and $t(13) = 2.2995$, $p = 0.039$, respectively (Figure 6).

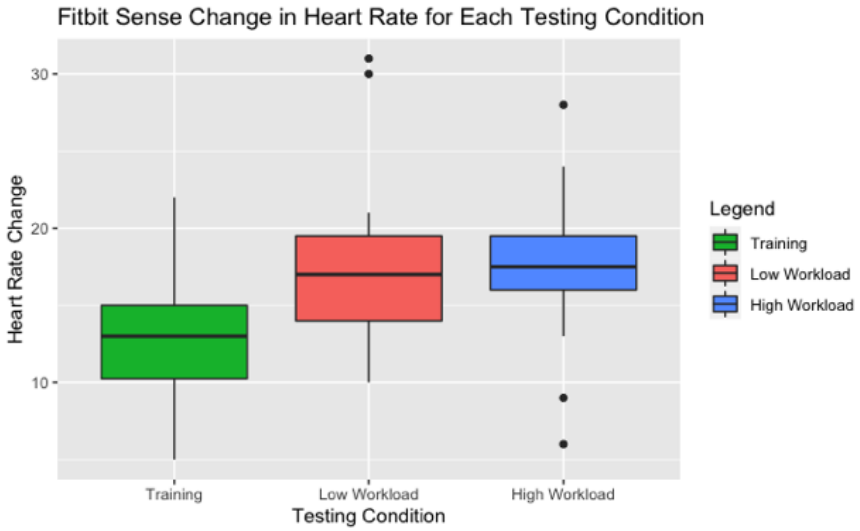

Figure 6. Differences in HR Increase between training and experimental modules using Fitbit Sense.

AliveCor Kardia vs. Apple Watch Series 6

The data gathered from both devices during each test condition was also evaluated through the use of paired samples t-tests. A significant difference was found between the HR data gathered during low workload condition from the Kardia ECG (M = 73.143, SD = 10.250) and the HR data gathered during low workload condition from the Apple Watch Series 6 ECG (M = 69.150, SD = 10.058) $t(13) = 2.9954$, $p = 0.0103$.

Using the same type of test, there was found to be no significant difference between the HR data gathered during the high workload condition from the Kardia ECG (M = 74.231, SD = 11.476) and the HR data gathered during the high workload condition from the Apple Watch Series 6 ECG (M = 72.451, SD = 11.803). From this result, an equivalency test was run to see if the two data streams could in fact be considered equal. Using calculated effect sizes of Cohen's d = -0.293 and d = 0.293, it was found that the observed effect falls outside the equivalence bounds and is not close enough to zero to be considered equivalent at the 90% confidence interval. (Seaman & Serlin, 1998). However, this may be due to the limited sample size.

Heart Rate Variability (HRV)

Heart rate variability was calculated using the Root Mean Square of Successive Differences (RMSSD) of the R-R intervals of heartbeats during the various ECG recordings. RMSSD was chosen because it has been shown to be the most accurate during short (30s or less) HRV measures (Thong et al.,

2003). In order to isolate the R-R intervals from the raw data, it was necessary to screen for local peaks and amplify the largest ones (QRS complex) (see Figure 7).

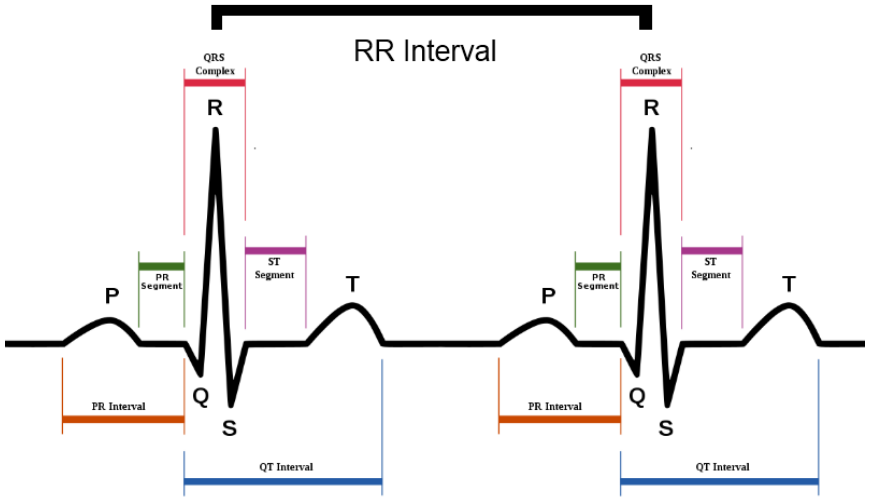

Figure 7. Depiction of QRS complex and R-R interval on example ECG data (Tawakal et al., 2012).

