A Validation Study: Fitbit Charge 2 Heart Rate Measurement at Rest and During Cognitive-Emotional Stressors

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN KINESIOLOGY & HEALTH SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO
August 2021

ABSTRACT

Previous studies emphasize concurrent validity of wearable devices during physical activity. We investigate device agreement between Fitbit Charge 2 (FBC2) Heart Rate (HR) and electrocardiography (ECG) under sedentary conditions. In a sample of 32 post-secondary students, agreement was observed during rest, negative affect images, deceptive pattern solving task, and a listening task. ECG HR and FBC2 HR was sampled continuously and transformed into 1120 paired 1-min epochs throughout a 35-minute protocol. Bland-Altman were adopted to evaluate agreement. FBC2 HR underestimated ECG HR as indicated by a mean error bias = -1.1 bpm (95% limits of agreement, -5.62 to 3.42), mean absolute percentage error 1.34 (SD = 1.85) %. Bland-Altman limits of agreement plot indicated minimal systematic error in the measurement range of 54 bpm to 117 bpm. When compared to gold-standard ECG HR the FBC2 HR demonstrated good agreement in healthy young adults.

ACKNOWLEDGEMENTS

I learned a great deal of interdisciplinary skills needed to investigate this research topic as a MSc thesis. I couldn't have accomplished these without my support system of professors, family, friends, and colleagues who helped me see this project through to completion. I gratefully acknowledge the assistance of all the persons and volunteers whose participation was essential in the successful completion of the study.

To my graduate supervisor Dr. Paul Ritvo and committee member Dr. Joel Katz, I thank you for exercising the patience and precision needed to foster progress through the many phases of this project. I would also like to thank Dr. Hugh McCague and the Statistical Consulting Service offered by the Institute for Social Research at York University for making statistics less intimidating and more understandable.

I thank my lab mates; Megan Kirk-Chang, Kevin Dang, Daniel Perez, Meysam Pirbaglou, Slobodan Stojanovic, Vina Mohabir, Zach Kelly, and Sabina Mirza. Persevering through the highs and lows of graduate school wouldn't have been possible without your support.

TABLE OF CONTENTS

LIST OF TABLES:

LIST OF FIGURES:

KEY TERMS & ABBREVIATIONS

ANSI-CTA	American National Standards Institute/ Consumer Technology Association
BA	Bland-Altman
bpm	Beats per minute
ECG	Electrocardiogram
HR/RHR	Heart Rate/ Resting Heart Rate
HRR	Heart Rate Reactivity
HRVO	Heart Rate Volatility
LOA	Limits of Agreement
ME	Mean Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
PPG	Photoplethysmography
WOHRM	Write-Worn Optical HR Monitoring Devices

1.0 INTRODUCTION

The increasing quantity of research on wearable (Fitbit) devices is indicated by the frequency of research studies (PubMed: 476 studies, ClinicalTrials.gov: 449 studies) on Fitbit-branded devices (Bent et al., 2020). According to the Fitabase Research Library (a repository for peer-reviewed Fitbit publications), only 25 validation studies (2016 – 2020) have focused on the Fitbit PurePulse^{TM} capacities in estimating HR (HR) (*Fitabase*, 2020). These proportions (476 research publications of which 25 pertain to validation) reflect a high ratio of studies that employ Fitbit-based measures in relation to studies that estimate how accurate and reliable Fitbit-based measures are. While this may seem the reverse of an optimal ordering (where measurement validation precedes measurement utilization), it's not uncommon due to the dual factors of publication pressure and insufficient distributions of statistical expertise (Bland & Altman, 1986a; Zaki et al., 2013; Gerke, 2020). In a letter, the British Journal of Sports Medicine called for 'evidence-based marketing claims', recommending that devices undergo rigorous evaluation prior to being launched in the market place, to ensure that wearable technology can be used safely and to its full potential (Sperlich & Holmberg, 2017). One of the scientists cited above (DA Altman) was the co-originator of the Bland-Altman Limits of Agreement (BA LOA) method of assessing measurement device validity, often used when a newer 'candidate' device is compared with a gold standard device (Bland & Altman, 1986a). Arguably, this method is the most thorough yet proposed and utilized currently (Gerke, 2020) and the method that is applied in this thesis in evaluating the validity of the Fitbit Charge 2 heart rate (FBC2 HR) estimations. Testing protocols to examine validity have been performed under sedentary conditions with a few studies emphasizing experimental stressor responses. In this thesis FBC2 HR measurements are compared with simultaneous electrocardiographic measurements (ECG HR). The specific

thesis objectives are to: (1) assess measurement agreement between FBC2 HR and ECG HR, and (2) to detect potential influences on measurement agreement due to cognitive-emotional stressors within a controlled lab setting.

2.0 LITERATURE REVIEW

2.1 Characteristics of Wrist Optical Heart Rate Monitoring (WOHRM) devices

Photoplethysmography (PPG) is a low-cost optical technique that can be used to detect blood volume changes in the microvascular bed of tissue. The PPG technology was first introduced to study skin vasculature in cold temperatures (Hertzman, 1938) but is commonly used to estimate oxygen saturation and blood pressure (Allen, 2007; Nitzan et al., 2014) and more recently, pulse rate, in commercial wearable devices (e.g. Fitbits, Apple Watches, Garmin devices) (Mühlen et al., 2021). Resting pulse rate is one of the key vital signs routinely measured in clinical practice. Although both PPG and ECG signals convey physiological information, the underlying physiology of PPG stems primarily from hemodynamics rather than the electrical activity of the heart reflected in the ECG signal. Pulse timing, amplitude and shape features compared to ECG are illustrated in Figure 1.

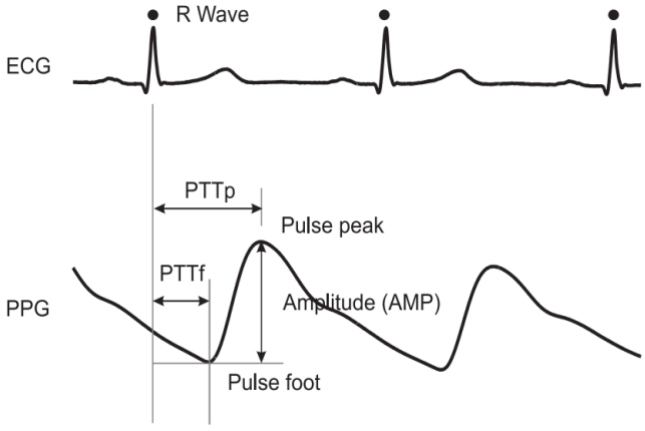

Figure 1. Characterization of PPG pulse timing, amplitude, and shape features as compared to ECG (Allen, 2007). The latency (pulse rate transit time) between biosignals is apparent when measuring HR and pulse rate simultaneously and continuously.

WOHRM devices calculate pulse rate when light, defined by varying wavelengths along a red-to-ultraviolet spectrum, penetrate human skin (Sliney, 2016). Depending on the wavelengths used to illuminate the skin, there are different capacities to monitor biological reactions relevant to pulse rate (Tsibadze et al., 2015). For example, *red* illumination is commonly used in pulse oximetry assessments, given its capability for penetrating deeply into living tissue (0.8-1.5mm), allowing for digitally-based observations of blood flowing within multiple arteries. In contrast, *green* light (540 nm) is absorbed well by the blood does not extend beyond a depth of 0.6 mm sampling changes in blood volume within the capillaries instead (Kamshilin & Margaryants, 2017). Commercial wearable devices that employ PPG use *green* light to estimate the radial pulse rate although the software algorithms to process this signal are proprietary and may vary by manufacturer. A simple illustration in Figure 2 depicts green light refracting from the blood back to the sensor.

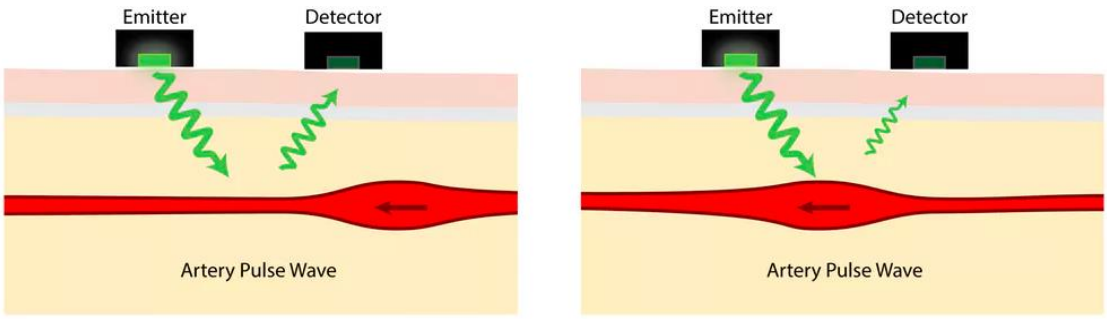

Figure 2. Visual depiction of green light refraction (i.e. pulse waves of blood absorbing light) (Pires et al., 2018).

2.2 Literature on Wrist Optical Heart Rate Monitoring (WOHRM) devices

This systematic review is, to some degree, based on a past systematic literature review undertaken by Mühlen et al. (2021), with the aim of establishing optimal practices for the measurement validation of consumer wearable heart rate devices (Mühlen et al., 2021). Their goal was to review all relevant published literature on PPG-based wearable HR devices (Mühlen et al., 2021). We updated their search by extending the end-date range of publications from March 2020 to June 2021, using a search strategy based on the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA). In the initial Mühlen review, PubMed, Web of Science, and Embase databases were accessed to identify relevant records meeting search criteria. In the updated review Embase was restricted due to home institution licencing access, while the two other databases were employed (PubMed, Web of Science).

Given the thesis focus, we narrowed the Mühlen et al., 2021 inclusion criteria to studies that solely investigated agreement between WOHRM and ECG HR monitoring during sedentary behaviours in adults (Tremblay et al., 2017). Both active and passive sedentary behaviours were included (e.g. sitting versus sitting and activity). However, if the task involved physical activity or recovery following an identified physical activity event, the study was not considered relevant to validating WOHRM devices for sedentary monitoring. Screening and data extraction were performed using an open-source reference software (Zotero). The process entailed saving search references, removing duplicates, and then screening for relevant articles.

The updated search led to the identification of N = 977 additional records (May 2020 – June 2021). Duplicates were identified (N = 30) and removed, leaving a total of 947 records. Then studies identified in the Mühlen et al. systematic review (N=44) and additional records (N=947) from the expanded search were pooled together (N= 991) and title/abstract screening

proceeded with three reporting criteria: 1) validation of WOHRM device(s), 2) ECG comparator, and 3) error bias calculations. As a result of this tripartite screening, 11 full-text studies were assessed for eligibility. From these 11 studies, two studies were excluded due to an assessment during exercise recovery (Jo et al., 2016; Falter et al., 2019), and 1 study was excluded for not reporting the protocol duration (Sequeira et al., 2019). In addition, 1 study reported a protocol duration that was less than the American National Standards Institute/ Consumer Technology Association (ANSI/CTA) standard of at least 5 minutes of testing per participant (*Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065)*, n.d.). This left a total of 7 studies for data extraction and literature review synthesis (Kroll et al., 2017; Dur et al., 2018; Bent et al., 2020; Hahnen et al., 2020; Konstantinou et al., 2020; Menghini et al., 2019; Schuurmans et al., 2020). Figure 3 outlines the steps involved in the inclusion process.

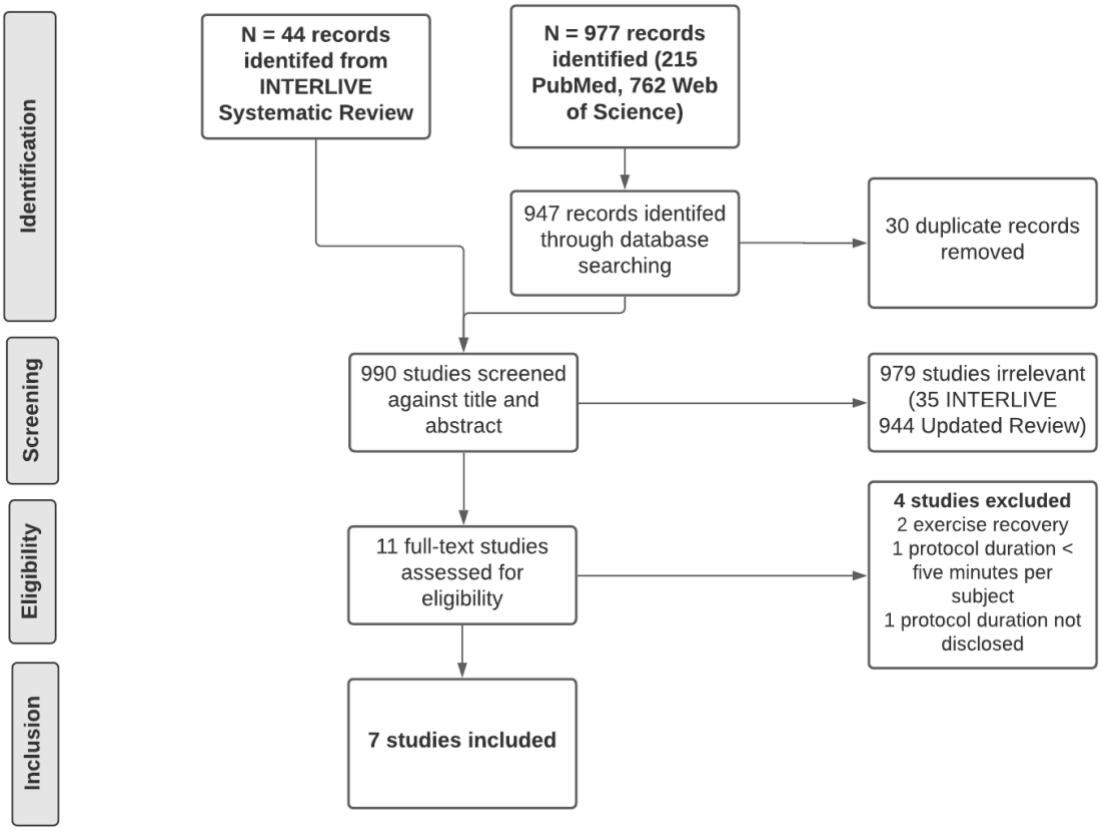

Figure 3. Diagram of steps involved in the identification, screening, and inclusion of relevant articles

We extracted the following data from each study: the author's last name, publication year, sample size, device type (WOHRM brand), comparator device, task type, number of repeated-observations per participant, testing environment, visual and numerical reporting of error bias and measurement agreement across devices (e.g. Bland-Altman Methods). See Table 1 for more information. The individual studies are discussed in Section 2.5: Individual Validation Study Review

TABLE 1
Data extraction Table of Key Study Characteristics from Systematic Review

Lead author (year of publication)	WOHRM models tested (versus ECG)	Sample Size	Study Tasks	Total Duration of Testing Protocol/Subject for Sedentary Tasks	Estimate Number of Paired-Values (PPG HR – ECG HR)
Bent (2020)	Empatica E4; Apple Watch 4; Round 2: Fitbit Charge 2; Round 3: Garmin Vivosmart 3, Xiaomi Miband, and Biovotion Everion	56	Rest (seated); typing; paced breathing	6 mins	336
Dur (2018)	Wavelet wristband	35	Rest (seated)	15 mins	525
Hahnen (2020)	Everlast TR10 smartwatch	85	Rest (seated)	5 mins	425
Konstantinou (2020)	Microsoft Band 2	43	resting period (baseline phase) and an experiment of pain induction (i.e., cold pressor task).	7 mins	371

Kroll (2017)	Fitbit Charge HR	50	sedentary behaviors while in the ICU	288 mins	14, 400
Menghini (2019)	Empatica E4	40	rest (seated), paced breathing; Stroop test; speech task; keyboard typing	15 mins	600
Schuurmans (2020)	Empatica E4	15	Resting (seated); watch the aquatic video for four minutes; Testing HR and HRV at various timepoints within game-based meditation intervention	23 mins	345

The collection of publications included in this review span nearly a half-decade (2017-2021), approximating the timeline for significant product releases of first-generation commercial fitness trackers with WOHRM capabilities. Six protocols tested validity within a controlled-lab setting, featuring one or multiple sedentary tasks while 1 study conducted testing within a healthcare setting (Kroll et al., 2017). The types of tasks assigned to participants varied: at rest, keyboard typing, paced-breathing, cold pressor test, Stroop test, speech preparation and recovery, watching a video, and meditation. In 2 of the 7 studies, testing involved mental stress and pain induction (Menghini et al., 2019; Konstantinou et al., 2020).

Mean sample size for the aggregate of studies was n = 46.3 (SD 21.5). In six studies testing was completed during a single visit of 23 minutes or less, while in another study testing was performed over a 24-hour period in an adult intensive care unit (Kroll et al., 2017). The median measurement duration for tasks involving sedentary behaviors was 15 minutes (range =

5-288). We noted that some studies did not provide total measurement time per participant, therefore these values were estimated manually, providing the number of participants and total paired-values included in the analysis were reported.

In accord with our inclusion criteria, absolute and relative error calculations (ME = mean error, MAE = mean absolute error, MAPE = mean absolute percentage error) and variance (SD) for sedentary behaviors were reviewed. Briefly put, most WOHRM devices demonstrated high levels agreement of < 5 bpm or < 10%, which is within the error threshold assumed allowable when ECG device reliability is evaluated. We report agreement in more detail in section 2.5 Individual Validation Study Review.

As a consequence of screening articles, we are aware of at least three WOHRM validation studies that reference ECG validation standards (Nelson & Allen, 2019; Bent et al., 2020; Hahnen et al., 2020). These reports are also mentioned in the Mühlen review (Mühlen et al., 2021). Guidelines for physical activity monitoring for HR using PPG and ECG methods are published as white papers with each edition indicating a similar standard with respect to agreement. We provide quotes from two reports. The first report is free to access providing specific cut-off values (as referenced from the original standard for HR accuracy for ECG monitors). Upon further inspection, the original source report is not open-access, and a $250 USD paywall is required to access the information, with a specific copyright clause that limits researchers' ability to publish criteria.

Secondary source:

"Accuracy is computed for the available data as defined above. A [physical monitoring device] PMD is considered accurate if the mean absolute percentage error (MAPE) is ≤10%.

*(*Association for the Advancement of Medical Instrumentation) standard *for HR accuracy for electrocardiography monitors (ECG) see Sections 3.2.11]"* (Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065), n.d.)

Primary source:

"Heart Rate range, accuracy, and QRS detection range [Mean Error] ME EQUIPMENT shall be equipped with means to detect and display the HR. The HR display range shall be at least 30/min to 200/min for adults and 30/min to 250/min for neonatal and pediatric use. The accuracy of the detected HR shall be ±10% or ±5/min, whichever is greater". (ANSI/AAMI/IEC 60601-2-27:2011 (R2016) - Medical Electrical Equipment - Part 2-27: Particular Requirements for the Basic Safety and Essential Performance of Electrocardiographic Monitoring Equipment, n.d.).

After reviewing both reports, we noted no supporting materials or external references were provided to explain how the authors determined error thresholds. We contrast this omission with several decades of method comparison research (Bland & Altman 1986, 1999, 2007, 2012) in medical device tests that include efforts to standardize reporting practices (Gerke, 2020). For this reason, we found it appropriate to seek out additional resources to better understand the relevant comparison research, enabling better understandings of the methods and statistics employed in the seven articles included for review.

2.3 Principles and Methods of Validity Testing

Reliability and *validity* are central research concepts in the social- and health sciences in developing useful measurement instrumentation. *Test-Retest Reliability* is the degree to which a

singularly assessed measure is judged to be sufficiently similar each time utilized. *Validity*, on the other hand, is used to assess how precise the measurement is when compared to competing measures, whereas *utility* refers to how useful the tool is in the field of study (Bannigan & Watson, 2009).

Criterion-related Validity

Of particular interest in this thesis is *criterion-related validity* which is demonstrated by a strong relationship or agreement between scores from two varying measurement approaches or devices. This type of validity is considered most appropriate when the measure of interest meets the criteria of being continuous and quantitative, and when a criterion standard is available (Karras, 1997a). A flow chart is provided below in Figure 4 to illustrate various types of validity

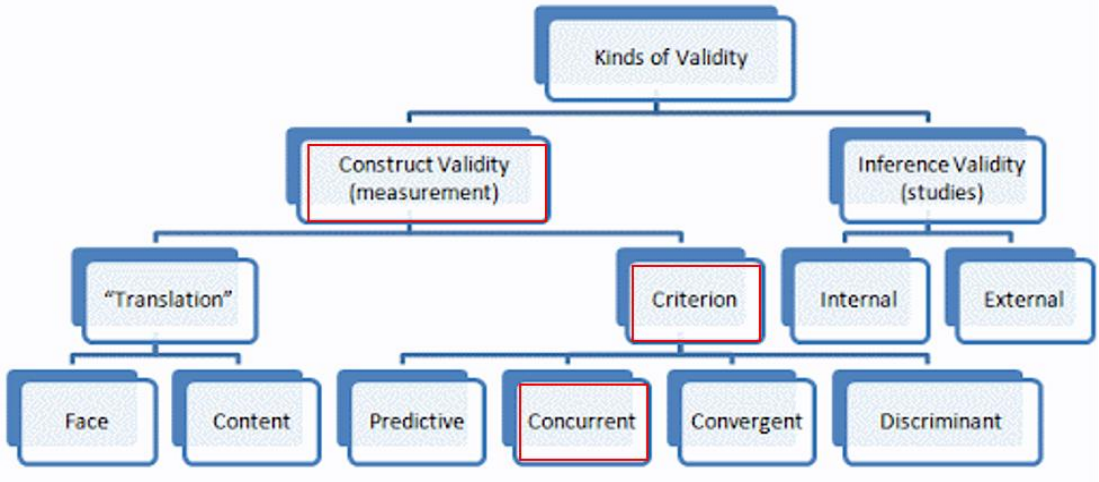

Figure 4. Types of validity. Red annotation indicates the tree-like decision-making structure the researchers consider prior to conducting validity testing

To analyze the criterion-related validity of a measurement procedure, a choice is made to either test *concurrent* or *predictive validity*, or both, in study designs. The related research questions pertain to either predicting an event in the present form (concurrence), or alternatively,

to predicting future events (predictive). Ultimately, the *characteristics* the test purports to measure inform a final judgment on validity assessment (Karras, 1997a).