A threshold in which all amplified QRS complexes surpassed could then be determined so that any point in time with an ECG measurement above that threshold indicated the R portion of the QRS complex (Figures 8-10). The R-R intervals could then be identified and RMSSD could be calculated.

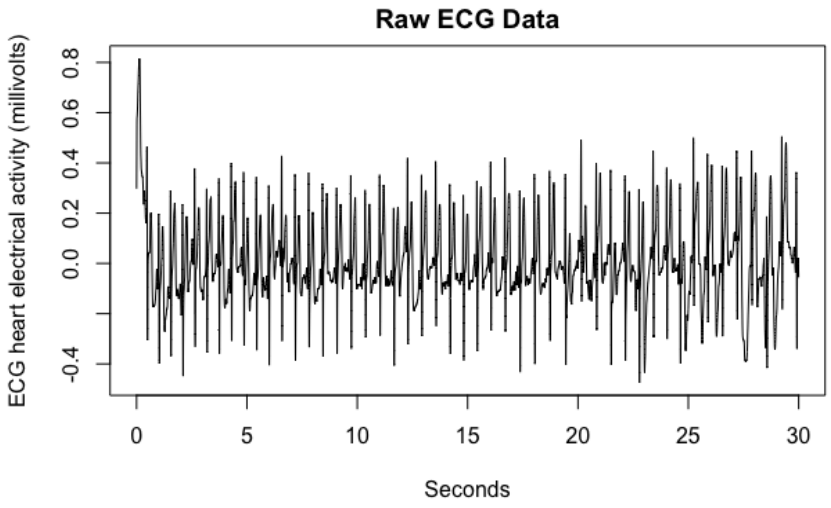

Figure 8. Example of raw ECG data gathered during low workload condition from Participant 16 using the AliveCor Kardia ECG.

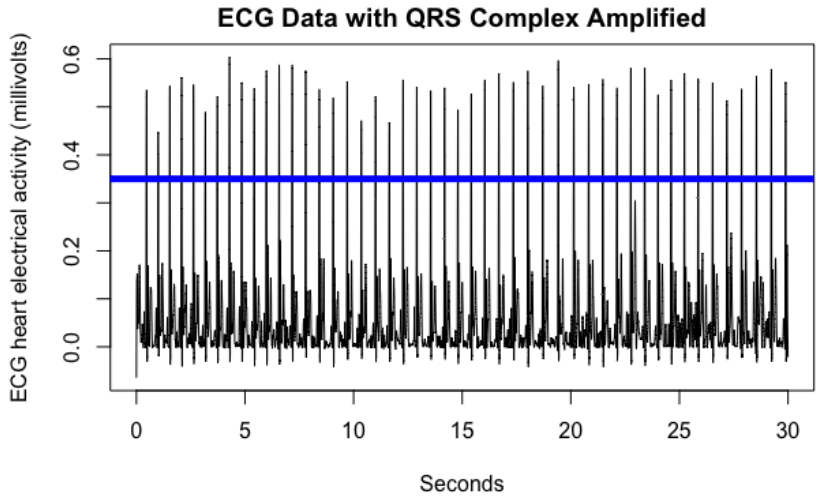

Figure 9. Low workload condition, Participant 16 AliveCor Kardia ECG data with QRS complexes amplified; Blue line indicates threshold in which only ECG R-waves surpass.