2.4 The Analysis of Differences: Bland-Altman Methods

Bland-Altman Methods **(**BA methods) offer a verifiable way to quantify and compare agreement parameters that inform clinicians about how closely a new measurement instrument agrees with a historical reference standard (Bland & Altman, 1999). For example, Dorosz et al (2012) performed a systematic review to evaluate the test performance of three-dimensional echocardiography (3DE) (Dorosz et al., 2012). A meta-analysis was performed for 23 studies comparing 3DE and 2DE vs. gold-standard cardiac magnetic resonance imaging in measuring left ventricular volumes and ejection fraction. Search selection criteria indicated that acceptable studies report BA methods of agreement for at least 1 of the 3 measurements "as correlation coefficients alone are misleading" (Dorosz et al, 2012). Mean Error (ME) scores and Limits of Agreement (LOA) were pooled to observe bias and variability of bias for each of the three measurements. The resulting agreement scores from pooled patient samples were considered for relevant clinical applications. This study highlights how BA methods are featured in a meta-analysis within a more mature field of study (cardiac imaging).

The original landmark Bland-Altman study was published in the Lancet and has received considerable attention with nearly 50,000 citations since the initial publication (Bland & Altman, 1986b). The majority of studies (85%) show BA methods are most frequently implemented for medical specialties (Zaki et al., 2012). According to a review of methodological reviews by Gerke 2020, between 2000-2019, seven different review studies developed preferred reporting items to improve the quality of method comparison research (five reviews have been published

since 2015) (Gerke, 2020). Altogether, 13 of 16 key items for reporting proposed by Abu-Arafeh et al. (2016) were considered the most comprehensive and potentially relevant to answer the question "what do authors recommend should be reported when a Bland and Altman analysis is presented?" (Abu-Arafeh et al., 2016). The authors of the aforementioned review systematically consulted N=111 articles to point out that poor reporting makes subsequent meta-analysis difficult. Errors of omission are more easily recognized when a checklist is used. From this review we concluded that the 13-item checklist was appropriate to consider when evaluating the quality of BA LOA analyses and reporting.

Each of the 13-items in the checklist is presented next, in combination with explanatory comments that guide validation perspectives (Abu-Arafeh et al., 2016). In some cases, relevant literature is used to expand on the description provided, followed by a discussion of the findings from our literature review of seven validation studies.

1. **Pre-established acceptable limits of agreement (LOA):** Abu-Arafeh (2016) found that clinically acceptable LOA were not established in 50% of publications reviewed. Of the studies included in our review, none reported acceptable limits (for HR comparisons) *apriori*. Hahnen (2020), for example, applied cut-off values for ECG-ECG method comparisons but did not reference BA methods and reported ±5bpm or ±10% in absolute mean error (Hahnen et al., 2020). Menghini (2019) indicated that LOA were based on 'arbitrary' criteria depending on ECG measures, as used in previous studies (Menghini et al., 2019). Relevant to the reference to 'arbitrariness', many articles that reference the original BA publications assert that LOA are defined by "clinical necessity, biological considerations or other goals" (Giavarina, 2015). Whether intervals are too narrow or too broad is viewed as less a statistical

consideration than a clinical one, dependent on area of study and intended application.

2. **Description of data structure (e.g. number of raters, replicates, block design):** From the studies included, all 7 studies in our review provided descriptions of data structure. Single-paired measurements were collected continuously and concurrently between one or more WOHRM devices and ECG. HR values were transformed to bpm. Biological signals such as HR values are considered not replicable because of multiple inputs that influence innate variations of cardiac output.

3. **Estimations of the repeatability of measurements if possible (mean of differences between replicates and respective standard deviations):** As discussed in item #2, replicates were not observed nor considered appropriate to estimating differences as evidence of repeatability. None of the reviewed studies disclosed information re: repeatability.

4. **Data plot, visually inspected for normality, absence of trend and constant variance across the measurements (e.g. histogram, scatter plot, etc.)**: If plots suggest a trend in the differences, regression analyses may be relevant (Abu-Arafeh et al., 2016). BA methods assume that differences (residual mean error) are normally distributed around the systematic trend (mean of mean error scores). The continuous measurement variables (e.g. FBC2 HR & ECG HR) need not be normally distributed, but their differences should be (Doğan, 2018). If the differences are normally

distributed, 95% of differences will lie between these limits. Three examples of visual inspections using a histogram, Q-Q plot, and a plot of mean differences are illustrated below using sample data from multiple publications.

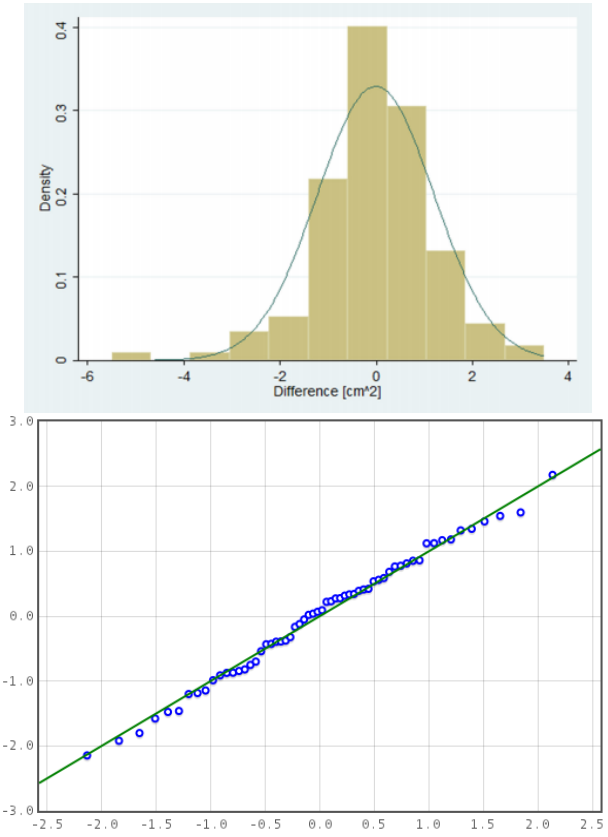

Figure 5 and Figure 6. Top: Histogram with approximating normal distribution of interrater differences (n = 140) (Gerke, 2020). The normal bell-curve is fitted to guide visual inspection. Bottom: Quantile-Quantile Plot Example (sample data: n= 20 individuals with 10 sample paired-measures per individual) (Pleil, 2016)

The purpose of inspecting a BA LOA plot is to discern the extent of systemic bias where the mean difference scores are consistently higher (or lower) than those spanning the measurement range (x axis 0-1200). Proportional bias is evident when mean differences

scores increase (or decrease) as a function of the amount proportional to the level of the measured variable (Ludbrook, 2010).

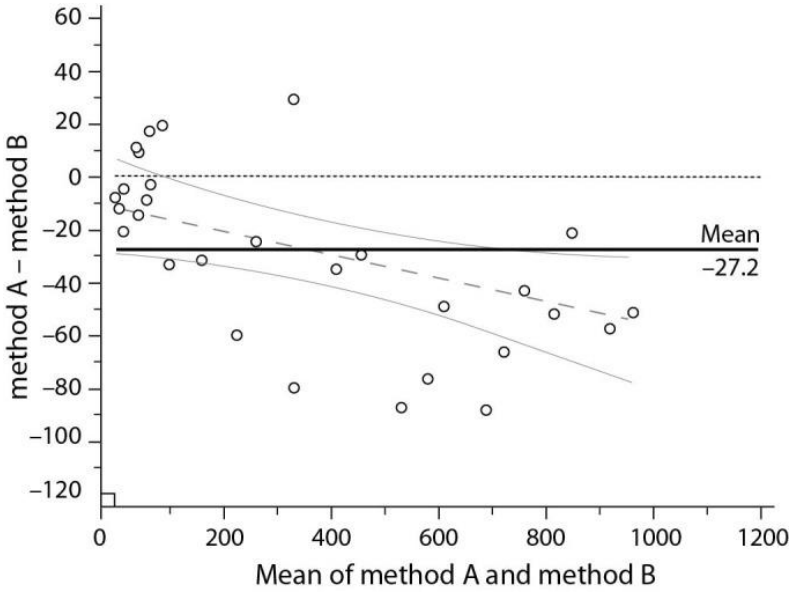

Figure 7. Example plot of mean differences comparing two methods across a range of scores. The solid black line indicates the mean error; the dotted line represents the regression line. Source: (Giavarina, 2015).

In reference to the included articles in this systematic review, 1 of the 7 studies provided some visual inspection for normality, absence of trend, and/or constant variance. Kroll et al (2017) employed density plots for both pooled group analysis and individual LOA plots as well as a histogram of pooled residual error, which, upon visual plot inspection, showed skewed distribution to the left and misleading bin width (below) for n = 14, 400 HR samples (Kroll et al., 2017).

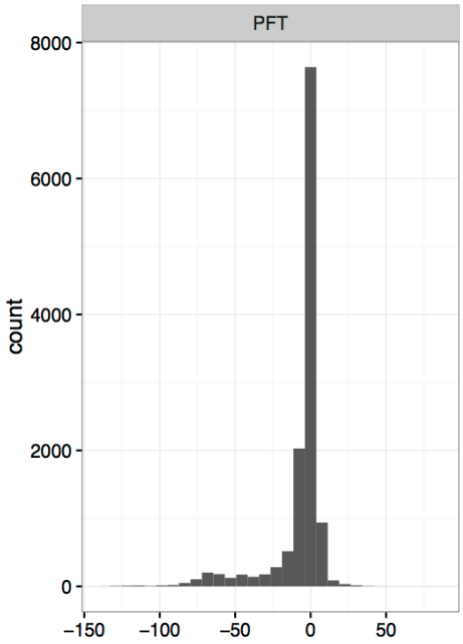

Figure 8. Histogram showing the distribution of the obtained HR differences (x-axis) Fitbit Charge when compared with ECG, in beats per minute. Source: (Kroll et al., 2017).

5. **Transformation of the data (if necessary):** As stated in item #4, normality can be observed by visualizing plots that show dispersion of residuals around the mean error over a measurement range. If differences are not normally distributed, a logarithmic transformation of original data can remove the skewness of the original data. However, this can compromise comparability of findings as the units of measurement are not the same. Accordingly, three studies computed statistical tests to determine assumptions about the distribution of residual error (Kroll et al., 2017; Menghini et al., 2019; Schuurmans et al., 2020). However, only one study transformed the data to establish normality using a logarithmic approach (Menghini et al., 2019).

6. **Distributional assumptions made (e.g., normal distribution of the differences):** A

number of statistical tests can be performed to assess the data for normal distribution. The Shapiro-Wilk test, Kolmogorov-Smirnov test, or D'Agostino-Pearson test can be used when visual evaluation of the histogram plot is inadequate (Giavarina, 2015; Doğan, 2018). Accordingly, two of the reviewed studies reported use of statistical tests to determine assumptions about the distribution of residual error. One study conducted a Wilcox signed test (Kroll et al., 2017) and a second study performed a Kolmogorov-Smirnov test (Schuurmans et al., 2020).

7. **Plotting and numerically reporting the mean of the differences (bias):** Of the studies included for the systematic review, most provided BA LOA plot and reported mean differences. One study showed a visual display of the Mean Error estimates with a BA LOA plot but did not numerically report mean error values. Instead, bias was numerically represented in absolute terms (mean absolute error, mean absolute relative error %) (Hahnen et al., 2020). In another study, mean error calculations were numerically represented, but BA LOA plots were not provided (Bent et al., 2020). Not plotting nor numerically reporting bias limit the comparisons between studies for meta-analytic purposes.

8. **Estimating the precision of the bias, i.e., standard deviation of the differences or 95% CI for the mean difference** Most of the studies provided SD or 95% CI. One study did not (numerically) disclose precision although the BA LOA plot of each task within the protocol was reported (Konstantinou et al., 2020). Another study reported absolute mean error bias and precision of error (SD) (Hahnen et al., 2020) in line with

CTA/AASI Standard (*Physical Activity Monitoring for HR (ANSI/CTA-2065)*, 2018.)

9. **Plotting and numerically reporting the BA LOA:** Of the studies included for the systematic review, 4 of the 7 studies reported both plotting and numerically calculating LOA (Kroll et al., 2017; Dur et al., 2018; Menghini et al., 2019; Schuurmans et al., 2020). One study did not report either of the reporting items (Bent et al., 2020).

10. **Estimation of the precision of the BA LOA by means of 95% confidence intervals:** Adequate evaluation of the results of comparison studies should not only provide the ‘‘limits of agreement’’, but also the precision of these estimates (Olofsen et al., 2015). In the example below, 95% CI LOA are shown as two solid black lines with (light blue areas), and regression fit of the differences on the means (as solid red line). Only 1 of the 7 studies included in the review reported this item (Kroll et al., 2017)

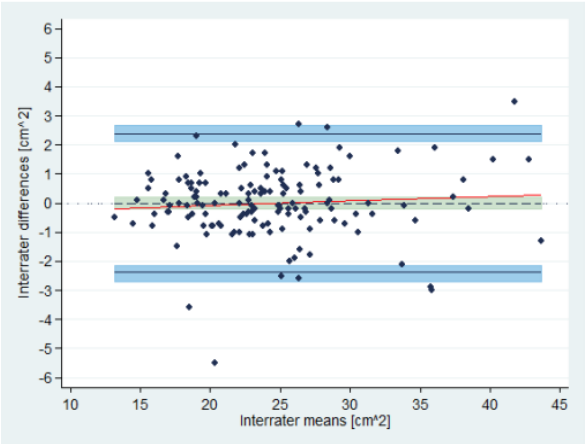

Figure 9. Example Plot of BA LOA with 95% CI for the LOA. Source: (Gerke, 2020)

11. **Assurance that the range of the mean values is sufficient. A narrow range of original values will result in agreement being inevitable:** Preiss and Fisher demonstrate that data range affects the LOA (Preiss & Fisher, 2008). No studies from the review discuss clinical significance of the range measurements observed from testing of HR.

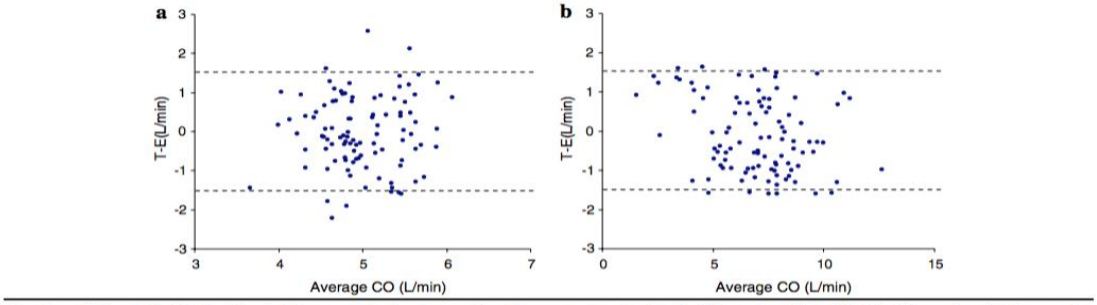

Fig. 1. a (left) and b (right). Bland Altman graph showing results of two experiments (Study #1 on the left and Study #2 on the right) comparing thermodilution (T) to an experimental (E) method for determining cardiac output. In Study #1 the range of cardiac output over which the observations are made is small compared to Study #2. The x-axes are deliberately scaled to make the experiments seem similar at first glance. In both graphs, dashed lines represent 95% limits of agreement.

Figure 10. Example of Two Bland-Altman Plots with Identical Scales.
Source: (Preiss & Fisher, 2008)

12. **Between- and within- subject variance or stating that the confidence intervals of the BA LOA were derived by taking the data structure into account:** This item is relevant when multiple observations are made per individual (i.e. repeated pairs of measurement on the same subject) and compared between individuals. Bland and Altman (2007) discuss their approaches to analyzing repeated measures within subjects for replicated data, or situations whereby the measure is subject to unstable variance (Bland & Altman, 2007). For the 7 studies included in this systematic review, none disclosed variance between- and within-subjects. Two studies provided

individual BA LOA plots for pooled individuals (Kroll et al., 2017; Konstantinou et al., 2020). In a web application: 'Bland-Altman Analysis Guide' by Olofsen (referenced in (Olofsen et al., 2015)) an example plot is provided to visually inspect residual versus subject ID plot (see below).

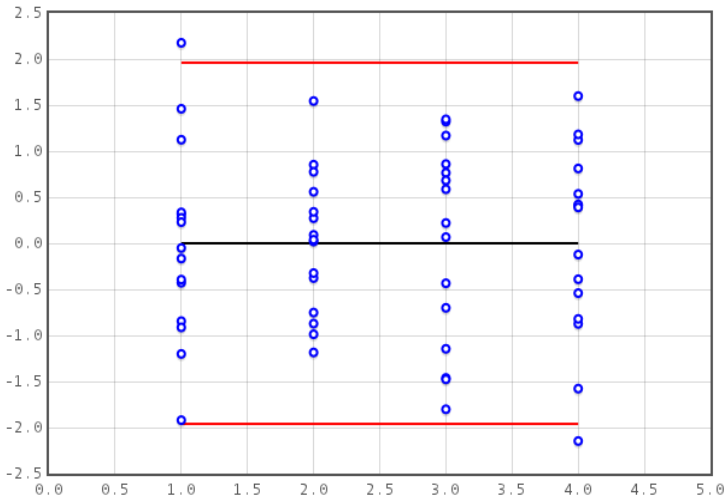

Figure 11. Example data showing residual mean error versus subject ID.

13. **Software package or computing processes used:** (Abu-Arafeh et al., 2016). The software used to compute plots and numerical calculations is not readily available and may lead to errors (Olofsen et al, 2014). Olofsen (2014) states that "the software available for calculating these descriptive statistics only implements the simpler methods, which are inappropriate for more complex data". 7/7 studies disclosed the software packages that were used to compute BA methods.

2.5 Individual Validation Study Review

In the study by Bent et al, (2020), multiple commercial and research grade WOHRM devices were tested across several movements (e.g. low-moderate physical activity) and sedentary tasks in a lab setting (Bent et al., 2020). This was the first study to investigate accuracy

in wearables across the full range of skin tones with an appropriate sample size. Error testing in heart rate across-skin tones and devices at rest showed no overall effect of skin tone. Significant differences were observed between WOHRM models and also between activity intensities. Mean absolute error calculations showed that error during activity was on average 30% higher than during rest. Amongst the commercial devices tested, the Apple Watch 4 performed more effectively during rest and physical activity as compared to the other devices (ME = -0.090 bpm (SD 2.8 bpm) & 0.79 bpm (SD 2.9 bpm), respectively. At rest, Fitbit Charge 2 showed relatively a nearly equivalent agreement of 0.34 bpm (SD 5.3). Surprisingly, the commercial wearables (4) performed better on average than research grade devices (2).

This study adds value to the literature along several dimensions. Within a single study, distinctions are drawn between high and low performing commercial and research grade devices. This has implications for safety, efficacy, and costs. In reference to the Bland-Altman 13 item checklist, the authors did not disclose visual or statistical tests of normality of the residual error. Moreover, BA LOA plots were noticeably absent from pooled and individual device comparisons (WOHRM vs. ECG). Lastly, 95% CI for LOA were not disclosed for pooled or individual devices which limits comparability between studies.

In another study by Dur et al, (2018), the Wavelet wristband was compared to ECG while healthy participants remained in a resting (seated) posture in a controlled-lab setting (Dur et al., 2018). HR, HRV, and spirometry data were collected throughout the study protocol. Agreement was established at 30, 45, and 60 second windows. BA LOA took into account multiple observations per participant. Skin-typing assessments were conducted but not featured in the error analysis. Results showed that PPG HR demonstrated good agreement but slightly overestimated ECG HR during resting conditions (ME = 0.3 (SD 1.1) bpm). Across the range of

recording windows, similar error outcomes were observed. The BA LOA plot provided ME and LOA and numerically reported these values (LOA = -2.6 -1.9 bpm). From visual inspections, there did not appear to be any evidence of systematic bias across the range of heart values. No rationale was given for the tasks chosen nor for the protocol length decided on.

Overall, this can be considered a valuable study demonstrating that in resting conditions the Wavelet wristband may be substituted for ECG for multiple sampling windows (30, 45, & 60 secs). But, neither a visual inspection nor statistical testing were performed to verify normality of residual error. Although the BA LOA plot and numerical error reports are apparent, 95% CI LOA were not. No rationales were provided for sample size nor protocol length.

In still another study by Hahnen et al, (2020), the Everlast TR10 was compared to ECG during quiet resting (seated) at a hospital preadmission testing center (Dur et al., 2018). Subjects were simultaneously screened for history of chronic disease. Self-report data indicated that the majority of patients had one or more cardiovascular complications. The investigators defined acceptable validity criteria a priori with HR measurements considered accurate if the MAE was within ±10% or ±5bpm. Results demonstrated the Everlast smartwatch did not meet the a priori criteria (MAE = 6.5 (SD 9.2) bpm and the MAPE = 9.9% (SD 14.3)). BA LOA plots were reported.

Overall, this study showed that Everlast smartwatch demonstrated low agreement and low precision (SD) during resting conditions. There were multiple study limitations: 1) Although subjects were screened for cardiovascular complications, it is not known what percentage with arrhythmias could have influenced the findings; 2) it was not disclosed whether the residual mean error met criteria for normality; 3) the authors did not report numerical calculations for ME, and the SD of ME, which were needed to calculate LOA and 95% CI LOA; 4) the size of

the plotted values in the BA LOA plot are such that systematic bias across the range of measurements is unclear. Lastly, by not providing the above calculations, it was unclear whether these results can be compared to similar studies.

In another study by Konstantinou et al. (2020), the Microsoft Band 2 was compared to ECG HR during rest and cold pressor (pain induction) conditions in a controlled-lab setting (Konstantinou et al., 2020). Healthy participants were exposed to a study protocol intended to elicit a stress response. Results showed good agreement in comparison with ECG for baseline resting (MAE = 3.78 bpm) and cold pressor conditions (MAE= 7.19 bpm). Paired-t-tests between baseline and experimental states showed significant increases in mean HR.

Overall, this study is one of two protocols to assess agreement during resting and stress reactivity with ECG as a comparator. However, there are notable study limitations. First, the authors did not report numerical values for ME (SD), LOA, or 95% CI LOA despite featuring some of these error values in the individual BA LOA plots. Second, they did not include pooled aggregate data contained within a single BA LOA plot. It is therefore challenging to fully comprehend findings. Third, there were no visual or statistical tests to demonstrate normal distribution for residual error. In summary, given the shortcomings of this study, we conclude that without sufficient BA methods, the results are unlikely to be comparable to related validation studies.

In another study by Menghini et al, (2019), the Empatica E4 was compared to ECG during sedentary behaviors within a controlled-lab setting with healthy subjects (Menghini et al., 2019). Skin-tone was objectively assessed but not featured in mean error calculations. Results demonstrated good agreement for each of the 7 sub-tasks (ME bias of -0.27-0.01 (LOA = -3.0 – 2.47) bpm) including during stress induction (Stroop task, speech preparation). BA LOA show

no evidence of systematic error throughout the range of measurements. Although the authors noted that residual error between devices was not normally distributed, they performed a logarithmic transformation that did not show improvement.

Overall, this study is one of two protocols to assess agreement during resting and stress reactivity in comparison with ECG. There was good agreement during rest and stress induction, including in elevated RHR values. The authors reported most BA items in detail for the reader to consider, however, the dataset was shown to not be normally distributed therefore the 95% CI LOA should be considered with caution.

In another study by Kroll et al, (2017), the Fitbit Charge was compared to ECG while participants remained in a resting (seated) posture in a hospital setting (Kroll et al., 2017). Participants were stable hospital inpatients in ICU. Repeated measures were taken over a 24 hour period. Results showed good agreement and poor precision (ME = -1.14 bpm, LOA = -31- 21 bpm). Performance was significantly better in patients in sinus rhythm than in those who were not (ME −0.99 bpm vs. −5.02 bpm, P=.02). Devices were recycled for use with other inpatients. The authors stated that they reused devices on average 9 times without seeing a decrement in performance. A histogram, density plots, and Wilcox signed ranked test indicated that the error distribution between devices was not significantly different.

Overall, this study is the only validation study to perform PPG HR testing in an ICU setting over a 24-hour period. Therefore, it is the most generalizable to the setting in which it was conducted. The total number of paired-samples included in the analysis is notable (N=12,358 or N=288 samples per inpatient). The authors add to the literature by providing results that indicate Fitbit Charge shows higher agreement for patients presenting sinus rhythms. The study could have been improved by plotting sinus rhythm vs. non-sinus rhythms and

providing the SD of the ME (in addition to LOA). The authors did not calculate 95% CI LOA. A challenge for the literature is to consider instances where too many paired-samples are collected such that the readability of the BA plot is compromised, as was the case in this study. Finally, although the patients were stable in the ICU, it is unclear from the procedures whether or not they remained in a true resting state throughout the 24 hour measurement period.

In another study by Schuurmans et al, (2020), the Empatica E4 was compared to ECG during rest (seated) within a controlled-lab setting with a clinical population of traumatized adolescents in residential care (Schuurmans et al., 2020). Devices were compared over multiple testing periods using a game-based meditation program. Normality was confirmed by Kolmogorov-Smirnov tests. Results showed good agreement between PPG HR and ECG (ME = -1.36 bpm, LOA = -2.47-5.18). Furthermore, the BA LOA showed no evidence of systematic error throughout the range of measurements.

Overall, this study adds to the literature by measuring adolescents during resting conditions. It is not clear in existing literature, how well PPG HR measures elevated RHR. While this study did include subjects with elevated RHR (mean = 88.44 (SD 12.51) bpm) the numerical calculations for the SD of the ME were omitted, as were 95% CI LOA.

2.6 Summary of Literature Review Findings

Given the previous review of 7 key studies, the reviewed literature appears inconsistent in reporting, designs and analyses. While this presents a limitation, given fewer studies to compare results with, it points to the need for the findings of this thesis. Wearable technologies have been seen as a potentially significant tool in health behaviour change (Sullivan & Lachman, 2017). It has been recently debated that they should be an integral part of large-scale efforts to increase the

sub-population who exercise healthfully for reduced disease risks (Hodkinson et al., 2021; Ritvo et al., 2021). While the focus of this thesis is validating sedentary HR estimates, existing literature is clear that multiple physical disease risks are elevated when RHRs are elevated (Aune et al., 2015, 2017). Furthermore, multiple psychiatric conditions involve elevated HR, either via short-duration spikes (e.g. Post Traumatic Stress Disorder, Panic Disorder) or via trends towards elevation associated with chronic social and general anxieties (Lader & Mathews, 1970; Taylor et al., 1982; Freedman, 1985; Taylor et al., 1986; Cameron et al., 1987; Gaffney, 1988; Meuret et al., 2011). For such individuals, valid observations of relative heart rate might offer another view of their experiences that could be incrementally modifiable, as self-observed in reductions of sedentary heart rate. While this is speculative, some instances of this type of assistance have been observed in past studies where Fitbits were a regular part of the provided intervention package (Ritvo et al., 2021). Such progress, however, depends on better validation studies demonstrating the precision of measurement in consumer wearables.

2.7 Hypothesis Testing

Null-hypothesis tests in concurrent validation studies involving BA methods are largely considered unsuitable, given the absence of theoretical or statistical predictions about specific populations. Nonetheless, the objective of this study is to test performance of the FBC2 in accurately estimating RHR when compared to (gold-standard) ECG while participants are seated and undergoing stress exposures. Testing is guided by statistical methods for assessing the BA LOA developed by Bland & Altman (Bland & Altman, 1986b, 1999, 2003, 2007) when the objective is testing two related devices using concurrent and continuous measurements. Satisfactory agreement is evaluated in the context of studies highlighted in the literature review

(Kroll et al., 2017; Dur et al., 2018; Menghini et al., 2019; Bent et al., 2020; Hahnen et al., 2020; Konstantinou et al., 2020; Schuurmans et al., 2020) and in ECG validation standards (*Physical Activity Monitoring for HR (ANSI/CTA-2065)*, 2018.). Similar levels of device performance for group level data are predicted as compared to outcomes demonstrated in the current literature for sedentary conditions. Based on two previous studies, we predict that agreement outcomes will present minor reductions of agreement with increasing intensity of heart rate volatility (HRVO) (Konstantinou et al., 2020; Menghini et al., 2019).

Ultimately agreement that approximates a true error score = 0 bpm could support device substitution when HR estimates under sedentary conditions are needed. The judgment of whether agreement outcomes are sufficient for applications is viewed as guided by statistical outcomes while dependent on intended applications.

3.0 METHODS

3.1 Study Design

Quantitative research methods were undertaken to perform an objective measurement of agreement within a controlled lab testing environment. The validity study protocol was designed to elicit a psychophysiological response to stress induced perturbations elicited through standardised procedures previously shown to induce cognitive-emotional stress.

Ethics statement

The protocol administered to study participants was reviewed and approved by the Human Participants Research Committee (HPRC) at York University (certificate - STU 2019-028) and follows the Declaration of Helsinki. Informed consent was obtained from all participants prior to study initiation.

Participants

Eligible study participants were screened and consented prior to conducting the experiment. Data collection was completed over a single study visit. Inclusion criteria: healthy undergraduate students from York University. Exclusion criteria: a self-reported history of (1) one or more cardiac related conditions, AND/OR, (2) a psychiatric disorder(s), OR (3) prior knowledge or familiarity of the deceptive stress task (Appendix B). Consenting participants were recruited throughout the spring and summer of 2019 using a secure online platform called 'Undergraduate Research Participation Pool' (URPP) maintained for research purposes by the York Psychology department (where students can earn up to 4% of their grade in PSYC 1010).

3.2 Sample Size

According to multiple sources, including the Bland Altman publications discussing method comparison studies and best practices for validating PPG HR, there appears no consensus for how to determine appropriate sample size (*Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065)*, n.d.; Gerke, 2020; Mühlen et al., 2021). For this reason, we compared the sample size data extracted from N= 7 relevant studies identified within the literature review to approximate a non-arbitrary sample size estimate. The results indicated a mean sample size of 46.3 participants (SD = 21.5) and a median (range) sample duration per participant of 15 (SD = 5-288) minutes. Therefore, we recruited N = 32 participants for a protocol of 35 minutes duration.

According to a recent systematic review published in 2020 intended to develop best-practice consumer wearable and smartphone step counter validation, if sufficient data are not available to define clinical maximum allowed differences regarding device agreement, a minimum of N = 15 participants and a maximum of N = 45 participants is advised after accounting for potential drop-out or data loss (Johnston et al., 2021).

Validation Protocol Design

We tested a device validation protocol informed by previous experimental protocols developed within York University (Azam et al., 2015, 2016) (Kirk-Change, 2021, in-press). Modifications were made to the delivery of the protocols where appropriate to ensure valid data collection and comparison between participants for group analysis. Table 2 provides a summary outlining the study protocol along with a list of the tasks performed.

Table 2
Validation Testing Protocol Outline

Phase	Testing Protocol (Task Names)	Duration (mins)
Pre-experiment	Self-Report (Demographics and Questionnaires); presentation (slide deck) for how the software program operates throughout the testing protocol	20
	Calibration Testing of ECG and FBC2	5
Time under observation (experiment)	(1) Resting Period 1	5
	(2) Viewing of Emotional Stimuli (IAPS)	10
	(3) Resting Period 2	5
	(4) Pattern Solving Task (PST)	5
	(5) Listening Task (LT)	10
Post-experiment	Self-report; disclosure of deception aspect of study	10
	Duration of Testing Period Total Study Duration	35 **70**

3.3 Lab-testing Procedures

Pre-experiment (25 minutes)

Eligible students were emailed prior to study visits. This correspondence asked them to refrain from the following behaviors prior to participating: (a) consuming or ingesting drugs/alcohol or caffeine for 6 hours prior to testing AND (b) engaging in moderate to vigorous physical activity for 2 hours prior to testing AND (c) consuming large meal within 1 hour prior to the testing period. A series of digitized self-report questionnaires and health inventories were completed (in-person) in the lab setting along with the consent obtained using desktop web application software (*Survey Monkey)*. Once participants had completed the self-report measures a short presentation was given by the researcher (SW) to verify that the participant understood the nature of the testing protocol and to explain how the software program would operate. Regarding postural considerations, a slide was shown that indicated a suggested posture to limit data loss with the ECG equipment (seated in a computer chair, palms facing the ceiling). Any

questions, concerns, and requests for clarity of instructions during the presentation were addressed by the researcher.

Next, the study equipment was fastened in place. This phase required multiple steps: 1) applying two ECG electrodes and ground flush to the skin; 2) fastening the respiration belt around the abdomen of the participant; 3) placing the ear headphones on the head of the participant to be worn throughout the duration of the testing protocol; 4) ECG HR, respiration, and FBC2 HR were verified using visual inspection of pilot input data.

The set up and calibration took approximately 5 minutes. For more details describing the equipment set up, see Appendix B.

Time under Observation (35 minutes)

***Resting Period 1*:** Study participants were seated in a chair. Participants observed a blank computer screen that was situated at about eye level for 5 minutes. An audio cue signalled though the headphones was provided to indicate an impending transition to a succeeding task.

***Viewing of Emotional Stimuli*:** Participants remained seated within a chair and were exposed to a standardized stimulus set containing full-colour pictures of actors/images extracted from the *International Affective Picture System* (IAPS) database to elicit a stress response (Lang et al., 1997). Participants were asked to attend to each image as it was presented while not focusing on any particular aspect but rather viewing the picture in its entirety with head and eyes oriented forward. A total of 75 images intended to elicit negative affect were shown on the computer screen (center-aligned) for a duration of 8 seconds per image with no wash out period. Image ordering and duration was consistent for all participants. Participants observed the negative

images for 10 minutes in duration. An audio cue signalled though the headphones indicated and impending transition to a next task.

***Resting Period 2*:** Study participants remained seated in chair. Participants observed a blank computer screen was positioned at eye level for 5 minutes. An audio cue signalled though the headphones was provided to indicate the impending transition to a succeeding task.

Pattern Solving Task: Study participants remained seated in chair with their right hand relocated from the resting position on the leg to the keyboard arrow keys on the right side of the keyboard with the forearm resting on the desk. Participants responded to the prompt by either pressing the left arrow key to indicate "FALSE" or the right arrow key to indicate "TRUE". This task was repeated for 35 trials within 5 minutes. The researcher was aware of the scoring (number correct vs. incorrect) and was situated to the left of the participant. An audio cue signalled though the headphones indicated an impending transition to a succeeding task. An example of the PST is shown in Figure 12.

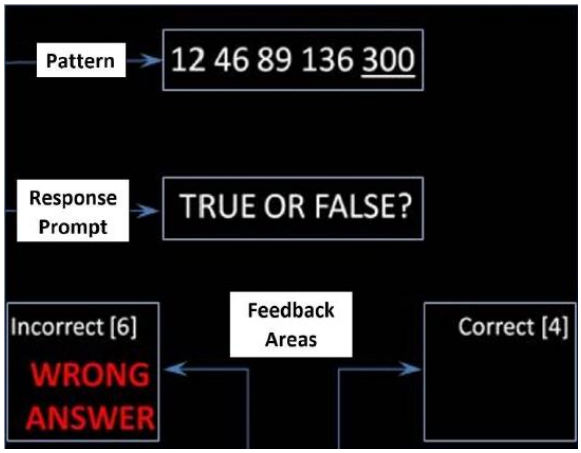

Figure 12. Example of Pattern Solving Task Interface

***Audio-description of mindfulness meditation*:** Following the end of Task 2, participants were exposed to a recorded mp3 file. The recording provided a description of mindfulness meditation instructions that explained the pedagogy of meditation. It was intended to be a description of rather than a guide to mindfulness. Participants were encouraged to listen attentively and to close their eyes if they saw fit. The end of this task coincided with the termination of data collection on all devices monitoring physiological biosignals (HR, respiration). Participants listened to the audio recording for 10 minutes. An audio cue signalled though the headphones was provided to indicate impending transition to a succeeding task.

Post-experiment (10 minutes)

As part of an exploratory analysis, participants were asked to complete a questionnaire after completing the testing procedure to determine the presence or absence of physiological changes during the listening task. (Note, we exclude these results from this survey in the data analysis.)

Subsequently a non-recorded semi-structured interview was conducted whereby the nature of deception task was disclosed stating that the outcome of the PST was 'fixed' with the intention to elicit an autonomic stress response. A disclosure document was signed by the participant stating that the participant would not disclose knowledge of the experiment outside the context of the study visit (see Appendix C). Equipment remained worn by the participant until the debriefing document was read and signed. Study equipment worn by participants was sanitized using disinfectant then wiped with a dry cloth. URPP credits were issued by email following completion of the study. The post-experiment period was about 10 minutes in duration

3.4 Equipment and Study Materials

Equipment:

1. *Fitbit Charge 2 (Google Inc.):* a commercial WOHRM device and fitness tracker used to collect measures of pulse rate through Fitbit's patented PurePulse$^{TM.}$ There are at least 45 manufacturers of wrist-worn devices. The FBC2 was selected due to its (a) widespread commercial adoption, (b) open-source Web API end-points, (c) 24-hr intraday HR, (d) extended battery life > 24 hrs, (e) Fitbit.Inc's long-standing investment in research-initiatives. The device was purchased and used according to manufacturer's recommendations. The original Fitbit Charge HR is featured in 476 studies in PubMed whereas the FBC2 was used in 61 PubMed Studies (Bent et al., 2020a).
2. *Electrocardiogram:* Continuous ECG recorded on a PC with ADInstruments (ADI) PowerLab 4-channel data acquisition system and LabChart V8 software through a 3-lead ECG.
3. *ADI Respirometry Equipment/Respiration belt:* was used to calibrate ECG and to observe HR patterns in relation to respiration (respiratory sinus arrhythmia).
4. *Wireless Headphones:* headphones provide noise-cancelling features to control for non-protocol related environmental stimuli. This device facilitated all audio related cues as part of transitioning between tasks and within tasks

Study Materials:

IAPS Database: 75 static (provocative) images were selected for low negative valence and issued at a ratio of picture: time of 1 picture every 8 seconds with no randomization of the image deck. The IAPS provides normative ratings of affect for a large set of images based on the Self-Assessment Manikin (SAM) 9-point rating scale. A score of 9 represents a high rating (i.e., high pleasure, high arousal), a score of 5 indicates a neutral rating, and a rating of 1 represents a low

pleasure, low arousal rating (Bradley & Lang, 1994). Overall, the 75-images had a moderately low pleasure rating (M = 3.05, SD = 1.63, Range = 1.79-4.30). It was decided that the content of the images which were intended to provoke negative emotions (e.g. fear, sadness, disgust, and anger) would be well tolerated for this study although less likely to impart uncomfortable (intense) levels of stress to participants (Lang et al., 1997).

Pattern Solving Task (PST): The PST is a deceptive task designed to elicit sympathetic activations or autonomic arousal in response to perceived error detection while mitigating individual differences in performance on basic laboratory stress tasks (Hajcak et al., 2004; Fechir et al., 2008). During this task, users were exposed to N = 35 trials with the researcher observing results within their line of vision but to one side. For each trial participants were allocated a total of eight seconds to anticipate a pattern amongst the 4 alphanumeric characters presented at the top of the screen, at which point to a 5th character displayed and the words "True" or "False" presented in the middle of the screen, prompting users to discern whether or not the 5^{th} set of charters fit the pattern. Next, users indicated a "True" response if they thought that the 5th character followed the "correct" pattern, or "False" if that was not the case, by actioning a corresponding keyboard press. Regardless of the key being pressed (True/False), the user's response was not scored until after 8-seconds had elapsed after the 5^{th} set of characters was displayed. The results for a given trial remained on the screen for 4 seconds followed by a subsequent trial. In addition, participants were made aware of the cumulative number of correct and incorrect responses remaining on the screen (corresponding to the location of the "Right Answer" and "Wrong Answer") throughout the duration of the task (see Appendix H).

This version of the PST is designed to create the false perception and the need for users to use the *right* hand to interact with a keyboard (off-hand to FBC2 placement) and to respond to screen-prompts, accordingly. Feedback for the task was configured using a predetermined random order (as standardized for all participants) to provide an equal number of "Right" and "Wrong" answers. Therefore, the best score possible was a result of 50%, shy of the 80% expectation stated in explicit terms prior to the experiment.

The alphanumeric characters were tailored to appear as if they follow simple or complex patterns, but ultimately all were 'pseudo' patterns created without a correct solution. In addition, users were also led to believe that the action of pressing a key stroke on the keyboard would be registered by the software and conveyed accordingly, when in fact, that was not the case. To this end, for all intents and purposes the *entire protocol was a video file mp4.* intended to create the perception of requiring users to engage when prompted to do so. Written documentation (e.g. consent form), tutorial, and verbal reiterating of task goals and expectations, was designed to help facilitate the deceptive nature of the task in convincing fashion.

Pre-recorded audio of mindfulness meditation description: The guided meditation followed the same script as previously validated (Azam, 2015; Azam et al., 2016b).

Beck Anxiety Inventory

Anxiety symptoms were assessed using the Beck Anxiety Inventory (BAI). The BAI is a 21-item self-report inventory to assess symptoms of anxiety (e.g., "feeling hot," "difficulty breathing," "nervousness") within the last two weeks. Respondents rate how much each of the 21 symptoms bothered them in the past month on a four-point ordinal Likert scale from a score of 0 = not at all, to 3 = severely. Scoring for the BAI ranges from a minimum of 0 to a maximum of

63 based on summing the scores for all items. The following cut-offs are empirically supported for the interpretation of scores: 0-9 = normal/no anxiety, 10-18 = mild to moderate anxiety, 19-29 = moderate to severe anxiety, 30-63 = severe anxiety. The BAI has wide application for both clinical and non-clinical populations and has demonstrated strong test-retest reliability over one week (.75) and internal consistency (Cronbach's alpha = .94). Content, construct and convergent validity is established (Steer & Beck, 1997; Sprinkle et al., 2002). Appendix D.