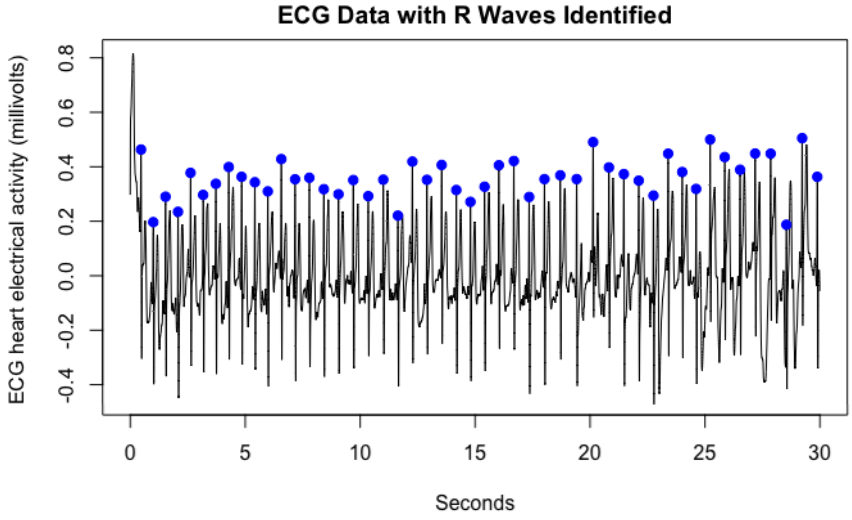

Figure 10. Low workload condition, Participant 16 AliveCor Kardia ECG data with R-R intervals identified; Blue dots represent the tops of the R-waves for each QRS complex.

AliveCor Kardia

Paired samples t-tests between the baseline ECGs (M = 0.0649, SD = 0.0413) and the two testing conditions, low workload Kardia ECGs (M = 0.0566, SD = 0.0317) and high workload Kardia ECGs (M = 0.0556, SD = 0.0413) showed no significant differences in RMSSD, $t(13) = 1.9874$, $p = 0.0684$, and $t(11) = 2.0088$, $p = 0.0698$, respectively. There was also no significant difference between low workload and high workload condition RMSSD using the Kardia ECG, $t(12) = 0.06172$, $p = 0.9518$.

Apple Watch Series 6

Paired samples t-tests between the baseline ECGs (M = 0.06492, SD = 0.0413) and the two testing conditions, low workload Apple Watch Series 6 ECGs (M = 0.06487, SD = 0.04284) and high workload Apple Watch Series 6 ECGs (M = 0.05699, SD = 0.03975) showed no significant differences in RMSSD, $t(12) = 0.33255$, $p = 0.7452$, and $t(13) = 0.63559$, $p = 0.5361$, respectively. There was no significant difference between low workload and high workload condition RMSSD using the Apple Watch Series 6 ECGs, $t(12) = 0.51582$, $p = 0.6154$, but due to the fact that it could not be shown that the differences between the pairs are normally distributed via the Shapiro-Wilk normality test, $W = 0.83791$, $p = 0.02$, this observation cannot be considered to be valid.

AliveCor Kardia vs. Apple Watch Series 6

The data gathered from both devices during each test condition was also evaluated through the use of paired samples t-tests. No significant differences were found between the RMSSD data gathered during the low workload (M = 0.0566, SD = 0.0317) or high workload (M = 0.0556, SD = 0.0413) condition using the Kardia ECG compared to the low workload (M = 0.06487, SD = 0.04284) or high workload (M = 0.05699, SD = 0.03975) condition using the Apple Watch Series 6, $t(13) = -0.83588$, $p = 0.4183$, and $t(12) = 0.4734$, $p = 0.644$, respectively. From these results, equivalency tests were run to see if the data streams from both devices could in fact be considered equal. Using calculated effect sizes of Cohen's d = -0.223 and d = 0.223 for low workload condition, and Cohen's d = -0.131 and d = 0.131 for high workload condition, it was found that the observed effect falls outside the equivalence bounds for both tests and is not close enough to zero to be considered equivalent at the 90% confidence interval. (Seaman & Serlin, 1998). However, this may be due to the limited sample size.

NASA Task Load Index (NASA-TLX)

Participants filled out a NASA-TLX questionnaire at the conclusion of each experimental portion of the MATB-II program to gauge subjective workload and stress levels. When comparing the responses from the low workload condition to the high workload condition through the use of paired samples T-Tests, all measured categories showed significant differences (all $p < 0.002$) (Figure 11).