Perfectionism Cognitions Inventory

The Perfectionism Cognitions Inventory (PCI) measures how often an individual engages in cognitions focused on the discrepancies between one's current self and ideal self. Correlational analyses found the PCI to be associated with psychological distress characterized by lack of self-reinforcement and positive perception of self (Flett et al., 2007). Respondents indicate how often they experienced perfectionistic thoughts (e.g. "I should never make the same mistake twice", and "I must be efficient at all times") in the previous week on a 5-point Likert scale. The PCI was recently used in studies on the relationship between perfectionism and distress, and an assessment of web-based CBT for maladaptive perfectionism (C. Arpin-Cribbie et al., 2012; C. A. Arpin-Cribbie et al., 2008). Lastly, the PCI is shown to demonstrate adequate validity and reliability. See Appendix D.

Five Facet Mindfulness Questionnaire (FFMQ):

To measure daily mindfulness the 39-item Five Facet Mindfulness Questionnaire (FFMQ) was administered to study participants (Baer et al., 2004). The FFMQ is established as a valid and reliable measure among undergraduate students and is commonly used to assess

changes in mindfulness tendencies (e.g., non-judging of inner experience) before and after a mindfulness-based intervention. Analysis of psychometric properties of the FFMQ demonstrates validity and reliability within satisfactory ranges (Gu et al., 2016). Five mindfulness-based factors are evaluated including: non-reactivity, observing, describing, non-judgment, and acting with awareness. The FFMQ is scored by summing all item responses and total scores ranging from 0 to 195, with higher scores indicating greater levels of mindfulness. It is unclear whether the semantics of the items' will be clear for those individuals without meditation experience (Grossman, 2008; Grossman et al., 2010). See Appendix D.

Patient Health Questionnaire-9:

The PHQ-9 is a multi-purpose instrument for screening, diagnosing, and monitoring the severity of depression and can be administered repeatedly, reflecting improving or worsening of symptoms in response to treatment. The PHQ-9 has been well validated and proven reliable when compared to the original PRIME-MD and is more efficient to use in practice (Spitzer, 1999). As a severity measure, the PHQ-9 score can range from 0 to 27. Each of the 9 items is scored within the range of 0 (not at all) to 3 (nearly every day) (Kroenke et al., 2001). See Appendix D for further background data.

Automatic Thoughts Questionnaire (ATQ)

The ATQ was developed to evaluate four dimensions of personal automatic negative statements: (1) personal maladjustment and desire for change (2) negative self-concepts and negative expectations, (3) low self-esteem, (4) helplessness. The ATQ consists of 30 items with scores ranging from 30 to 150. Respondents are asked to self-assess the frequency of the thought

on a 5-point likert scale. The ATQ is demonstrated to reliably distinguish depressed from non-depressed groups, with a Cronbach's alpha of 0.96 (Hollon & Kendall, 1980). See Appendix D.

Fitzpatrick Skin Type Scale (photo-typing) (FSTS)

The primary application for the FSTS is to create a classification system for dermatologists to assign patients to in order to determine a person's risk to ultraviolet radiation exposure and therefore the potential for resulting irritations, burns, and hyperpigmentation. The FSTS is a self-report measure used to subjectively categorize study participants into a numerical skin-typing classification system (I-VI). Scores are used to estimate the characteristics (i.e. melanin) of the skin's ability to alter properties of ultraviolet (light) waves directed at the skin's surface (Sachdeva, 2009).The FSTS self-report measure consists of a two classification structures which include: (1) skin complexion and (2) the degree of sun exposure relating to the amount of melanin pigment in the skin composed of three subscales; (a) Genetic Disposition, (c) Sun exposure, (c) Tanning Habits (Fitzpatrick, 1988). According to this system there are six skin types classified from *I to VI*, which is essentially a numerical classification schema for human skin color. Responses are tallied and summating to total score ranging between 0 and 36. Higher scores correlate with increased levels of skin pigmentation and consequently a reduction developing skin cancer from sun damage (Sachdeva, 2009). See Appendix D

3.5 Data Collection

ECG-derived HR: Recordings of ECG data were taken using ADInstruments (ADI) PowerLab 4-channel data acquisition system. LabChart Pro software by ADI is integrated with the ECG unit and was used to obtain computer digitized ECG signal from the PowerLab unit at a sampling

rate of 1000 Hz. ADI spirometry equipment was used to observe respiratory parameters. LabChart Pro software by ADI is integrated with the ECG unit which enabled the research team to compute RHR. ECG data was collected and prepared in accordance with the standards set by the Task Force of the European Society of Cardiology (TFESC) the North American Society of Pacing Electrophysiology. LabChart Pro uses automated ECG interpretation which indicates ectopic and normal beats. Data were visually inspected to ensure each R-R heartbeat interval had clear PQRST formation. The "beat classifier view" feature in LabChart 8 Pro was utilized to inspect ectopic beats. R-R frequency parameters were transformed to 1-minute epochs, from 0-59 seconds (bpm) in LabChartPro then exported as an .xls file in Excel.

Wearable-derived HR: A dedicated Fitbit study account was created by the research team on the study computer to store raw HR data from a single FBC2 device. We used the Fitbit Developer API to create a custom web application to enable downloading of intraday HR data at minute-level granularity from a single FBC2 study device, using an open-source software (Python). R-R intervals were computed internally using Python from Fitbit Web API and then exported as .xls file in Excel. Each time the script for the computer program was run, individual HR values were extracted using date and time endpoints to select for the appropriate test date (e.g. 11:00:00-11:35:00AM). We provide the calendar dates range and firmware version for data collected between April 28 - July 05, 2019 (22.55.2).

For an outline of the steps taken to collect and analyze data for ECG and FBC2 see the diagram flow in Figure 13.

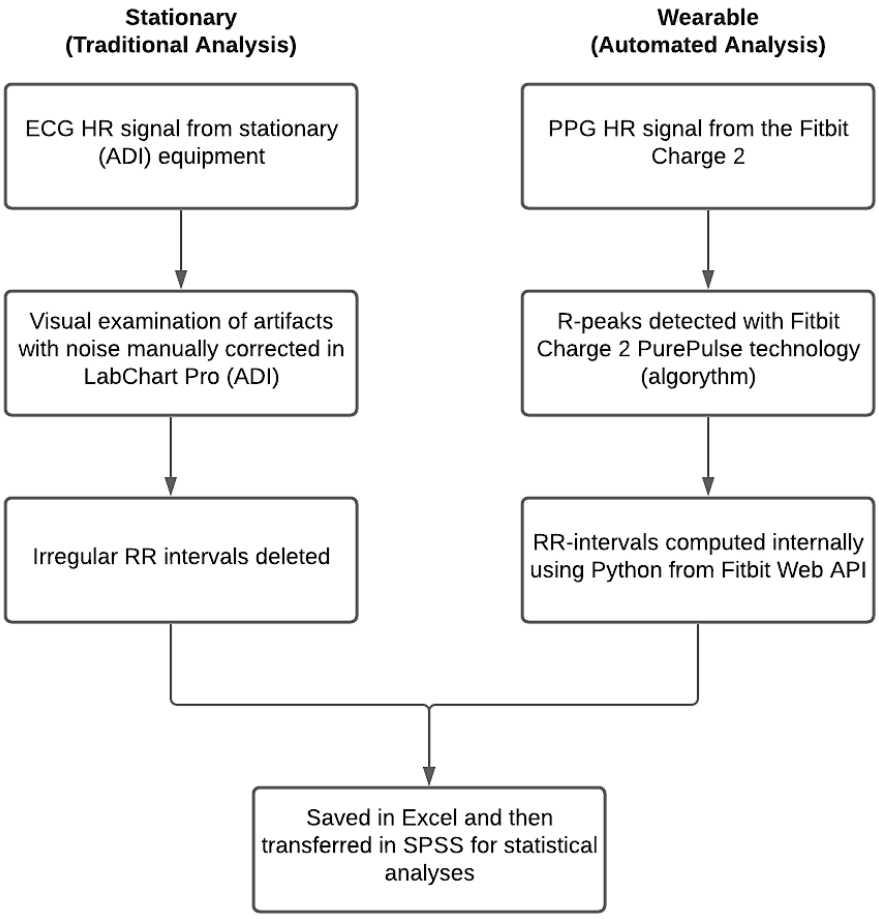

Figure 13. Procedure followed by each method for analyzing the ECG and FBC2 HR data.

3.6 Data Analyses

3.61 Device Agreement

We analyzed HR data in aggregate at the group level across all participants as well as the between- and within-subject variance. A simple data structure consisted of paired-values sampled concurrently and continuously consisting of four columns (Subject ID, Time, Method A, Method B) by 35 rows with each row representing HR averaged to bpm. As additional data

was collected, we manually appended those values to the initial dataset resulting in a 4 (columns) x 1120 (rows) matrix. From this initial structure error calculations were transformed into new variables (ME, ME%, MAE, MAPE%) with SPSS Statistical software.

Pooled ME values, precision of error (standard deviation and limits of agreement (LOA)), and correlation analysis were computed from the residual error values between ECG HR and FBC2 HR. We also performed the above calculations for each of five subtasks (Resting Period 1, IAPS, Resting Period 2, PST, & LT).

Congruent with the Bland-Altman 13-item checklist, we visually inspected the residual ME to determine whether an approximately normal distribution was apparent by using SPSS to plot a histogram and Q-Q plot.

SPSS was also used to compute a BA LOA Plot to visually inspect for systematic error (e.g. evidence of fixed or proportional error bias). Python (Pandas.Dataframe package) was used to plot a Box and Whisker plot to observe relative error between study tasks and to inspect for extreme scores (outliers).

As previously mentioned, this dataset contained multiple measurements per participant. To determine the between- and within-subject variance we used a free custom web application referenced in and discussed in a Bland-Altman publication (Bland & Altman, 2007) to calculate and plot residual ME versus subject ID (https://sec.lumc.nl/method_agreement_analysis/). De-identified data were uploaded to the web application run on JavaScript programming language via Google Chrome and did not require a user account or fees to perform analysis.

Descriptive statistics from the demographics, psychometrics, and related self-report survey were exported from Survey Monkey. The data was aggregated in Excel before performing

calculations in SPSS. SPSS was used to compute mean and standard deviation as well as cut-off values where appropriate.

3.62 Exploratory Analysis: Heart Rate Volatility (HRVO) & Device Agreement

For this exploratory analysis we expand on the dataset by using SPSS to transform FBC2 HR into a new variable: HRVO. First, a histogram was plotted to show the distribution of aggregate changes in individuals' HR on a minute-to-minute basis throughout the study protocol [$x^{\Delta} = x^2 - x^1$]. Second, to determine the influence of the HRVO intensity on agreement we tranformed new categorical variable called 'HRVO Intensity Intervals' consisting of six separate categories (+0-1 bpm^1, +0-2bpm^2… >5bpm^5). With the new variable created, residual MAE was compared by HRVO intensity to produce a table that indicated MAE (SD) and frequency count.

If assumptions of ANOVA were met we computed a one-way ANOVA (analysis of variance) between six HRVO intervals. The Tukey Test was a post hoc test used to perform a pairwise comparison of the means to see whether a significant difference was found between intervals.

In addition, we plot two line graphs as time series (ECG HR-FBC2 HR) consisting of the pooled mean heart rate values per minute over time to observe relationship HRVO between measurement devices.

Statistical Criteria for Validating WOHRM Devices

The following benchmark criteria were adopted for discerning agreement from group analysis applied to the group dataset:1) **Mean Absolute Error** (MAE) ≤ 5 bpm or **Mean Absolute Error** (MAE) or MAPE% of ≤ 10% (Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065), 2018.)**,** (2) Pearson R correlation ≥0.90, (3) BA LOA Plot was visually

inspected for indication of systematic error across the HR value range (Karras, 1997a, 1997b).

Statistical Techniques and Formulae for Evaluating Concurrent Validity (Agreement)

Statistical analysis and plots were performed using multiple software tools: IBM SPSS (Version: 24), Microsoft Excel, and Python (*Jupyter Notebook: version 3.65*). Below are the error calculations performed for the analysis along with the formula.

Error Calculations	Formula
Mean (Directional) Error	
The ME was calculated as the difference between FBC2 measurement and the gold standard ECG	*ME = (FBC2 HR – ECG HR)/ # of samples*
Limited of Agreement (LOA)	
LOA between the two methods are defined by a 95% prediction interval of a particular value of the difference.	*LOA = ME±1.96SD*
Mean Absolute Error (MAE)	
The MAE was calculated as the average absolute difference between the device and the gold standard.	*MAE = ABS(FBC2 HR – ECG HR)/ # of samples*
Mean Absolute Percentage Error (MAPE)	
The MAPE relative to the ECG was calculated by averaging the individual absolute percent errors.	*MAPE = ABS(FBC2 HR – ECG HR *100)/ FBC2)/ # of samples*
Bland Altman LOA Plots	
Difference plots, Calculate 95% LOA for which the spread of ME between device and reference values are plotted.	*x-coordinates(mean average (bpm)), y-coordinates(mean residual error(FBC2-ECG))*

Benefits/Risks to Study Participants

Benefits/compensation: no substantial benefits were associated with participating in this study. As described in the study consent form anticipated that this investigation would lead to a better understanding of how WORHM could be incorporated for various clinical applications involving event-based monitoring and subsequent intervention (See Appendix B). URPP were issued to study participants at the time of consent which took place during the study visit.***Risks:*** Both ECG and Fitbit PurePulse^{TM} technology are shown to be safe, non-invasive health monitoring equipment that present minimal risks to study participants. According to Fitbits official website, "Allergy to visible light are extremely rare and the PurePulse^{TM} LED lights are on the visible spectrum, similar to the lights in your home or office. The LED lights produce very low power so they won't burn your skin, and they're programmed to shut down if your device freezes or can't find a signal." Moreover, participants received a maximum exposure of less than one hour. In addition, few consumers of wearable devices report irritation or contact dermatitis as a result of the silicon band pressing against the skin. These symptoms are more readily apparent and exacerbated with repeated use combined with activities resulting in increased sweat production during movement, none of which applies to the current experiment.

4.0 RESULTS

Study summary

A group of 32 participants successfully completed the entire study protocol (4 males, 28 females; ages 21.6 (SD 2.57) yrs.) and were an ethnically diverse student group. Self-reported psychometrics indicated elevated levels of anxiety and depression. Recruitment took place over the course of an 8-week period between April 11/2019 and June 7/2019. Table 3 shows sample characteristics as descriptive statistics. This protocol was designed to elicit cognitive-emotional stress. Five sub-tasks were assessed: 1) Resting Period 1 (5 mins), 2) viewing negative images (10 mins), 3) Resting Period 2 (5 mins), PST (5 mins), and a listening task (10 mins).

The aggregate dataset that was included for group analysis resulted in 1120 minutes of data collection (32 participants x 35-minute testing period). All paired-values of HR estimations observed between devices (FBC2 HR vs. ECG HR) were matched and included in analyses. No missing data or unmatched paired-values were observed prior to computing mean averages for each 1-minute bin. A single wearable device was re-used for each participant. The battery of the FBC2 was kept at or near-full charge per participant. Overall, the study protocol was well tolerated with no attrition during testing. There were minor concerns re: the possible risks associated with measuring biosignals (e.g. ECG conductivity), and there was observed some occasional sleepiness or drowsiness during the Resting Periods 1 & 2 and the Listening Task.

Table 3 *Study Sample Characteristics*	**Value (N=32)**
Age (yrs.)	21.6 (SD 2.57)
Gender	

Male	4
Female	28
Ethnic Backgrounds	
Caucasian	3 (9.38 %)
South Asian	10 (31.25 %)
Chinese	5 (15.63 %)
Black	6 (18.75 %)
Aboriginal	1 (3.13 %)
Other	7 (21.88%)
Fitzpatrick Skin Type Scale	
Level 1-2	0
Level 3	2 (6.3%)
Level 4	25 (78.1%)
Level 5	5 (15.6%)
Level 6	0
Number of valid 1-min epochs	
Within-lab protocol	1120

Psychometric Assessment	**Mean (SD)**	**Score Range/Cut-off (N)**
Beck Anxiety Inventory	13.94 (11.97)	
Normal/no anxiety		0-9 (13)
Mild to moderate anxiety		10-18 (10)
Moderate to severe anxiety		19-29 (7)
Severe anxiety		30-63 (2)
Perfectionism Cognitions Inventory	46.2 (21.46)	
Average Item Score		1.84/4

Five Facet Mindfulness Questionnaire	105.6 (21)	
Average Item Score		2.7/5
Patient Health Questionnaire	9.7 (6.7)	
Minimal depression		1-4 (10)
Mild depression		5-9 (8)
Moderate depression		10-14 (7)
Moderately severe depression		15-19 (6)
Severe depression		20-27 (1)
Center for Epidemiologic Studies Depression Scale	18.6 (7.9)	
Low risk		<16 (12)
At risk for clinical depression		>16 (20)

Table 3: Study sample characteristics included the following variables: 1) Age, 2) Gender, 3) Ethnicity, and 4) FST-Scale. The total number of matched pairs values is equal to 32 participants sampled x 35 mins/ participant = 1120 paired records. All of the above values are presented in the following format: means (SD) for numeric calculations and frequency, % for categorical variables.

Device Error Estimates

We visually inspected two different plots to verify whether or not the data was approximately normally distributed. The histogram below in Figure 14 illustrates the distribution of the residual mean differences between both methods. The plot illustrates the frequency distribution of ME scores expressed in bpm. A slight left skew is observed indicating the FBC2 HR underestimated ECG HR in minor proportions.

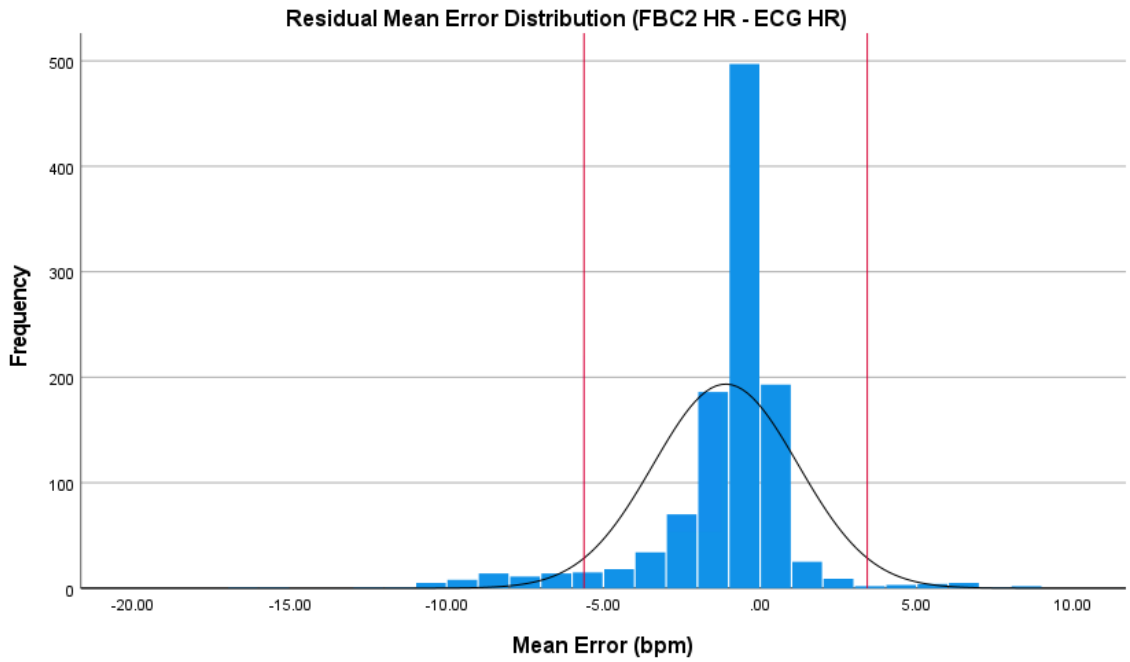

Figure 14. A histogram illustrates the distribution of ME between methods (FBC2-ECG). The red line represents perfect agreement between devices Two vertical blue lines are used to indicate the proportion of HR observations contained within the 95% LOA. The black (bell-shaped) curve is used to compare a normal sampling distribution.

As shown in Figure 15 we performed a quantile-quantile (Q-Q) plot to interpret whether residual ME was approximately normally distributed. Upon visual inspection, the points appear not to be positioned on the 45-degree reference line that suggests possible non-normal distribution. As a result, two statistical tests of normality were run to verify whether residual mean error was normally distributed. Results showed that both the Kolmogorov-Smirnov and Shapiro-Wilk test p-value were statistically significant ($p = 0.00$) for a large sample size ($N=1120$). Therefore, we rejected the null hypothesis that the residual error between devices was normally distributed.