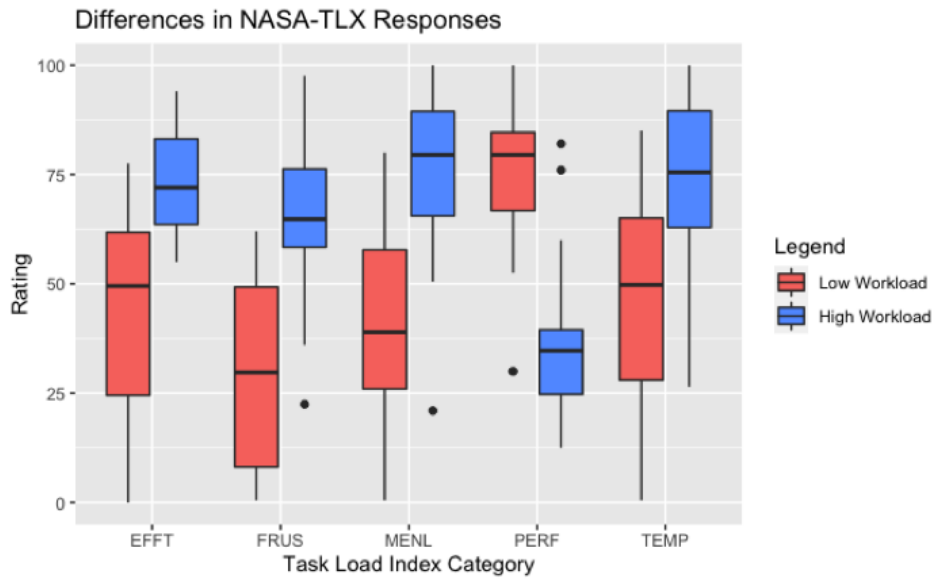

Figure 11. Differences in NASA-TLX responses by category from low workload condition to high workload condition; EFFT: Effort, FRUS: Frustration, MENL: Mental Demand, PERF: Performance, TEMP: Temporal Demand.

Effort (EFFT) (M = 28.821, SD = 25.966), $t(13) = 4.1531$, $p = 0.0011$, Frustration (FRUS) (M = 34.107, SD = 22.677), $t(13) = 5.6277$, $p < 0.001$, Mental Demand (MENL) (M = 34.107, SD = 26.245), $t(13) = 4.8625$, $p < 0.001$, and Temporal Demand (TEMP) (M = 28.50, SD = 26.997), $t(13) = 3.9499$, $p = 0.0017$, all showed significant increases in response ratings from the low workload condition to the high workload condition. Performance (PERF) (M = -37.53, SD = 20.093), $t(13) = -6.9899$, $p < 0.001$, showed a significant decrease in rating from low workload condition to high workload condition. This shows that the perceived significant decrease in performance was accurately reflected by the actual MATB-II task scores of the two workload conditions. Physical Demand (PHYS) was not included in evaluation because it could not be shown that the differences between the pairs are normally distributed via the Shapiro-Wilk normality test, $W = 0.81708$, $p = 0.008$.

Debriefing Questionnaire Responses

Participants were asked to rate how stressed they currently felt on a scale from 0-10 (0 being not stressed at all, 10 being most stressed) as part of the demographic survey at the beginning of the experimental session, and again upon completion of all trials during the debriefing questionnaire at the end of the experimental session. Pre-trial responses to this question resulted in an average stress level of 3.625 (mode = 5). Post-trial responses resulted in an average stress level of 4.625 (mode = 6, 7).

DISCUSSION

The goal of this research is to determine the extent to which two of the leading smartwatches in the technology industry, the Apple Watch Series 6 and the Fitbit Sense, can accurately detect and measure stress responses as a function of workload. Additionally, this research aimed to determine the accessibility of measured user health data so that the average user may better understand how their body responds to short-term stress.

Effects of Workload Changes

The results of this study confirmed that heart rate is affected by workload and supports our first hypothesis: as user workload increases, user stress responses increase. Heart rate variability was impacted by increases in workload in this study, but further experimentation with an increase in sample size could lead to significance, as $p < 0.07$ both low and high workload conditions. The participant responses from the NASA-TLX questionnaire demonstrate a significant increase in participant perceived workload from the low workload condition to the high workload condition. Additionally, pre-trial and post-trial questionnaires show a subjective increase in user overall stress levels on a 10-point scale. Pairing these findings with the stress responses measured using the AliveCor Kardia ECG during the low workload condition and the high workload condition, our second hypothesis is supported: an increase in perceived user stress will result in a measurable increase in physiological stress responses. This body of work demonstrates that workload affects stress levels and that changes in HR and HRV hold the potential to accurately measure stress responses in humans (Taelman et al., 2009; Okada et al., 2013; Cranwell-Ward & Abbey, 2005; Wainwright & Calnan, 2002).