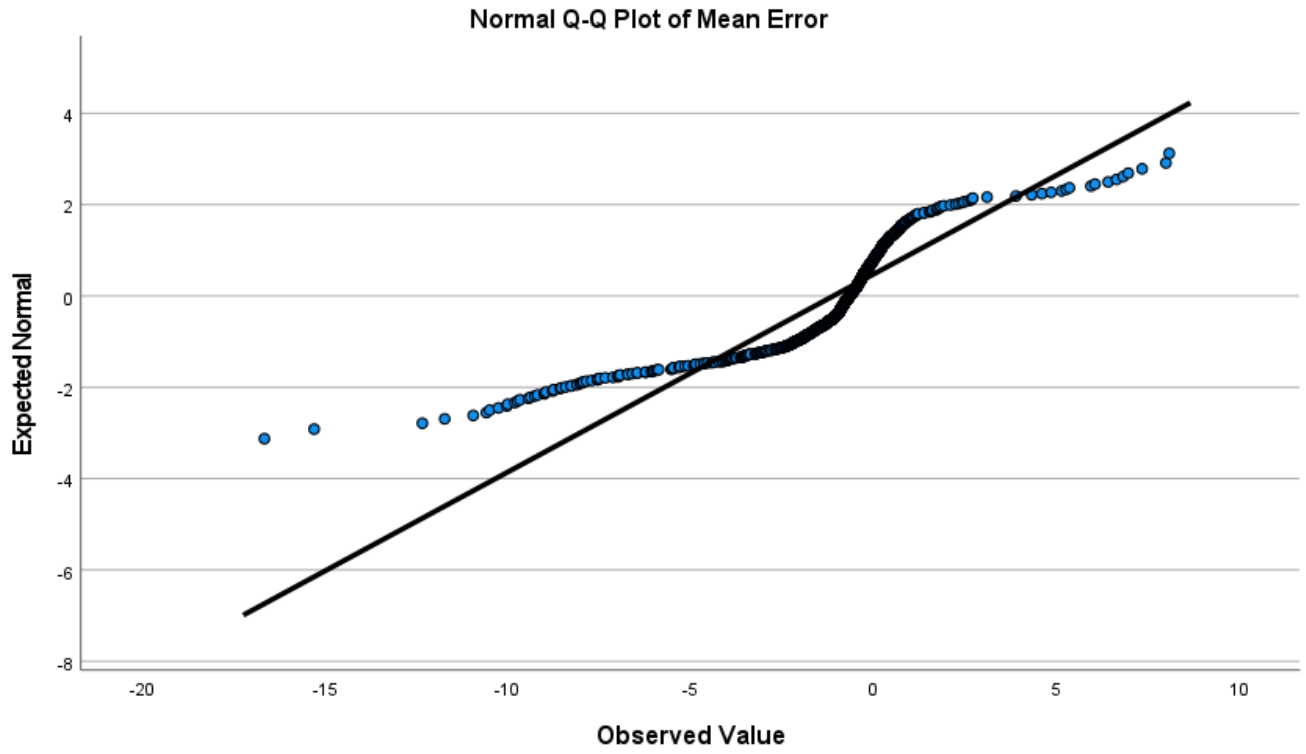

Figure 15. Quantile-Quantile (Q-Q) Plot. Blue points indicate individual paired-values. Darker shaded points indicate overlap where the majority of error values fall. A normal distribution is indicated by points that approximate or fall on the 45-degree reference line (black diagonal line).

Device Error Summary

Table 4
Benchmark Criteria

Statistical Measure	Benchmark Applied	Study Results
Pearson r correlation	≥ 0.90	0.97
Mean Absolute Error	≤ 5 bpm	1.5 bpm
Mean Absolute Percentage Error	≤ 10 %	1.34%
Visual Inspection of BA LOA Plot	No evidence of systematic error	No evidence of systematic error

Table 4. Proposed statistical benchmarks for agreement between FBC2 HR and ECG HR. The values contained within the third column reflect exclusively, the overall aggregate findings for the 35-minute protocol. In addition to mathematical manipulation, qualitative value judgment as applied to BA plots is used to determine the extent to which agreement and bias are present.

Two related methods of estimating HR were tested for concurrent validity throughout the completion of five consecutive tasks. Paired mean samples aggregated to 1-min; both good agreement, precision (standard deviation), and strong positive association were established between devices for the group analysis; ME = -1.1 (SD 2.3) bpm, MPE = 1.01 (SD 0.04) %, MAE =1.5 (SD 2.07) bpm, MAPE% = 1.34 (SD 1.85) %, P(r)=.97. See Table 4 to compare the results achieved in present study versus proposed benchmark criteria. Results demonstrate a relatively small mean error variance for each of the sub-tasks with the highest and lowest mean error reported during the Listening Task (ME = 1.07 (SD 2.42) bpm), IAPS (ME = 1.17 (SD 2.2) bpm) tasks, respectively. Table 5 shows a further specification of the above results by study task, using multiple statistical parameters.

Table 5

Error Bias Reporting by Study Task

Study Tasks		**Values**
Resting Period 1	Duration	5 minutes
	Pairs, N	160
	FBC2 HR, Mean (SD)	73.27(9.98)
	ECG HR, Mean (SD)	74.27(9.63)
	Mean Error (SD)	- 0.99(2.39)
	% Mean Error (SD)	-0.89(2.1)

	% Mean Absolute Error (SD)	1.37(1.86)
	Pearson coefficient correlation (r)	0.97
IAPS Images Task	Duration	5 minutes
	Pairs, N	320
	FBC2 HR, Mean (SD)	72.83(9.98)
	ECG HR, Mean (SD)	74(10.19)
	Mean Error (SD)	- 1.17(2.2)
	% Mean Error (SD)	-1.05(1.96)
	% Mean Absolute Error (SD)	1.25(1.84)
	Pairs, N	.98
Resting Period 2	Duration	5 minutes
	Pairs, N	160
	FBC2 HR, Mean (SD)	73.44(9.79)
	ECG HR, Mean (SD)	74.81(9.83)
	Mean Error (SD)	-1.37(2.16)
	% Mean Error (SD)	1.22(1.93)
	% Mean Absolute Error (SD)	1.36(1.84)
	Pairs, N	.98
Pattern Solving Task	Duration	5 minutes
	Pairs, N	160
	FBC2 HR, Mean (SD)	73.93(9.84)
	ECG HR, Mean (SD)	74.85(9.75)
	Mean Error (SD)	-.92(2.33)

	% Mean Error (SD)	-0.8(.2)
	% Mean Absolute Error (SD)	1.31(1.81)
	Pairs, N	.972
Listening Task	Duration	10 minutes
	Pairs, N	320
	FBC2 HR, Mean (SD)	73.18(9.89)
	ECG HR, Mean (SD)	74.24(9.46)
	Mean Error (SD)	-1.07(2.42)
	% Mean Error (SD)	-.95(2.2)
	% Mean Absolute Error (SD)	1.46(1.85)
	Pairs, N	.97

ECG and FBC2 HR outcomes, Bland-Altman Analysis and Limits of Agreement

BA methods were used to determine agreement between devices (i.e. the distribution of mean bias within 95% limits of agreement (LOA)). No evidence was found for a systematic error through the range of HR measures (range= 50-111 bpm), i.e. neither fixed nor proportional biases were evident. On average FBC2 HR slightly underestimated ECG HR (ME = -1.1(2.3) bpm. A total of 1044/1120 (93.2%) data points fell within the pre-defined LOA (-calculated [mean difference $\mp$ 1.96 X SD of the differences] OR LOA = [-5.62, 3.42] bpm; 95% CI LOA (lower LOA = [-6.65, -4.94] bpm, upper LOA = [2.74, 4.45] bpm). LOA containing 95% of paired-sampled observations are shown in Figures 16.

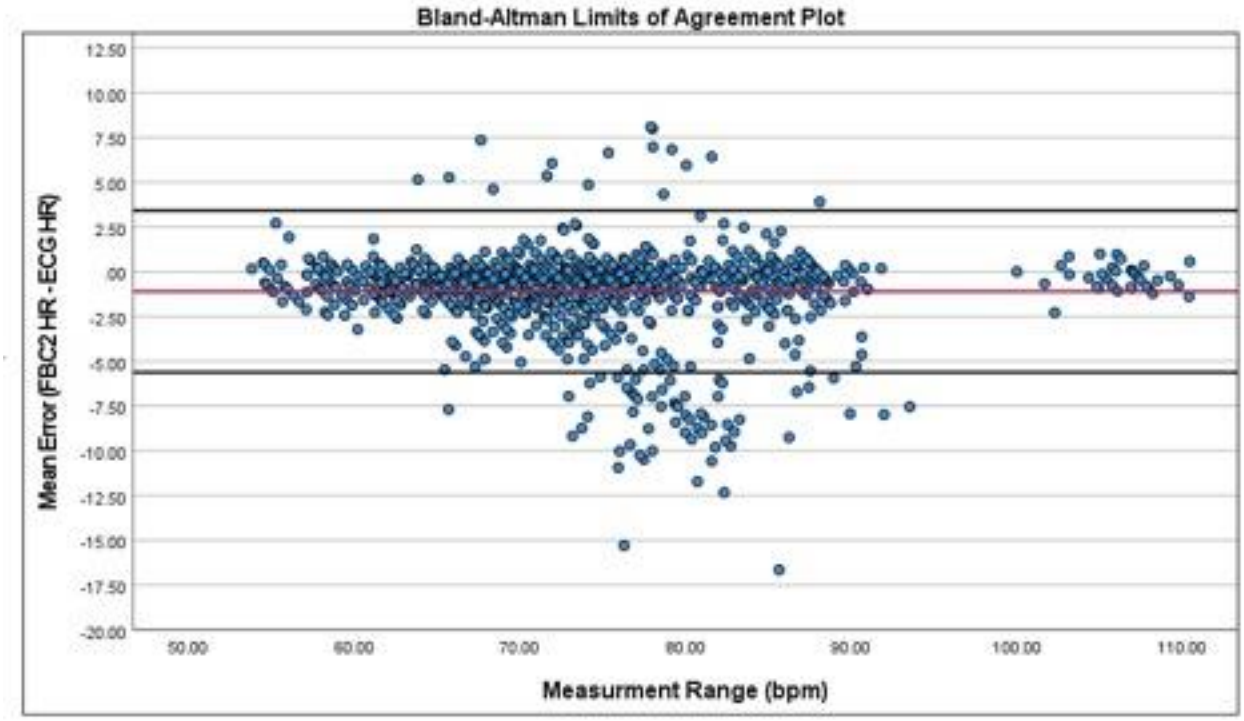

Figure 16. BA LOA Plot of differences against averages. Mean HR difference bpm are shown on the x-axis; red and black (horizontal) lines indicate mean difference (bias) between the measurement for criterion vs. reference standard and limits of agreement (LOA). ME = 0 on the y-axis reflects the theoretical line of equality between methods (FBC2 HR vs. ECG HR).

Percent Error

Across the 35-minute lab-based study protocol, the FBC2 produced a MAPE of 1.34%. During **Resting Periods 1 & 2**, the FBC2 had a MAPE of 1.37% and 1.36%, respectively. During the **Image Viewing (IAPS) task**, the FBC2 had a MAPE of 1.25%. During the **PST** (cognitive-stress task), the FBC2 had a MAPE of 1.31%. During the **Listening Task,** the FBC2 had a MAPE of 1.46%. (see Table 5 for percent error statistics and Figure 17 for ME, % by device across activities).

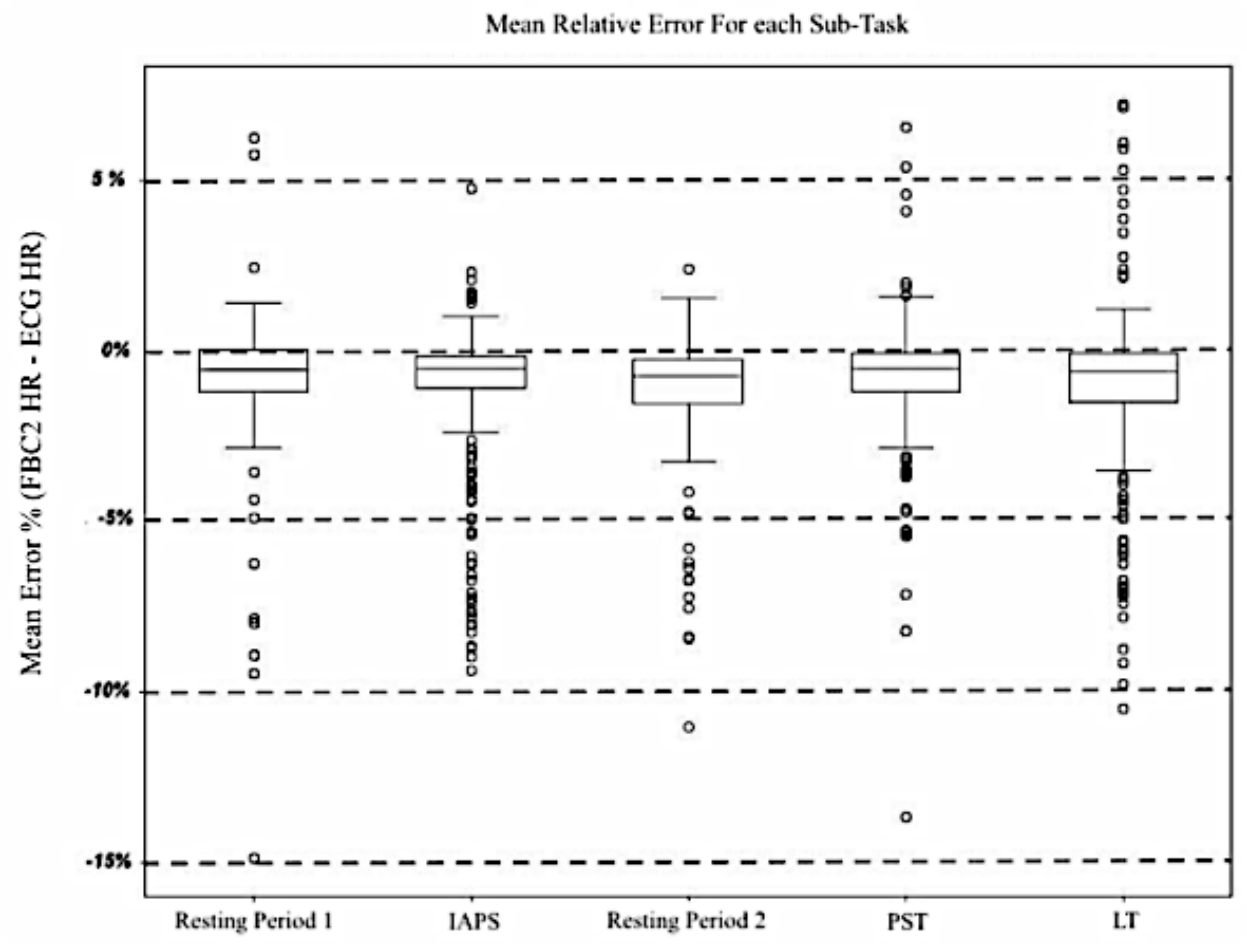

Figure 17. Percent relative mean error for pooled-group analysis in Heart Rate across tasks. Dashed lines along the y-axis reflect error threshold and its multiples for device agreement (5, 10, 15%). For each task - median error value, interquartile

To consider the within- and between-subject variance regarding agreement, we plotted the residual error versus individual study ID across the study protocol (N=32 participants, number of paired-values = 35). Within-subject variance was equal to 2.79 (SE 0.12) bpm, whereas the between-subject variance was equal to 2.62 (SE 0.69) bpm). See Figure 18 for a

graphical representation of within- and between- subject variance.

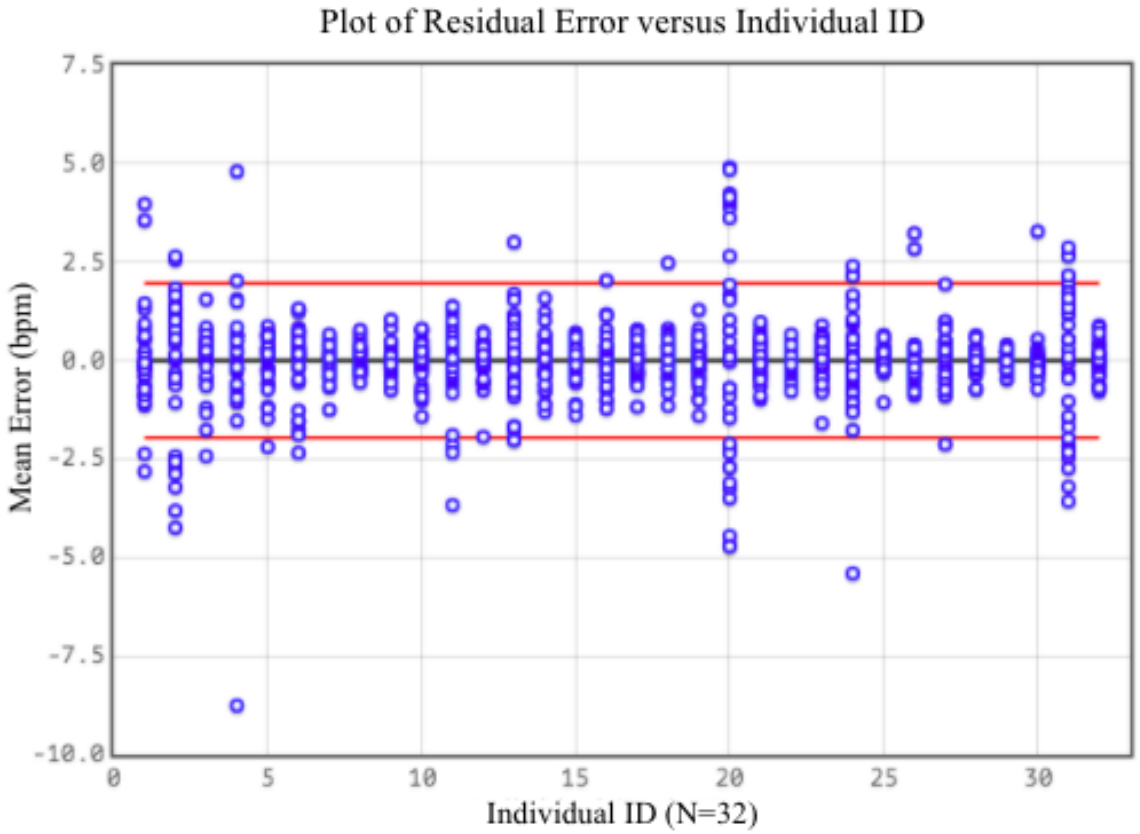

Figure 18. Scatter plot of residual error versus Individual ID. Blue circles = individual mean error scores between paired values (FBC2 HR vs. ECG HR). Black line = true error score. Red lines = variance.

Stress Reactivity Paradigm: HR Reactivity and Agreement

To illustrate the relationship between FBC2 HR and ECG HR under the presence and absence of stress reactivity, we plotted two line graphs for each measurement device along a time series. Each individual point indicates grand mean (or pooled mean) derived of individual HR values at a given time point throughout the duration of the 35-minute validation protocol. For a graphical representation of this relationship see Figure 19.

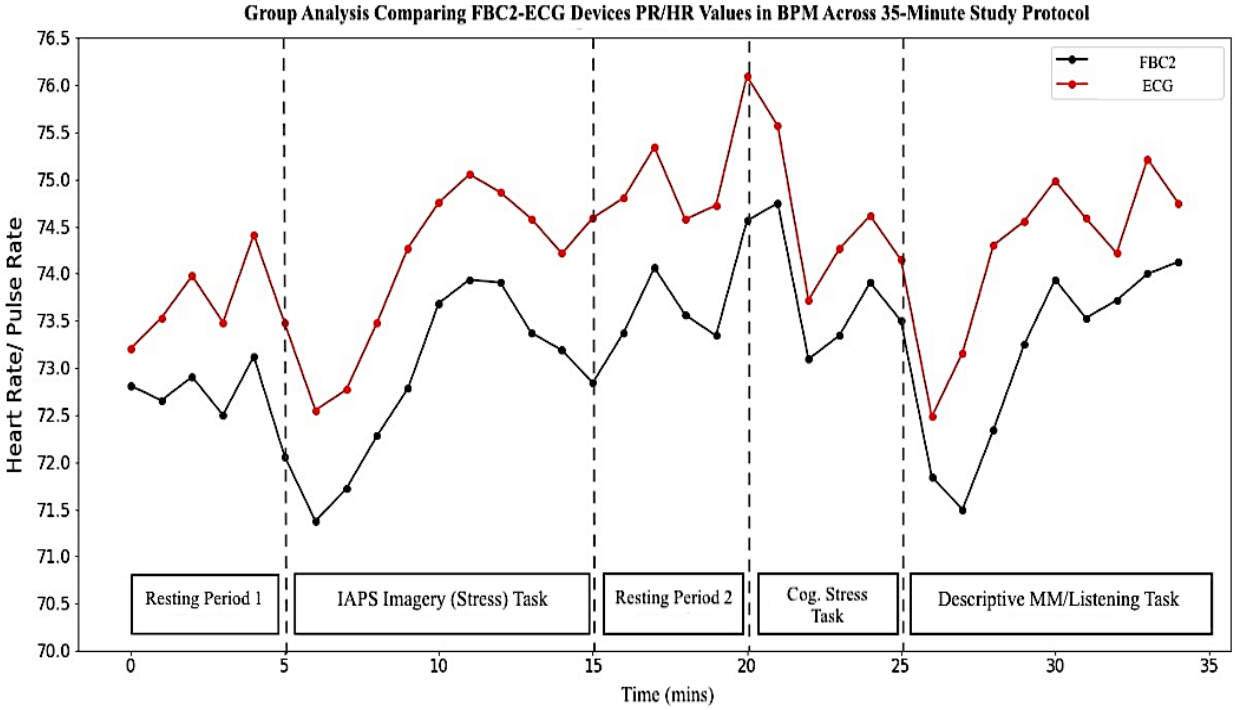

Figure 19. Time-Series Plot of Group Mean changes in HR across the study protocol. Time represented in mins (x-axis) vs. mean HR bpm (y-axis) for both FB Charge 2 and ECG devices across 5 sub-tasks for all study participants (N=32). Vertically-oriented dotted lines reflect transitions to a subsequent task.

Altogether, the influence of heart rate volatility (HROV) in relation to agreement was assessed during sedentary conditions in a controlled lab environment. Minute level HR observations were extracted from the FBC2 dataset and compared to a subsequent observation to calculate a net value (bpm). A total of 34 of 1120 records were removed from the dataset to account for measures sampled at the 0- and 35-minute mark as there were no matched paired values [$x^{\Delta} = x^2 - x^1$] to calculate the difference scores, coinciding with the beginning and end of the testing protocol for each participant, respectively. Visualizations shown in Figure 20 were

used to determine HRVO categories.

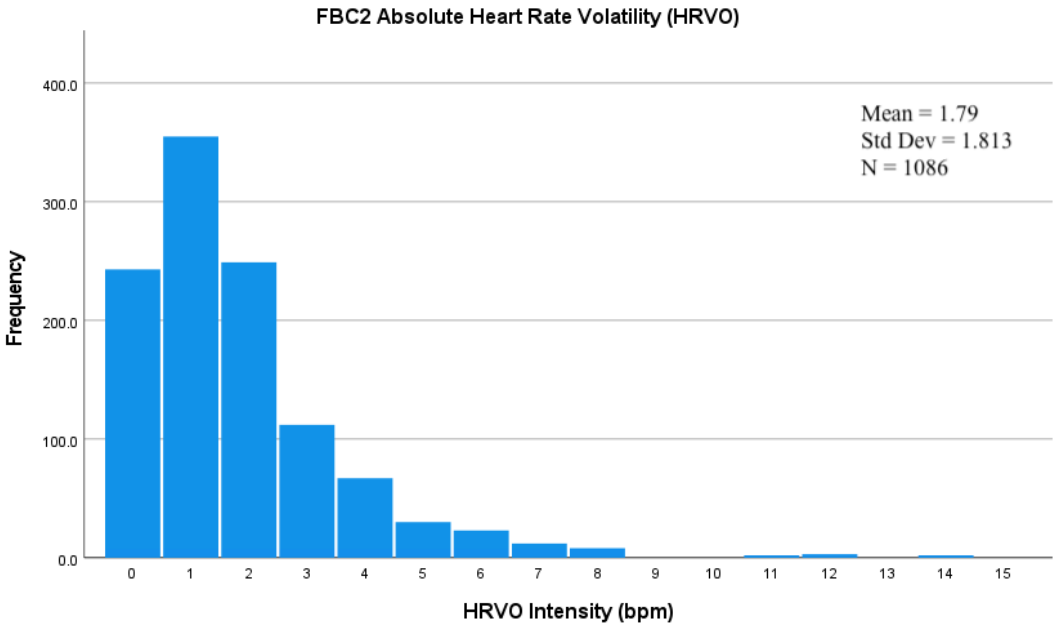

Figure 20. Histogram illustrating the distribution of FBC2 absolute HRVO

To quantify the relationship between HRVO intensity on observed agreement and precision of the error bias (standard deviation), Table 6 shows a trend for increasing HRVO and the influence on error outcomes but a one-way ANOVA was not performed.

Table 6
Influence of Heart Rate Volatility on Error Bias

FBC2 Heart Rate Volatility Intensity (bpm)	MAE	SD	COUNT
+ 0 – 1	1.32	1.82	243
+ 1 – 2	1.30	1.85	335
+ 2 – 3	1.36	1.78	249
+ 3 – 4	1.81	2.30	112
+ 4 – 5	1.93	2.42	67
+ > 5	2.62	3.28	80
Total			**1086**

5.0 DISCUSSION

Because the current wearable validation literature is inconsistent in reporting practices, research design and analyses, there are few studies with which to compare the results of this thesis. Nonetheless, the consistent efforts in this thesis to use appropriate validation methods can point to how future studies may evolve. Wearable technologies, in any case, are a consistent health care factor, likely to exert similar or increased influences for an indefinite period (Sullivan & Lachman, 2017). Therefore, validation is an important concern, even if the goal is to alert the public to unreliable information. Given the established market interest, 'wearables' could be integrated into large-scale efforts to increase healthy, risk-reducing exercise regimens (Hodkinson et al., 2021; Ritvo et al., 2021). Similarly, if sedentary HR estimates are sufficiently valid, and responsive to stress exposures, they could be usefully applied in populations vulnerable to unhealthy HR elevations. For example, existing literature indicates that multiple disease risks elevate when RHR elevates (Aune et al., 2015, 2017). In the most recent meta-analysis of n = 87 studies (Aune (2017), the Relative Risk (RR) per 10 beats/minute elevation in resting heart rate (RHR) was 1.07 for coronary heart disease, 1.09 for sudden cardiac death, 1.18 for heart failure, 1.06 for stroke, 1.15 for cardiovascular disease, 1.14 for cancer and 1.17 for all-cause mortality. While it seems unrealistic to expect large populations to reduce their RHR to reduce disease risks, these data demonstrate that risks increase with RHR increases. Therefore, if methods were discovered to reduce RHR behaviourally (e.g. mindfulness meditation, other 'relaxation' methods, yoga, mild exercise interventions like Tai Chi and dance), one motivation to engage might be physical disease risk reduction. Furthermore, these data point to the potential for pharmacological interventions aimed at modest RHR reductions (and the potential for a combined behavioural therapy-pharmacotherapy reduction protocol).

Similar to the physical disease literature, elevated HR seems to increase risks for problematic psychiatric conditions. Elevated heart rates, either through brief spikes of elevation (e.g. Post Traumatic Stress Disorder, Panic Disorder) or chronic/frequent elevations (chronic social and general anxieties (Lader & Mathews, 1970; Taylor et al., 1982; Freedman, 1985; Taylor et al., 1986; Cameron et al., 1987; Gaffney, 1988; Meuret et al., 2011) seem to be a precipitating and/or symptomatic pattern. Because of concurrent fight-flight experiences (with HR spikes and/or modest elevations), psychiatric patients can find it difficult to approach these experiences objectively and instead often manifest an irrational helplessness in pursuing self-modification efforts. In interventions with patients with major depressive disorder (Ritvo et al., 2021) where Fitbits were distributed as part of the behavioural antidepressant package, case observations suggest that reframes of anxiety episodes as episodes of elevated HR are helpful. They provide a context for incremental improvements such that patients engage in relaxation processes when not symptomatically dominated and find more empowered mental states as a result. Nonetheless, despite such observations, if the HR data being processed is not valid, the therapeutic effects observed may be largely due to placebo effects that are highly vulnerable to eventual disruption. Given the thesis data collection, device testing and analyses, the major question is how valid Fitbit Heart Rate Charge 2 (FBC2) devices seem to be when used to assess heart rate under sedentary conditions?

The most persuasive evidence favouring Fitbit validity in this thesis are the very low error rates in relation to standard ECG. Study results (Table 5) indicate a Mean Absolute Error (MAE) of 1.5 bpm (or 3.5 bpm less than the rigorous benchmark applied). When the MAE is transformed to a Mean Absolute Percentage Error (MAPE), the error is considerably lower than the benchmark of <10%, as it comprises only 1.34%. Lastly, there did not appear to be

indications of systematic error after visual inspection of the BA LOA plot. Given the above indications, it is not surprising that the Pearson correlation is r = .97.

It is also persuasive that (see Figure 19 Line Graph) there is little impact on relative measurement at the various highs and lows of HR evoked during the protocol. While it would be ideal to assess a range beyond the 71 bpm to the 76 bpm observed, the highest HR evoked, i.e. 76 bpm in ECG, was only 2.5 bpm discrepant from the 74.5 bpm that concurrently registered on the FBC2. Similarly, the lowest HR observed, i.e. 72.5 bpm per ECG, was only 1.0 bpm discrepant from the 71.5 bpm observed on the FBC2. While we might prefer observations with greater ranges, the ranges observed represent adequate consistency.

Furthermore, most of the error bias, when observed, involved a small subset of study participants (n = 4 participants). The average MAE for this subset of participants was higher at an MAE of 5.62 bpm (range = 2.52 – 7.62 bpm). Nonetheless, as shown in Figure 16, the BA LOA plot indicates an absence of fixed or proportional bias across the range of HR measures. In other words, observations of 1120 paired values did not indicate a trend of consistently higher or lower observations or error that increases in proportion to the level of the measured variable. In BA analyses, the absence of consistent bias is a positive indicator, suggesting that a minor degree of systematic (problematic) error is evident.

Although these results are promising, it is unclear how the LOA observed can be generalized to real-world applications, as the field of device validation is relatively new and without systematic cut-off values for event-based monitoring (Sequeira et al., 2019). Accordingly, we referenced the cut-off values for medical-grade ECG which is <10% MAPE (Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065), 2018.). Again, the MAPE

observed in FBC2 is well below < 10% MAPE.

Study Strengths, Weaknesses and Limitations

One strength of the FBC2 device tested is the capacity to observe a full day of HR when the client wears the device consistently for 24 hours. Therefore, both minute-level HR and the longer episode statistical derivatives can potentially support meaningful insights into autonomic functioning (e.g. morning vs. afternoon vs. evening time periods). Another strength, of course, was the surprising, previously stated, correspondence with ECG signalling when participants remained sedentary.

A key study weakness is that the results show that the residual ME = -1.1 (SD 2.3) bpm results were not apparently normally distributed as verified by Figures 14 and Figure 15. Despite this observation, a final decision was to **not** log transform the data in order to preserve the original measurement units (in bpm) and to maintain a point of reference for future cut-off score consensus discussion. Furthermore, preserving the original units facilitates comparisons with other validation studies, including future ones. While the shape of the distribution and numerical error calculations indicate a majority of observations slightly underestimate the line of equality (ME = 0 bpm), this pattern is consistent and should not present a problem in understanding the relative differences of subject's heart rate reactivity or volatility in relation to ECG-type comparisons. Also, because this study has a relatively large sample size of N=1120 paired-observations, the outlier scores beyond the LOA (containing 95% of paired-values) reflect a small proportion of observations (i.e. just 6.8% of all observations).

The analyses derived from the results provided cannot account for random error produced by either device and therefore the findings should be assessed in the larger context of other

publications involving similar study protocols and analysis. At the time of study conduct (2018/2019), few studies had conducted experiments in a psychosocial context. Furthermore, the limited guidelines for PPG HR in biobehavioral research suggesting a need for more standardized testing protocols. The status quo is changing as new recommendations and a checklist for consumer wearable HR devices were recently published (Mühlen et al., 2021).

Regarding sample size, we measured HR in bpm as this was shown in reviews to be the most common approach and one that could account for the variable sampling rate of FBC2 (Nelson et al., 2020; Mühlen et al., 2021). It is unknown whether higher or lower agreement would have been established in the present study had the sampling rate (and number of samples) been increased from 1-min bins to 30 seconds (1120 → 2240 data points) over the same durations. One study suggests that sampling rate changes (30 and 45 seconds as compared 1-min bins) result in reduced agreement outcomes (Dur et al., 2018). While greater volatility might have been observed, the higher sensitivity associated with a faster sampling rate might also have resulted in a 'smoothing' of the average error. We also note that while the number of paired-values falling beyond the LOA was low (6.8%), it is unclear why the majority was concentrated within a small subset of participants. Error trends shown in Table 6 are intended to account for the role of heart rate volatility (HRVO) in relation to agreement.

In terms of the generalizability of our findings, the sample was disproportionately female (7:1) reducing the power to generalize results to young (healthy) male adults. Shcherbina et al, (2017) demonstrated that the error rates for males were significantly higher than for females, although their study protocol emphasized physical activity (Shcherbina et al., 2017).

Although our study collected demographics by ethnicity and self-reported skin typing, there was not sufficient power to verify the relationship between skin-reflecting properties and

agreement. Nonetheless, group and individual results were analyzed, as shown in Table 4 and Figure 18. A more thorough investigation of the influence of skin-typing and agreement is explored in another validation study of WOHRM (Bent et al., 2020).

Future Research Opportunities

The present study on commercial wearables emphasized utility tasks implicating a mental health focus. The latter follows decades of credible HRR research that consistently shows that activation of the autonomic nervous system in response to acute cognitive-emotional stress increases metabolic demand on the cardiovascular system (Brouwer & Hogervorst, 2014). This wearable focus has evolved into a rapidly evolving area of applied research with expert opinions and checklists taken into consideration when conducting validation studies with wearable devices (Mühlen et al., 2021). Now that this technology is commercially available and rapidly advancing to address the needs of chronically diseased populations, this study can invigorate interests in exploring clinical HR via a mental health lens. In pursuing this research there is promise for developing cost-effective and timely treatment protocols that address issues of perceptual processing of visceral events.

Open-source Fitbit API enables users to produce multiple physiological and behavioral inputs such as steps, HR, and sleep along the same time-series index allowing for greater specificity of programming to control and isolate for relevant biomarkers or key behavioral indicators. Understanding the nature of how each input is processed and the extent to which the computed values are valid (Haghayegh et al., 2019) is essential to unlocking novel approaches to advancing the field. To our knowledge, while HR can be sampled at variable rates, the resulting data from the Fitbit API with regards to steps and sleep is limited to exactly 1-min bins.

Some other examples of research topics where WOHRM could advance study include the following: sleep patterns/quality of sleep for various clinical populations known to express insomnia or frequented night terrors (Guerrero & Crocq, 1994; Kaup et al., 1994); employee work stress during interpersonal communication such as meetings and presentations (e.g. Zoom meetings); ambulatory panic assessment (Leibold & Schruers, 2018); determining responders vs. non-responders during meditation. More recently, data mining of consented user intraday HR from Fitbit devices was used to predict "long covid symptoms" symptoms after viral infections (Radin et al., 2021). Given the possible expansion of uses, there is the prospect of developing a theoretical framework for how the measurement of biosignals (termed "digital biomarkers"), enabled by wearables, could be integrated with existing mHealth programming to improve low engagement trends (Ng et al., 2019; Melcher et al., 2020).

Lastly, we acknowledge that the newer and more robust technologies that may be coming to the commercial and research markets (and not limited to Fitbit branded devices) could accomplish multiple goals with greater sensitivity and specificity for mental stress events (Bent et al., 2020). Nonetheless, the assumptions of the 'new' technologies must be tested in order to properly inform the public.

6.0 CONCLUSION

The FBC2 appeared to adequately estimate HR in young healthy adults within a controlled laboratory setting for data collection. This study showed that FBC2 HR slightly underestimated ECG HR yet performed significantly better than the guidelines for acceptable device error (ECG – ECG testing). These findings are generalizable to adults who exhibit mental stress during bouts of inactivity and sedentariness. A wide array of real-world applications could positively impact psychiatry and clinical psychology assuming the results observed in this thesis are replicated and then generalized to free-living conditions.

REFERENCES

Abu-Arafeh, A., Jordan, H., & Drummond, G. (2016). Reporting of method comparison studies: A review of advice, an assessment of current practice, and specific suggestions for future reports. *British Journal of Anaesthesia*, *117*(5). https://doi.org/10.1093/bja/aew320

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1–R39. https://doi.org/10.1088/0967-3334/28/3/R01

ANSI/AAMI/IEC 60601-2-27:2011 (R2016)—Medical electrical equipment—Part 2-27: Particular requirements for the basic safety and essential performance of electrocardiographic monitoring equipment. (n.d.). Retrieved July 18, 2021, from https://webstore.ansi.org/Standards/AAMI/ANSIAAMIIEC60601272011R2016

Arpin-Cribbie, C. A., Irvine, J., Ritvo, P., Cribbie, R. A., Flett, G. L., & Hewitt, P. L. (2008). Perfectionism and Psychological Distress: A Modeling Approach to Understanding their Therapeutic Relationship. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *26*(3), 151–167. https://doi.org/10.1007/s10942-007-0065-2

Arpin-Cribbie, C., Irvine, J., & Ritvo, P. (2012). Web-based cognitive-behavioral therapy for perfectionism: A randomized controlled trial. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *22*(2), 194–207. https://doi.org/10.1080/10503307.2011.637242

Aune, D., Ó Hartaigh, B., & Vatten, L. J. (2015). Resting heart rate and the risk of type 2 diabetes: A systematic review and dose--response meta-analysis of cohort studies. *Nutrition, Metabolism, and Cardiovascular Diseases: NMCD*, *25*(6), 526–534. https://doi.org/10.1016/j.numecd.2015.02.008

Aune, D., Sen, A., ó'Hartaigh, B., Janszky, I., Romundstad, P. R., Tonstad, S., & Vatten, L. J. (2017). Resting heart rate and the risk of cardiovascular disease, total cancer, and all-cause mortality – A systematic review and dose–response meta-analysis of prospective studies. *Nutrition, Metabolism and Cardiovascular Diseases*, *27*(6), 504–517. https://doi.org/10.1016/j.numecd.2017.04.004

Azam, M. A. (2015). *The Psychophysiology of Maladaptive Perfectionism and Mindfulness Meditation: An Investigation Using Heart Rate Variability*. https://yorkspace.library.yorku.ca/xmlui/handle/10315/30688

Azam, M. A., Katz, J., Fashler, S. R., Changoor, T., Azargive, S., & Ritvo, P. (2015). Heart rate variability is enhanced in controls but not maladaptive perfectionists during brief mindfulness meditation following stress-induction: A stratified-randomized trial. *International Journal of Psychophysiology*, *98*(1), 27–34. https://doi.org/10.1016/j.ijpsycho.2015.06.005

Azam, M. A., Katz, J., Mohabir, V., & Ritvo, P. (2016a). Individuals with tension and migraine headaches exhibit increased heart rate variability during post-stress mindfulness meditation practice but a decrease during a post-stress control condition – A randomized, controlled experiment. *International Journal of Psychophysiology*, *110*, 66–74. https://doi.org/10.1016/j.ijpsycho.2016.10.011

Azam, M. A., Katz, J., Mohabir, V., & Ritvo, P. (2016b). Individuals with tension and migraine headaches exhibit increased heart rate variability during post-stress mindfulness meditation practice but a decrease during a post-stress control condition—A randomized, controlled experiment. *International Journal of Psychophysiology: Official Journal of the*

International Organization of Psychophysiology, *110*, 66–74. https://doi.org/10.1016/j.ijpsycho.2016.10.011

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky inventory of mindfulness skills. *Assessment*, *11*(3), 191–206. https://doi.org/10.1177/1073191104268029

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020a). Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digital Medicine*, *3*, 18. https://doi.org/10.1038/s41746-020-0226-6

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020b). Investigating sources of inaccuracy in wearable optical heart rate sensors. *Npj Digital Medicine*, *3*(1), 1–9. https://doi.org/10.1038/s41746-020-0226-6

Bland, J. M., & Altman, D. G. (1986a). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*, *1*(8476), 307–310.