Data Accessibility of Each Device

AliveCor Kardia

The AliveCor Kardia device is a mobile, FDA-approved ECG device that was used as the official ECG comparison to the smartwatches selected in this study. Each ECG that was taken with this device is automatically uploaded to the Kardia application on a connected smartphone via bluetooth and to the AliveCor Kardia official website (https://app.alivecor.com/). Users can easily view and extract the ECG data in the form of a European Data Format (EDF) file, which is the standard format used for the exchange and storage of medical time series data (Kemp & Olivian, 2003).

Apple Watch Series 6

The Apple Watch Series 6 has the ability to constantly track user heart rate and take an ECG in the span of 30 seconds. Heart rate and ECG information is automatically uploaded to the Apple Health

application on a connected smartphone. Continuous heart rate data and summaries can be viewed on the Apple Health application in increments as short as 1 minute, but cannot be extracted for raw data analysis. HRV can be calculated over extended periods of time wearing the device. ECG information can be exported and mailed to a desired location and comes in the form of a comma-separated values (CSV) file composed of heart electrical activity readings (measured in millivolts) paired with time stamps.

Fitbit Sense

The Fitbit Sense has the ability to constantly track user heart rate and take an ECG in the span of 30 seconds, similar to the Apple Watch Series 6. Heart rate and ECG information is automatically uploaded to the Fitbit application on a connected smartphone. Continuous HR data and summaries can be viewed on the Fitbit application. Through the use of a third-party application (Pulse Watch), users can extract all measured HR data in increments as small as 5 seconds. HRV can be calculated over extended periods of time wearing the device. ECG information is stored in the form of a Portable Document Format (PDF). The raw data behind the ECG readings are unable to be extracted for data analysis.

For all devices utilized in this study, HRV was calculated by converting the data into a time series format, screening for local peaks and amplifying the largest ones (QRS complex), and then isolating the points in time in which the peaks surpassed a determined threshold, revealing the R-R intervals so that RMSSD could be calculated. This method is mentioned in depth in the Results section.

Data Stream Comparison

The AliveCor Kardia device showed a significant increase in stress responses during low workload and high workload test conditions when compared to the baseline, but no significant difference between the low workload condition and high workload condition directly. This suggests that participants were equally stressed for the low workload as they were for the high workload based on heart rate, despite the difference in demand. This could be attributed to the fact that both the low workload and high workload conditions were being scored, and that a financial reward was said to be given to the highest performers on these trials. The promise of a reward based on the level of success creates a high-stakes testing environment, which has been shown to increase levels of intrinsic motivation and stress (Tagher & Robinson, 2016; Cameron et al., 2005). During the low workload and high workload conditions, the Apple Watch Series 6 was unable to detect a significant difference in stress responses as compared to the baseline measurements. When comparing the low workload and high workload conditions directly, the Apple Watch Series 6 detected a significant increase in HR, but no significant change in HRV. The significant increase in HR could be explained by a number of variables such as the limited number of participants, extraneous movement during ECG measurement, or time delay from trial end to ECG start,

among others. Equivalency tests between the AliveCor Kardia ECG datastreams and the Apple Watch Series 6 ECG datastreams show that they cannot be considered equivalent, which contradicts our third hypothesis: the data streams produced from the smartwatch ECGs will align with the data streams from the official, FDA-approved ECG device. These findings contribute to a new and growing body of work that aims to test the validity of health features on wrist-wearable smart devices. Further, this research supports the findings of variability in data collection in current smartwatches (Siirtola, 2019; Ciabattoni, 2017).

Limitations and Future Work

This study was affected by multiple limitations which must be considered. It is important to consider the differences in confidence levels of participants and their individual perception of how equipped they are to handle the MATB-II tasks presented to them. It has been shown that if an individual feels they have the necessary skills or tools to achieve a task, their experience will be less stressful when compared to an individual that has a lower confidence level, regardless of the workload intensity (Selye, 1984). This can result in variability in data collection between participants, even though each participant was subjected to the same two MATB-II evaluations.

Another limitation to this study was the availability of the data streams from the two industry leading smartwatches. As mentioned in the Data Stream Accessibility section, the Apple Watch Series 6 is unable to provide extractable continuous HR data, and the Fitbit Sense is unable to provide extractable ECG data. Due to the limitations of the devices used, we are not currently able to directly compare the two devices in regards to HR and HRV in a testing environment. This limitation may reduce the potential knowledge gained in the understanding of the stress detection capabilities of these devices. The data streams that were made available for data analysis are not exported in a uniform file format, so there is also potential for error in the data conversion process.

The ECG data in this study was collected immediately following the completion of each trial in order to capture the stress responses directly elicited by the tasks of the MATB-II evaluation. As a result of this, we must consider the limitation that the smartwatches are unable to passively record ECGs while the participant is actively taking the examination. To take an ECG in this experiment, the participant had to stop what they were doing, place their index finger on the smartwatch sensors, and remain still in this position for the duration of the ECG recording (30 seconds). It has been shown that heart rates can decrease up to 22 beats per minute (bpm) in 60 seconds of recovery, so it may be likely that by the time the ECGs have finished recording, the participant was not at their peak stress level (Shetler et al., 2001). This claim is supported by the fact that on average, the HRs at the time of ECG recording were 13 bpm less than the maximum HRs experienced during the examination periods in this study. This can help

explain why there were no significant detectable differences in HRV between the low workload and high workload conditions.

Finally, the ECG data is affected by user movement during the ECG recording process, as the devices are non-intrusive and rely on tactile contact with the user to generate a recording. It has been shown that there is a significant correlation between HRV measurement error and user movement, so any movement or brief lack of contact with the sensors in the smartwatches could contribute to error in the data (Maritsch et al., 2019).

It is critical for future work to be informed of potential sources of error in the data production and collection processes with the smartwatch technology that is currently available to the public. For these reasons, we recommend that future research gathers data on a larger scale to offset some of the variations in data seen in this study. This will hopefully highlight the capabilities and validity of the extractable health metrics in these devices as technology continuously improves.

KEY POINTS

- Stress responses were highest during examination conditions, supporting previous work showing that workload can affect stress responses in the human body.
- Measured stress responses aligned with the participant's perception of how hard they were working and how stressed they were feeling. This supports assertions from previous work measuring changes in heart rate and heart rate variability that these measures hold the potential to accurately measure stress responses in humans.
- The accessibility of data gathered using current smartwatches for in depth analysis is very limited, and these devices are susceptible to variability when heart measurements are being taken.
- Stress and the ability to detect stress-related responses is crucially important, and smartwatches must be considered as viable health tracking options, especially as technology continuously improves.

REFERENCES

Apple Inc. (2020). *Apple Watch Series 6: User Guide.* Cupertino, CA: Author.

Cameron, J., Pierce, W. D., Banko, K. M., & Gear, A. (2005). Achievement-based rewards and intrinsic motivation: A test of cognitive mediators. *Journal of educational psychology*, *97*(4), 641.

Childs, E., White, T. L., & de Wit, H. (2014). Personality traits modulate emotional and physiological responses to stress. *Behavioural pharmacology*, *25*(5 0 6), 493.

Chuah, S. H. W., Rauschnabel, P. A., Krey, N., Nguyen, B., Ramayah, T., & Lade, S. (2016). Wearable technologies: The role of usefulness and visibility in smartwatch adoption. *Computers in Human Behavior*, *65*, 276-284.

Ciabattoni, L., Ferracuti, F., Longhi, S., Pepa, L., Romeo, L., & Verdini, F. (2017, January). Real-time mental stress detection based on smartwatch. In *2017 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 110-111). IEEE.

Cranwell-Ward, J., & Abbey, A. (2005). The Most Stressful Jobs. In *Organizational Stress* (pp. 63-71). Palgrave Macmillan, London.

Fitbit LLC. (2020). *Fitbit Sense: User Manual.* San Francisco, CA: Author.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2014, September). Workload overload modeling: An experiment with MATB II to inform a computational model of task management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, No. 1, pp. 849-853). Sage CA: Los Angeles, CA: SAGE Publications.

Harwood, J., Dooley, J. J., Scott, A. J., & Joiner, R. (2014). Constantly connected–The effects of smart-devices on mental health. *Computers in Human Behavior*, *34*, 267-272.

Johnson, S., Cooper, C., Cartwright, S., Donald, I., Taylor, P., & Millet, C. (2005). The experience of work‐related stress across occupations. *Journal of managerial psychology*.

Kemp, B., & Olivan, J. (2003). European data format 'plus'(EDF+), an EDF alike standard format for the exchange of physiological data. *Clinical neurophysiology*, *114*(9), 1755-1761.