Bland, J. M., & Altman, D. G. (1986b). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*, *1*(8476), 307–310.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the*

International Society of Ultrasound in Obstetrics and Gynecology, *22*(1), 85–93. https://doi.org/10.1002/uog.122

Bland, J. M., & Altman, D. G. (2007). Agreement Between Methods of Measurement with Multiple Observations Per Individual. *Journal of Biopharmaceutical Statistics*, *17*(4), 571–582. https://doi.org/10.1080/10543400701329422

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Brouwer, A.-M., & Hogervorst, M. A. (2014). A new paradigm to induce mental stress: The Sing-a-Song Stress Test (SSST). *Frontiers in Neuroscience*, *8*, 224. https://doi.org/10.3389/fnins.2014.00224

Cameron, O. G., Lee, M. A., Curtis, G. C., & McCann, D. S. (1987). Endocrine and physiological changes during "spontaneous" panic attacks. *Psychoneuroendocrinology*, *12*(5), 321–331. https://doi.org/10.1016/0306-4530(87)90061-8

Doğan, N. Ö. (2018). Bland-Altman analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine*, *18*(4), 139–141. https://doi.org/10.1016/j.tjem.2018.09.001

Dorosz, J. L., Lezotte, D. C., Weitzenkamp, D. A., Allen, L. A., & Salcedo, E. E. (2012). Performance of 3-Dimensional Echocardiography in Measuring Left Ventricular Volumes and Ejection Fraction. *Journal of the American College of Cardiology*, *59*(20), 1799–1808. https://doi.org/10.1016/j.jacc.2012.01.037

Dur, O., Rhoades, C., Ng, M. S., Elsayed, R., van Mourik, R., & Majmudar, M. D. (2018). Design Rationale and Performance Evaluation of the Wavelet Health Wristband:

Benchtop Validation of a Wrist-Worn Physiological Signal Recorder. *JMIR MHealth and UHealth*, *6*(10), e11040. https://doi.org/10.2196/11040

Falter, M., Budts, W., Goetschalckx, K., Cornelissen, V., & Buys, R. (2019). Accuracy of Apple Watch Measurements for Heart Rate and Energy Expenditure in Patients With Cardiovascular Disease: Cross-Sectional Study. *JMIR MHealth and UHealth*, *7*(3), e11889. https://doi.org/10.2196/11889

Fechir, M., Schlereth, T., Purat, T., Kritzmann, S., Geber, C., Eberle, T., Gamer, M., & Birklein, F. (2008). Patterns of Sympathetic Responses Induced by Different Stress Tasks. *The Open Neurology Journal*, *2*, 25–31. https://doi.org/10.2174/1874205X00802010025

Fitabase. (2020). Fitabase. https://www.fitabase.com/research-library/

Fitzpatrick, T. B. (1988). The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology*, *124*(6), 869–871. https://doi.org/10.1001/archderm.1988.01670060015008

Flett, G. L., Hewitt, P. L., Whelan, T., & Martin, T. R. (2007). The Perfectionism Cognitions Inventory: Psychometric properties and associations with distress and deficits in cognitive self-management. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *25*(4), 255–277. https://doi.org/10.1007/s10942-007-0055-4

Freedman, R. R. (1985). Ambulatory Monitoring of Panic Disorder. *Archives of General Psychiatry*, *42*(3), 244. https://doi.org/10.1001/archpsyc.1985.01790260038004

Gaffney, F. A. (1988). Hemodynamic, Ventilatory, and Biochemical Responses of Panic Patients and Normal Controls With Sodium Lactate Infusion and Spontaneous Panic Attacks. *Archives of General Psychiatry*, *45*(1), 53. https://doi.org/10.1001/archpsyc.1988.01800250063008

Gerke, O. (2020). Reporting Standards for a Bland–Altman Agreement Analysis: A Review of Methodological Reviews. *Diagnostics*, *10*(5), 334. https://doi.org/10.3390/diagnostics10050334

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Grossman, P. (2008). On measuring mindfulness in psychosomatic and psychological research. *Journal of Psychosomatic Research*, *64*(4), 405–408. https://doi.org/10.1016/j.jpsychores.2008.02.001

Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2010). Mindfulness-based stress reduction and health benefits: A meta-analysis. *Focus on Alternative and Complementary Therapies*, *8*(4), 500–500. https://doi.org/10.1111/j.2042-7166.2003.tb04008.x

Gu, J., Strauss, C., Crane, C., Barnhofer, T., Karl, A., Cavanagh, K., & Kuyken, W. (2016). Examining the factor structure of the 39-item and 15-item versions of the Five Facet Mindfulness Questionnaire before and after mindfulness-based cognitive therapy for people with recurrent depression. *Psychological Assessment*, *28*(7), 791–802. https://doi.org/10.1037/pas0000263

Guerrero, J., & Crocq, M.-A. (1994). Sleep disorders in the elderly: Depression and post-traumatic stress disorder. *Journal of Psychosomatic Research*, *38*(Suppl 1), 141–150. https://doi.org/10.1016/0022-3999(94)90144-9

Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R., & Castriotta, R. J. (2019). Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *21*(11), e16273. https://doi.org/10.2196/16273

Hahnen, C., Freeman, C. G., Haldar, N., Hamati, J. N., Bard, D. M., Murali, V., Merli, G. J., Joseph, J. I., & Helmond, N. van. (2020). Accuracy of Vital Signs Measurements by a Smartwatch and a Portable Health Device: Validation Study. *JMIR MHealth and UHealth*, *8*(2), e16811. https://doi.org/10.2196/16811

Hajcak, G., McDonald, N., & Simons, R. F. (2004). Error-related psychophysiology and negative affect. *Brain and Cognition*, *56*(2), 189–197. https://doi.org/10.1016/j.bandc.2003.11.001

Hertzman, A. B. (1938). THE BLOOD SUPPLY OF VARIOUS SKIN AREAS AS ESTIMATED BY THE PHOTOELECTRIC PLETHYSMOGRAPH. *American Journal of Physiology*. https://doi.org/10.1152/ajplegacy.1938.124.2.328

Hollon, S. D., & Kendall, P. C. (1980). Cognitive self-statements in depression: Development of an automatic thoughts questionnaire. *Cognitive Therapy and Research*, *4*(4), 383–395. https://doi.org/10.1007/BF01178214

Jo, E., Lewis, K., Directo, D., Kim, M. J., & Dolezal, B. A. (2016). Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. *Journal of Sports Science & Medicine*, *15*(3), 540–547.

Johnston, W., Judice, P. B., García, P. M., Mühlen, J. M., Skovgaard, E. L., Stang, J., Schumann, M., Cheng, S., Bloch, W., Brønd, J. C., Ekelund, U., Grøntved, A., Caulfield, B., Ortega, F. B., & Sardinha, L. B. (2021). Recommendations for determining the validity of consumer wearable and smartphone step count: Expert statement and checklist of the INTERLIVE network. *British Journal of Sports Medicine*, *55*(14), 780–793. https://doi.org/10.1136/bjsports-2020-103147

Kamshilin, A. A., & Margaryants, N. B. (2017). Origin of Photoplethysmographic Waveform at Green Light. *Physics Procedia*, *86*, 72–80. https://doi.org/10.1016/j.phpro.2017.01.024

Karras, D. J. (1997a). Statistical Methodology: II. Reliability and Validity Assessment in Study Design, Part A. *Academic Emergency Medicine*, *4*(1), 64–71. https://doi.org/10.1111/j.1553-2712.1997.tb03646.x

Karras, D. J. (1997b). Statistical methodology: II. Reliability and validity assessment in study design, Part B. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, *4*(2), 144–147. https://doi.org/10.1111/j.1553-2712.1997.tb03723.x

Kaup, B. A., Ruskin, P. E., & Nyman, G. (1994). Significant life events and PTSD in elderly World War II veterans. *The American Journal of Geriatric Psychiatry*, *2*(3), 239–243. https://doi.org/10.1097/00019442-199400230-00008

Kirk, M., Taha, B., McCague, H., Hatzinakos, D., Dang, K., Katz, J., and Ritvo, P. (in press) Online Cognitive Behavioral Therapy, Mindfulness Meditation, and Yoga (CBT-MY) Intervention for Posttraumatic Stress Disorder: Psychometric and Psychophysiology Outcomes, Journal of Medical Internet Research - Mental Health (in press)

Konstantinou, P., Trigeorgi, A., Georgiou, C., Gloster, A. T., Panayiotou, G., & Karekla, M. (2020). Comparing apples and oranges or different types of citrus fruits? Using wearable versus stationary devices to analyze psychophysiological data. *Psychophysiology*, *57*(5), e13551. https://doi.org/10.1111/psyp.13551

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroll, R. R., McKenzie, E. D., Boyd, J. G., Sheth, P., Howes, D., Wood, M., Maslove, D. M., & for the WEARable Information Technology for hospital INpatients (WEARIT-IN) study group. (2017). Use of wearable devices for post-discharge monitoring of ICU patients: A feasibility study. *Journal of Intensive Care*, *5*(1), 64. https://doi.org/10.1186/s40560-017-0261-9

Lader, M., & Mathews, A. (1970). Physiological changes during spontaneous panic attacks. *Journal of Psychosomatic Research*, *14*(4), 377–382. https://doi.org/10.1016/0022-3999(70)90004-8

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In *Attention and orienting: Sensory and motivational processes* (pp. 97–135). Lawrence Erlbaum Associates Publishers.

Leibold, N. K., & Schruers, K. R. (2018). Assessing Panic: Bridging the Gap Between Fundamental Mechanisms and Daily Life Experience. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00785

Ludbrook, J. (2010). Confidence in Altman-Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology & Physiology*, *37*(2), 143–149. https://doi.org/10.1111/j.1440-1681.2009.05288.x

Melcher, J., Camacho, E., Lagan, S., & Torous, J. (2020). College student engagement with mental health apps: Analysis of barriers to sustained use. *Journal of American College Health: J of ACH*, 1–7. https://doi.org/10.1080/07448481.2020.1825225

Menghini, L., Gianfranchi, E., Cellini, N., Patron, E., Tagliabue, M., & Sarlo, M. (2019). Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology*, *56*(11), e13441. https://doi.org/10.1111/psyp.13441

Meuret, A. E., Rosenfield, D., Wilhelm, F. H., Zhou, E., Conrad, A., Ritz, T., & Roth, W. T. (2011). Do Unexpected Panic Attacks Occur Spontaneously? *Biological Psychiatry*, *70*(10), 985–991. https://doi.org/10.1016/j.biopsych.2011.05.027

Mühlen, J. M., Stang, J., Lykke Skovgaard, E., Judice, P. B., Molina-Garcia, P., Johnston, W., Sardinha, L. B., Ortega, F. B., Caulfield, B., Bloch, W., Cheng, S., Ekelund, U., Brønd, J. C., Grøntved, A., & Schumann, M. (2021). Recommendations for determining the validity of consumer wearable heart rate devices: Expert statement and checklist of the INTERLIVE Network. *British Journal of Sports Medicine*, *55*(14), 767–779. https://doi.org/10.1136/bjsports-2020-103148

Nelson, B. W., & Allen, N. B. (2019). Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study. *JMIR MHealth and UHealth*, *7*(3). https://doi.org/10.2196/10828

Nelson, B. W., Low, C. A., Jacobson, N., Areán, P., Torous, J., & Allen, N. B. (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *Npj Digital Medicine*, *3*(1), 1–9. https://doi.org/10.1038/s41746-020-0297-4

Ng, M. M., Firth, J., Minen, M., & Torous, J. (2019). User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. *Psychiatric Services (Washington, D.C.)*, *70*(7), 538–544. https://doi.org/10.1176/appi.ps.201800519

Nitzan, M., Romem, A., & Koppel, R. (2014). Pulse oximetry: Fundamentals and technology update. *Medical Devices (Auckland, N.Z.)*, *7*, 231–239. https://doi.org/10.2147/MDER.S47319

Olofsen, E., Dahan, A., Borsboom, G., & Drummond, G. (2015). Improvements in the application and reporting of advanced Bland-Altman methods of comparison. *Journal of*

Clinical Monitoring and Computing, *29*(1), 127–139. https://doi.org/10.1007/s10877-014-9577-3

Physical Activity Monitoring for Heart Rate (ANSI/CTA-2065). (n.d.). Consumer Technology Association®. Retrieved May 17, 2021, from https://shop.cta.tech/products/physical-activity-monitoring-for-heart-rate

Pires, I. M., Oniani, S., Woolley, S., & Collins, T. (2018). *How reliable is your wearable heart-rate monitor?* The Conversation. http://theconversation.com/how-reliable-is-your-wearable-heart-rate-monitor-98095

Pleil, J. D. (2016). QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics. *Journal of Breath Research*, *10*(3), 035001. https://doi.org/10.1088/1752-7155/10/3/035001

Preiss, D., & Fisher, J. (2008). A Measure of Confidence in Bland–Altman Analysis for the Interchangeability of Two Methods of Measurement. *Journal of Clinical Monitoring and Computing*, *22*(4), 257–259. https://doi.org/10.1007/s10877-008-9127-y

Radin, J. M., Quer, G., Ramos, E., Baca-Motes, K., Gadaleta, M., Topol, E. J., & Steinhubl, S. R. (2021). Assessment of Prolonged Physiological and Behavioral Changes Associated With COVID-19 Infection. *JAMA Network Open*, *4*(7), e2115959. https://doi.org/10.1001/jamanetworkopen.2021.15959

Ritvo, P., Knyahnytska, Y., Pirbaglou, M., Wang, W., Tomlinson, G., Zhao, H., Linklater, R., Bai, S., Kirk, M., Katz, J., Harber, L., & Daskalakis, Z. (2021). Online Mindfulness-Based Cognitive Behavioral Therapy Intervention for Youth With Major Depressive Disorders: Randomized Controlled Trial. *Journal of Medical Internet Research*, *23*(3), e24380. https://doi.org/10.2196/24380

Sachdeva, S. (2009). Fitzpatrick skin typing: Applications in dermatology. *Indian Journal of Dermatology, Venereology, and Leprology*, *75*(1), 93. https://doi.org/10.4103/0378-6323.45238

Schuurmans, A. A. T., de Looff, P., Nijhof, K. S., Rosada, C., Scholte, R. H. J., Popma, A., & Otten, R. (2020). Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: A Comparison to Electrocardiography (ECG). *Journal of Medical Systems*, *44*(11), 190. https://doi.org/10.1007/s10916-020-01648-w

Sequeira, N., Giglia, V., Chow, C., Mariano, Z., Angaran, P., & Dorian, P. (2019). COMMON WEARABLE FITNESS DEVICES DEMONSTRATE VARIABLE ACCURACY IN MEASURING HEART RATE DURING SUPRAVENTRICULAR TACHYCARDIA. *Canadian Journal of Cardiology*, *35*(10), S110–S111. https://doi.org/10.1016/j.cjca.2019.07.508

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., & Ashley, E. A. (2017). Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*, *7*(2), 3. https://doi.org/10.3390/jpm7020003

Sliney, D. H. (2016). What is light? The visible spectrum and beyond. *Eye*, *30*(2), 222–229. https://doi.org/10.1038/eye.2015.252

Sperlich, B., & Holmberg, H.-C. (2017). Wearable, yes, but able…?: It is time for evidence-based marketing claims! *British Journal of Sports Medicine*, *51*(16), 1240–1240. https://doi.org/10.1136/bjsports-2016-097295

Spitzer, R. L. (1999). Validation and Utility of a Self-report Version of PRIME-MDThe PHQ Primary Care Study. *JAMA*, *282*(18), 1737. https://doi.org/10.1001/jama.282.18.1737

Sprinkle, S. D., Lurie, D., Insko, S. L., Atkinson, G., Jones, G. L., Logan, A. R., & Bissada, N. N. (2002). Criterion validity, severity cut scores, and test-retest reliability of the Beck Depression Inventory-II in a university counseling center sample. *Journal of Counseling Psychology*, *49*(3), 381–385. https://doi.org/10.1037/0022-0167.49.3.381

Steer, R. A., & Beck, A. T. (1997). Beck Anxiety Inventory. In *Evaluating stress: A book of resources* (pp. 23–40). Scarecrow Education.

Sullivan, A. N., & Lachman, M. E. (2017). Behavior Change with Fitness Technology in Sedentary Adults: A Review of the Evidence for Increasing Physical Activity. *Frontiers in Public Health*, *4*, 289. https://doi.org/10.3389/fpubh.2016.00289

Taylor, C. B., Sheikh, J., Agras, W. S., Roth, W. T., Margraf, J., Ehlers, A., Maddock, R. J., & Gossard, D. (1986). Ambulatory heart rate changes in patients with panic attacks. *The American Journal of Psychiatry*, *143*(4), 478–482. https://doi.org/10.1176/ajp.143.4.478

Taylor, C. B., Telch, M. J., & Havvik, D. (1982). Ambulatory heart rate changes during panic attacks. *Journal of Psychiatric Research*, *17*(3), 261–266. https://doi.org/10.1016/0022-3956(82)90004-8

Tremblay, M. S., Aubert, S., Barnes, J. D., Saunders, T. J., Carson, V., Latimer-Cheung, A. E., Chastin, S. F. M., Altenburg, T. M., & Chinapaw, M. J. M. (2017). Sedentary Behavior Research Network (SBRN) – Terminology Consensus Project process and outcome. *The International Journal of Behavioral Nutrition and Physical Activity*, *14*, 75. https://doi.org/10.1186/s12966-017-0525-8

Tsibadze, A., Chikvaidze, E., Katsitadze, A., Kvachadze, I., Tskhvediani, N., & Chikviladze, A. (2015). [VISIBLE LIGHT AND HUMAN SKIN (REVIEW)]. *Georgian Medical News*, *246*, 46–53.

Zaki, R., Bulgiba, A., Nordin, N., & Azina Ismail, N. (2013). A Systematic Review of Statistical Methods Used to Test for Reliability of Medical Instruments Measuring Continuous Variables. *Iranian Journal of Basic Medical Sciences*, *16*(6), 803–807.

APPENDICES

Appendix A: Study Consent Form

Study Consent Form

Study Name: Comparing the Fitbit Charge 2 HR Monitor with electrocardiography during Sedentary Behaviours, Cognitive-Emotional Stress, and Mindfulness Meditation: A Validation study

Principal Investigator:	Spencer Williams, BSc., MSc., (cand). williams.spencer15@gmail.com 138 Chemistry Building York University
Supervisor:	Dr. Paul Ritvo pritvo@yorku.ca 136 Chemistry Building York University

You are being asked to take part in a research study. Before agreeing to take part in this study, it is important that you read and understand the following explanation of the proposed study procedures. The following information describes the purpose, procedures, potential harms/risk/discomforts, and the benefits associated with this study. It also describes your right to refuse to participate or to withdraw from any part of the study at any time.

In order to decide whether you wish to participate in this research study, you should understand enough about it to make an informed decision. This is known as the informed consent process. Please ask the researcher to explain any words you do not understand before signing this consent form. Make sure all your questions have been answered to your satisfaction before signing this document.

Purpose

You have been asked to participate in a study designed to validate the efficacy of Fitbit devices (Fitbit Charge 2) through a series of passive and active tasks using a computer and guided audio for listening to mindfulness meditation. Electrocardiogram (ECG) recordings and respiration

measurements will be taken.

Role of Research Participant

Your participation in this study will entail a time commitment of 65 minutes. If you agree to participate in this study, you will be asked to provide your up-to-date medical history during the pre-screening. These questions will help the researcher determine if you are eligible to participate in the study. Feel free to ask the researcher to clarify the questions, if necessary.

If you choose to partake in this study, you will be asked to complete a total of 8 questionnaires, view still-frame images, complete a cognitive task and listen to a 15-minute recording related to mindfulness meditation.

Questionnaires:

(i) Demographic & Health Questionnaire

(ii) Five Facet Mindfulness Questionnaire (FFMQ): is commonly used to assess changes in mindfulness tendencies

(iii) Beck Anxiety Inventory: examines baseline anxiety levels

(iv) Center for Epidemiologic Studies Depression Scale: assesses depressive symptoms

(v) Automatic Thoughts Questionnaire: evaluates automatic negative thoughts

(vi) Perfectionism Cognitions Inventory (PCI): assess the frequency of cognitions focused on the discrepancies between one's current and ideal self.

(vii) Qualifying Mindfulness Meditation Self-Report (QMM-SR): assess the perceived quality of the experience while practicing formal meditation

(viii) Fitzpatrick Skin Type Scale (photo-typing) (FSTS): create a classification system for dermatologists assign to patients in order to determine a person's risk to ultraviolet light radiation

(ix) Patient Health Questionnaire (PHQ-9): provides a measure of overall mental health

(x) Mind-Wandering Inventory (MWI): assessing the frequency of different types of mind-wandering events

Cognitive Task:
Over a period of 5-minutes, you will be asked to solve a pattern consisting of four alphanumeric characters presented at the top of the screen that will be followed by a 5th character. You will be asked to click either "True" or "False", respectively, depending on whether or not the 5th character follows the same pattern as the previous four characters. HR and respiration measures will be taken during this task.

IAPS Imagery:
You will be asked to sit quietly and still observing for 10 minutes as images are presented on a computer screen. HR and Respiration will be taken during this task as well.

Listening Task:
You will be asked to sit quietly and still for 10-minutes while listening to a recording related to mindfulness meditation. HR and respiration measures will be taken during this task.

Potential Harms, Risks or Discomforts

Some of the questions ask about private matters such as whether or not you have been clinically diagnosed with a mental disorder or if you have any illnesses/diseases. To mitigate this potential risk, you may choose not to answer the question. There is no risk for undergoing ECG recordings or respiration measurements, however, some participants may feel uncomfortable performing a cognitively demanding task while attached to various pieces of equipment. Moreover, the light emitting diodes nor the band material has shown to reliable hard consumers of such wearables. If at any point you feel uncomfortable, please let the researcher know and you may take breaks at any point.

Potential Benefits:
There are no substantial benefits associated with participating in this study. We hope this investigation will lead to a better understanding of how wearables devices (i.e. Fitbits) can be incorporated for clinical applications.

Voluntary Participation and Withdrawal

Your participation in this study is completely voluntary and you may choose to stop participating at any time. Your decision not to volunteer, to stop participating, or to refuse to answer particular questions will not influence the nature of the ongoing relationship with York University either now, or in the future. If you decide to stop participating, you may withdraw without penalty, financial or otherwise and you will still receive the promised inducement (URPP credit). In the

event you withdraw from the study, all associated data collected will be immediately destroyed wherever possible. Should we wish to withdraw after the study you will have the option to also withdraw your data up until the analysis is complete.

Confidentiality:

Unless you choose otherwise, all the information you supply during the research will be held in confidence and unless you specifically indicate your consent, your name will not appear in any report or publication of the research. The experimental data acquired in this study is collected in an anonymized form that cannot be connected back to you. This research is used for teaching purposes be presented at meetings, published, shared with other scientific researchers or used in future studies. Your name or other identifying information will not be used in any publication or teaching materials without your specific permission. Data will be collected via ECG, Fitbit Charge 2, respiration belt, and the computer. Handwritten notes will not have any identifying information. Written data will be safely stored in a locked filing cabinet and only the researcher will have access to this information. Any data collected via computer will be securely stored on a password protected USB keys and laptops with solely the researcher having access to this data. Electronic data will be retained on a password protected USB in a locked filing cabinet.

The researcher acknowledges that the host of the online survey (Survey Monkey) may automatically collect participant data without their knowledge (i.e., IP addresses). Although this information may be provided or made accessible to the researchers, it will not be used or saved without the participant's consent on the researcher's system. Further, behavior this project employs e-based collection techniques, data may be subject to access by third parties as a result of various security-based legislation now in place in many countries and thus the confidentiality and privacy of data cannot be guaranteed during web-based transmission.