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry investigation*, *15*(3), 235–245. https://doi.org/10.30773/pi.2017.08.17

Maritsch, M., Bérubé, C., Kraus, M., Lehmann, V., Züger, T., Feuerriegel, S., ... & Wortmann, F. (2019, September). Improving heart rate variability measurements from consumer smartwatches with machine learning. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (pp. 934-938).

Nidiffer, F. D., & Leach, S. (2010). To hell and back: Evolution of combat-related post traumatic stress disorder. *Dev. Mental Health L.*, *29*, 1.

Okada, Y., Yoto, T. Y., Suzuki, T. A., Sakuragawa, S., Sakakibara, H., Shimoi, K., & Sugiura, T. (2013). Wearable ECG recorder with acceleration sensors for monitoring daily stress. *J. Med. Biol. Eng*, *33*(4), 420-426.

Quick, J. D., Horn, R. S., & Quick, J. C. (1987). Health consequences of stress. *Journal of Organizational Behavior Management*, *8*(2), 19-36.

Reeder, B., & David, A. (2016). Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of biomedical informatics*, *63*, 269-276.

Richter, M., & Gendolla, G. H. (2009). The heart contracts to reward: Monetary incentives and preejection period. *Psychophysiology*, *46*(3), 451-457.

Robertson, M. F., & Ruiz, L. E. (2010). Perceptions of stress among collegiate aviation flight students. *The Collegiate Aviation Review International*, *28*(1).

Schwerdtfeger, A., & Friedrich-Mai, P. (2009). Social interaction moderates the relationship between depressive mood and heart rate variability: evidence from an ambulatory monitoring study. *Health Psychology*, *28*(4), 501.

Samaha, M., & Hawi, N. S. (2016). Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. *Computers in human behavior*, *57*, 321-325.

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock Jr, J. R. (2011). The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide.

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological methods*, *3*(4), 403.

Selye, H. (1984). *The stress of life*. McGraw-Hill.

Selye, H. (1956). What is stress. *Metabolism*, *5*(5), 525-530.

Seoane, F., Mohino-Herranz, I., Ferreira, J., Alvarez, L., Buendia, R., Ayllón, D., ... & Gil-Pita, R. (2014). Wearable biomedical measurement systems for assessment of mental stress of combatants in real time. *Sensors*, *14*(4), 7120-7141.

Shetler, K., Marcus, R., Froelicher, V. F., Vora, S., Kalisetti, D., Prakash, M., ... & Myers, J. (2001). Heart rate recovery: validation and methodologic issues. *Journal of the American College of Cardiology*, *38*(7), 1980-1987.

Siirtola, P. (2019, September). Continuous stress detection using the sensors of commercial smartwatch. In *Adjunct Proceedings of the*

2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (pp. 1198-1201).

Stanford University. (2007, February 25). Why Do Humans And Primates Get More Stress-related Diseases Than Other Animals?. *ScienceDaily*. Retrieved October 26, 2021 from www.sciencedaily.com/releases/2007/02/070218134333.htm

Taborsky, B., English, S., Fawcett, T. W., Kuijper, B., Leimar, O., McNamara, J. M., ... & Sandi, C. (2021). Towards an evolutionary theory of stress responses. *Trends in ecology & evolution*, *36*(1), 39-48.

Taelman, J., Vandeput, S., Spaepen, A., & Van Huffel, S. (2009). Influence of mental stress on heart rate and heart rate variability. In *4th European conference of the international federation for medical and biological engineering* (pp. 1366-1369). Springer, Berlin, Heidelberg.

Tagher, C. G., & Robinson, E. M. (2016). Critical aspects of stress in a high-stakes testing environment: A phenomenographical approach. *Journal of Nursing Education*, *55*(3), 160-163.

Tawakal, M. I., Suryana, M. E., Noviyanto, A., Satwika, I. P., Alvissalim, M. S., Hermawan, I., ... & Jatmiko, W. (2012, December). Analysis of multi codebook GLVQ versus standard GLVQ in discriminating sleep stages. In *2012 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 197-202). IEEE.

Thong, T., Li, K., McNames, J., Aboy, M., & Goldstein, B. (2003, September). Accuracy of ultra-short heart rate variability measures. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)* (Vol. 3, pp. 2424-2427). IEEE.

Wainwright, D., & Calnan, M. (2002). *Work stress: The making of a modern epidemic*. McGraw-Hill Education (UK).