Questions

If you have questions about the research in general or about your role in the study, please feel free to contact Spencer Williams by email (williams.spencer15@gmail.com) or Dr. Paul Ritvo by telephone at (416) 736-2100 ext. 22396 or by email (pritvo@yorku.ca). You can also contact the Department of Kinesiology and Health Science by telephone (416) 736-5728 or email (kahs@yorku.ca).

This research has received ethics review and approval by the Delegated Ethics Review Committee, which is delegated authority to review research ethics protocols by the Human Participants Review Sub-Committee, York University's Ethics Review Board, and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines. If you have any questions

about this process, or about your rights as a participant in the study, please contact the Sr. Manager & Policy Advisor for the Office of Research Ethics, 5th Floor, Kaneff Tower, York University (telephone 416-736-5914 or email ore@yorku.ca).

Legal Rights and Signatures:

I, _____, consent to participate in this research study, "Comparing the Fitbit Charge 2 HR Monitor with electrocardiography during Sedentary Behaviours, Cognitive-Emotional Stress, and Mindfulness Meditation: A Validation study" conducted by Spencer Williams. I have understood the nature of this project and wish to participate. I am not waiving any of my legal rights by signing this form. My signature below indicates my consent.

_____	_____	_____
Name of Participant	**Participant Signature**	**Date**

Appendix B: Study Protocol Set-up Guide

Setting up ECG/Respiration Belt

Respiration Belt

1. **Setting up the respiratory belt transducer- better to put on belt FIRST**
   - "Please secure the respiratory belt transducer using the Velcro around their waist (under their shirt) below their rib cage, with the black strap facing the front of the diaphragm.
   - "Please make sure the belt is secure enough so that it moves in and out as they breathe, but that it's not too tight so as to obstruct comfortable breathing."

ECG

1. After the participant has completed the consent form, have them seated, and **explain the experimental protocol:**
   - *"We will be collecting an electrocardiogram (ECG) and respiration data for 10 minutes while you are resting, 10 minutes while you are viewing images, 5 while you complete a cognitive task (PST) that measures your intelligence, and while you listen to a 10 minute audio.*

2. **Prepare the electrodes:**
   - Location: box under the desk
   - Attach the leads (red, black, white) on to the metallic node of the electrodes.
   - Remove sticker backing one at a time while explaining where the electrodes belong

3. **Explain where the electrodes go:**
   - *"The negative (white) electrode must be placed on the inside of your right wrist"*
   - *"The positive (red) electrode must be placed on the inside of your left wrist"*
   - *"The ground (black) electrode must be placed on the left ankle (inside facing, tibia)"*

4. **Inform participant about the electrodes:**
   - *"Are the stickers secure? Are you comfortable with them on?"*
   - *"Please let me know if the stickers feel as if they're slipping off at any time."*
   - Remember to emphasize that when they are sticking the electrodes to their bodies they should rub around the metallic node

ECG & Respiration Sampling

1. **Open the "Sampling" system file**
   - In the 'Welcome Center' of LabChart, you should see system files labelled "Sampling". When you open "sampling" you should see 'chart view' with 3 channels on the screen:

2. **Sample their ECG & Respiration briefly on LabChart** (ensure the recording is coming through clearly and with as little noise as possible).
    - **Press the 'Start' button** on your screen to collect a recording (top right). Look at the dots above the spikes in the 'chart view' of Channel 1:
        - there should be prominent 'R' spikes
        - Blue dots: reflect normal beat identification
        - Yellow dots: the beats are being marked 'ectopic' (i.e., a disturbance of the cardiac rhythm)
    - **Select the beat classifier** (under the HRV section) to determine whether heartbeats are falling out of normal ranges (600 ms – 1000 ms), and proceed with discretion as in the beginning the participant's HR may simply be high. (There is no need to alert the participant that their heartbeats are 'irregular' or 'not normal').
    - Go back to chart view. In Channel 2 you should see the respiration signal.
        - *"Please take a deep breathe in and out"*
        - You should see the wave ascend with inhalation and descend with exhalation.
        - Don't worry if the wave appears 'jagged' or 'erratic' in the beginning, the belt transducer is most likely calibrating and will gradually smooth out. In the chart view for the respiration signal
            - ! Ensure you are seeing white dots being placed above each cycle of breathing.

Pattern Solving Task (PST) – 5 minutes

1. Open the paradigm file "PST_Mar24" on the desktop computer.
    - The "Pattern Solving Task" paradigm file should be open
2. Open up PowerPoint with PST instructions
    - Read the instructions and give them a practice round
    - *"You need to solve a pattern consisting* ***of 4 alphanumeric characters presented at the top of the screen for 8 seconds****. After 8 seconds, a* ***5th*** *character will be shown and the words* ***"True or False"*** *will appear in the middle of the screen. Press "****T" (True)*** *if you believe the 5th character follows the "correct" pattern or* ***"F" (False)*** *if it does not. Immediately after the response, you will receive feedback about whether your answer was right or wrong. A number indicating total* ***"Correct" and "Incorrect" answers (up to the current trial) will remain on the screen also in the bottom right and bottom left areas*** *(corresponding to where the "Right Answer" and "Wrong Answer" feedback appears) throughout the duration of the task*
    - "If you saw the number sequence "2, 4, 8, 16" and then "32" appeared, would you have pressed "T" or "F"? (Answer: T – true)
    - Do you have any questions?
3. Open the "PST" file in **LabChart** on the laptop
4. In **Paradigm**, click the green play button
    - Enter in participant ID > Press Continue
5. Before beginning measure the illumination in the room (in lux using the luxometer)

6. Read the following instructions to the participant: *"You are now going to be completing a cognitive task. This task will measure your intelligence. It is important that you complete this task to the best of your ability. I will be measuring your HR and breathing while you complete this task. I will remain in the room while you complete this task."*
7. *"This task will measure your capacity for cognitive processing which is related to intelligence. The average score on this task is 80-85% correct answers with response time of less than 2 seconds. This is the required minimum performance and your individual performance must be close to or equal the average performance of all subjects, therefore please* ***respond as quickly but as accurately as possible****! Also, please do not move or speak during this task. Do you have any questions?"*
8. Let them read over the instructions one more time. When they begin press start in **LabChart**

9. After the experiment, save the data as PARTICIPANT#_PST_DATE

Appendix C: Post-Debriefing Consent Form for Studies Involving Deception

POST-DEBRIEFING CONSENT FORM FOR STUDIES INVOLVING DECEPTION

Study Title: A Validation Study: A Validation Study: Fitbit Charge 2 Device Used For Measuring HR During Sedentary Behaviours, Cognitive-Emotional Stressors and Mindfulness Meditation

Principal Investigator: Spencer Williams, BSc., MSc. (cand.)

During the debriefing session, I learned that it was necessary for the researchers to disguise the real purpose of this study. I realize that this was necessary since having full information about the actual purpose of the study might have influenced the way in which I responded to the tasks and this would have invalidated the results. Thus, to ensure that this did not happen, some of the details about the purpose of the study initially were not provided (or were provided in a manner that slightly misrepresented the real purpose of the study). However, I have now received a complete verbal and written explanation as to the actual purpose of the study and have had an opportunity to ask any questions about this and to receive acceptable answers to my questions.

I have been asked to give permission for the researchers to use my data (or information I provided) in their study, and agree to this request. I am aware that I may withdraw this consent by notifying the Principal Investigator.

This study has been reviewed and received ethics clearance through the Human Participants Review Committee (HPRC). If you have questions for the Committee contact the Sr. Policy Advisor, Research Ethics, Office of Research Ethics, at 416-736-5914 or ore@yorku.ca

For all other questions contact (Paul Ritvo, C.Psych (416-580-8021, paul.ritvo@gmail.com).

Participants Name: _____
Participants Signature: _____
Date: _____
Witness' Name: _____
Witness' Signature: _____

Text adapted in whole or in part from the original University of Waterloo, "POST-DEBRIEFING CONSENT FORM
FOR STUDIES INVOLVING DECEPTION (In Lab)"

Appendix D: Demographics and Psychometric Questionnaires

Audio/ Visual Comprehension & Recent Behaviors

1. English as a first language (Y/N)
2. Does this person have difficulty seeing (if prescription glasses/ contacts are required to be worn and are readily available to you for this trial pleases select 'No')? (Y/N)
3. Does this person have difficulty hearing (if a hearing device is required to be worn and is readily available to you for this trial select 'No') (Y/N)
4. If any, please indicate if you've engaged in the following behaviours within the last 6 hrs:
    1. Alcoholic consumption
    2. Smoking
    3. Medications known to increase HR or alter breathing
    4. Consumption of a large meal (within the last hour)
    5. Consumption of caffeine
    6. Moderate or high intensity physical activity
    7. Do you have a cold or any other condition whereby breathing whereby breathing through either your mouth or nose is impaired? (Y/N)

Beck Anxiety Inventory (BAI; Beck, Epstein, Brown, Steer, 1988).

Below is a list of common symptoms of anxiety. Please carefully read each item in the list. Indicate how much you have been bothered by that symptom during the past month, including today, by circling the number in the corresponding space in the column next to each symptom.

Items	Not at All	Mildly, but it didn't bother me much	Moderately – it wasn't pleasant at times	Severely – it bothered me a lot
1. Numbness or tingling	□	□	□	□

2. Feeling Hot	☐	☐	☐	☐
3. Wobbliness in legs	☐	☐	☐	☐
4. Fear of the worst happening	☐	☐	☐	☐
5. Dizzy or lightheaded	☐	☐	☐	☐
6. Heart pounding/racing	☐	☐	☐	☐
7. Unsteady	☐	☐	☐	☐
8. Terrified or afraid	☐	☐	☐	☐
9. Nervous	☐	☐	☐	☐
10. Feeling of choking	☐	☐	☐	☐
11. Hands trembling	☐	☐	☐	☐
12. Shaky/unsteady	☐	☐	☐	☐
13. Fear of losing control	☐	☐	☐	☐
14. Difficulty in breathing	☐	☐	☐	☐
15. Nervous	☐	☐	☐	☐
16. Fear of dying	☐	☐	☐	☐
17. Scared	☐	☐	☐	☐
18. Indigestion	☐	☐	☐	☐
19. Faint/lightheaded	☐	☐	☐	☐

20. Face flushed	□	□	□	□
21. Hot/Cold Sweats	□	□	□	□

Automatic Thoughts Questionnaire (ATQ; Hollon & Kendall, 1980).

Listed below are a variety of thoughts that pop into people's heads. Please read each thought and indicate how frequently, if at all, the thought occurred to you over the last week. Please read each item carefully and circle the appropriate answers on the answer sheet in the following fashion (1 = "not at all", 2 = "sometimes", 3 = "moderately often", 4 = "often", and 5 = "all the time"). Then, please indicate how strongly, if at all, you tend to believe that thought, when it occurs. On the right hand side of the page, circle the appropriate answers in the following fashion (1 = "not at all", 2 = "somewhat", 3 = "moderately ", 4 = "very much", and 5 = "totally).

Frequency					**Items**	**Degree of Belief**				
1	2	3	4	5	1. I feel like I'm up against the world	1	2	3	4	5
1	2	3	4	5	2. I'm no good.	1	2	3	4	5
1	2	3	4	5	3. Why can't I ever succeed?	1	2	3	4	5
1	2	3	4	5	4. No one understands me.	1	2	3	4	5
1	2	3	4	5	5. I've let people down	1	2	3	4	5
1	2	3	4	5	6. I don't think I can go on.	1	2	3	4	5
1	2	3	4	5	7. I wish I were a better person	1	2	3	4	5

1	2	3	4	5	8.	I'm so weak.	1	2	3	4	5
1	2	3	4	5	9.	My life's not going the way I want it to.	1	2	3	4	5
1	2	3	4	5	10.	I'm so disappointed in myself.	1	2	3	4	5
1	2	3	4	5	11.	Nothing feels good anymore.	1	2	3	4	5
1	2	3	4	5	12.	I can't stand this anymore.	1	2	3	4	5
1	2	3	4	5	13.	I can't get started.	1	2	3	4	5
1	2	3	4	5	14.	What's wrong with me?	1	2	3	4	5
1	2	3	4	5	15.	I wish I were somewhere else	1	2	3	4	5
1	2	3	4	5	16.	I can't get things together	1	2	3	4	5
1	2	3	4	5	17.	I hate myself.	1	2	3	4	5
1	2	3	4	5	18.	I'm worthless.	1	2	3	4	5
1	2	3	4	5	19.	Wish I could just disappear.	1	2	3	4	5
1	2	3	4	5	20.	What's the matter with me?	1	2	3	4	5
1	2	3	4	5	21.	I'm a loser.	1	2	3	4	5
1	2	3	4	5	22.	My life is a mess.	1	2	3	4	5

1	2	3	4	5	23.	I'm a failure.	1	2	3	4	5
1	2	3	4	5	24.	I'll never make it.	1	2	3	4	5
1	2	3	4	5	25.	I feel so helpless.	1	2	3	4	5
1	2	3	4	5	26.	Something has to change	1	2	3	4	5
1	2	3	4	5	27.	There must be something wrong with me	1	2	3	4	5
1	2	3	4	5	28.	My future is bleak.	1	2	3	4	5
1	2	3	4	5	29.	It's just not worth it.	1	2	3	4	5
1	2	3	4	5	30.	I can't finish anything.	1	2	3	4	5

Perfectionism Cognitions Inventory (PCI; Flett, Hewitt, Whelan, & Martin, 2007)

Listed below are a variety of thoughts about perfectionism that sometimes pop into people's heads. Please read each thought and indicate how frequently, if at all, the thoughts occurred to you over the last week. Please read each item carefully and circle the appropriate number, using the scale below.

0 = Not At All, 1 = Sometimes, 2 = Moderately Often, 3 = Often, 4 = All of the Time

Items	Not at All	Sometimes	Moderately Often	Often	All of the Time

1. Why can't I be perfect	0	1	2	3	4
2. I need to do better	0	1	2	3	4
3. I should be perfect	0	1	2	3	4
4. I should never make the same mistake twice	0	1	2	3	4
5. I've go to keep working on my goals	0	1	2	3	4
6. I have to be the best	0	1	2	3	4
7. I should be doing more	0	1	2	3	4
8. I can't stand to make mistakes	0	1	2	3	4
9. I have to work hard all the time	0	1	2	3	4
10. No matter how much I do, it's never enough	0	1	2	3	4
11. People expect me to be perfect	0	1	2	3	4
12. I must be efficient at all times	0	1	2	3	4
13. My goals are very high	0	1	2	3	4
14. I can always do better, even if things are almost perfect	0	1	2	3	4
15. I expect to be perfect	0	1	2	3	4

16. Why can't things be perfect?	0	1	2	3	4
17. My work has to be superior	0	1	2	3	4
18. It would be great if everything in my life was perfect	0	1	2	3	4
19. My work should be flawless	0	1	2	3	4
20. Things are seldom ideal	0	1	2	3	4
21. How well am I doing?	0	1	2	3	4
22. I can't do this perfectly	0	1	2	3	4
23. I certainly have high standards	0	1	2	3	4
24. Maybe I should lower my goals	0	1	2	3	4
25. I am too much of a perfectionist	0	1	2	3	4

Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977).

Below is a list of some of the ways you may have felt or behaved. Please indicate how often

You've felt this way during the past week. Respond to all items. Place a check mark in the appropriate column.

During the Past Week…	Rarely or none of the time (less than 1 day)	Some or a little of the time (1-2 days)	Occasionally or a moderate amount of time (3-4 days)	All of the time (5-7 days)
1. I was bothered by things that usually don't bother me	□	□	□	□
2. I did not feel like eating; my appetite was poor	□	□	□	□
3. I felt that I could not shake off the blues even with help from my family	□	□	□	□
4. I felt that I was just as good as other people	□	□	□	□
5. I had trouble keeping my mind on what I was doing	□	□	□	□
6. I felt depressed	□	□	□	□
7. I felt that everything I did was an effort	□	□	□	□
8. I felt hopeful about the future	□	□	□	□
9. I thought my life had been a failure	□	□	□	□
10. I was fearful	□	□	□	□
11. My sleep was restless	□	□	□	□
12. I was happy	□	□	□	□
13. I talked less than usual	□	□	□	□

14. I felt lonely	□	□	□	□
15. People were unfriendly	□	□	□	□
16. I enjoyed life	□	□	□	□
17. I had crying spells	□	□	□	□
18. I felt sad	□	□	□	□
19. I felt that people disliked me	□	□	□	□
20. I could not "get going"	□	□	□	□

Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001).

er the *last 2 weeks*, how often have you been thered by any of the following problems?	**Not at all**	**everal Days**	**ore than Half the Days**	**early every day**
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself – or that you are a failure or have let yourself or family down	0	1	2	3

7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite 0 being so fidgety or restless that you have moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or hurting yourself in some way	0	1	2	3
+ ___ + _ = ___ **TOTAL SCORE**				

If you checked off *any* problems, how *difficult* have these problems made it for you to take care of things at home, or get along with other people?

Not difficult at all Somewhat difficult Very Difficult Extremely Difficult

Fitzpatrick Skin Types Scale and Scoring:

What are the colour of your eyes? (1-5)	1. Light blue or green, grey 2. Blue, green, grey 3. Dark blue or green, light brown (hazel) 4. Dark Brown 5. Brownish-black
What is the colour of your hair (naturally and before aging)? (1-5)	1. Red 2. Blonde 3. Chestnut or dark blonde 4. Dark brown 5. Black

What is the colour of your skin (unexposed area)? (1-5)	1. Pink 2. Very pale 3. Light brown or olive 4. Brown 5. Dark Brown
Do you have freckles on unexposed areas? (1-5)	1. Many 2. Several 3. Few 4. Rare 5. None
What happens to your skin if you stay in the sun for extended periods? (1-5)	1. Severe burns, blistering, peeling 2. Moderate burns, blistering, peeling 3. Burns sometimes followed by peeling 4. Rare burns 5. No burns
Do you turn brown after sun exposure? (1-5)	1. Hardly or not 2. Light tan 3. Medium tan 4. Dark tan 5. Verk dark tan
Is your face sensitive to the sun? (1-5)	1. Very sensitive 2. Sensitive 3. Mildly sensitive 4. Resistant 5. Very resistant
How often do you tan? (1-5)	1. Never 2. Rarely 3. Sometimes 4. Often 5. Always
When did you last expose your skin to the sun or artificial tanning sources (tanning beds)? (1-5)	1. More than three months ago 2. In the last 2-3 months 3. In the last 1-2 months 4. In the last week 5. In the last day

- **Type I** (scores 0–6) always burns, never tans (palest; freckles).
- **Type II** (scores 7–13) usually burns, tans minimally
- **Type III** (scores 14–20) sometimes mild burn, tans uniformly
- **Type IV** (scores 21–27) burns minimally, always tans well (moderate brown)
- **Type V** (scores 28–34) very rarely burns, tans very easily (dark brown)
- **Type VI** (scores 35–36) Never burns (deeply pigmented dark brown to darkest brown)

Qualification of Mindfulness Meditation - 17-Item Self-Report <Likert-scale 1-5>

(1) Right now the surface of my skin feels warm
(2) I was completely immersed in what I was doing
(3) Right now I feel relaxed
(4) Right now I feel more sensitive to body sensations
(5) My sense of self was diminished for an extended period of time (less self-conscientious as a result)
(6) Right now I feel safe
(7) Right now I feel a heightened sense of control
(8) This mediation session felt long
(9) Right now I feel content with myself
(10) This meditation session felt effortless
(11) Right now I feel energized
(12) Right now my body (i.e. muscles) feels less tense
(13) The meditation session was intrinsically regarding in and of itself
(14) During the meditation session my skills were equivalent to the challenge at hand
(15) Throughout the meditation sessions, I did not worry about others
(16) I would practice formal meditation again

We are interested in your experiences while doing the breath attention task. Please read each item and indicate the response that best represents how frequently you experienced it during the mindful attention task. Please use the following 5-point scale when rating each item in the provided boxes.

Mind-Wandering Inventory (39-Questions)

① – Never
② – Rarely
③ – Sometimes
④ – Often
⑤ – Very often

During the mindful attention task …

1. I worried about things that might happen in the future

2. I thought I should be doing better on the mindful attention task

3. I was aware of outside noises

4. I was aware of my breath

5. I thought about the future

6. I thought about pleasant things that happened in the past

7. I was aware of pain in my body

8. I felt bored

9. I thought about pleasant things that might happen in the future

10. I had daydreams or fantasies involving others

11. I thought about pain in my body

12. I thought about solutions to my problems

13. I thought about unpleasant bodily sensations

14. I had conversations with my "inner voice"

15. I felt sleepy

16. I imagined having conversations with others
17. I thought about using my phone

18. I was aware of pleasant body sensations

19. I experienced images of colors/shapes or other non-verbal experiences

20. I was simultaneously aware of my breath and another experience(s)
21. I thought about how difficult it was to attend to my breath
22. I thought about things that I need to do
23. I was aware of unpleasant feelings/emotions
24. I wished I was doing something other than the mindful attention task
25. I thought about past experiences
26. I noticed how awake I felt
27. I worried about things that happened in the past
28. I was aware of pleasant thoughts/emotions
29. I had daydreams or fantasies involving me
30. I thought about my performance on the mindful attention task
31. I counted to myself
32. I thought about using my computer or phone
33. I was aware of experiences that I cannot label or describe
34. I thought about new ideas
35. I noticed how attentive I was
36. I "heard" music in my head
37. I thought about how easy it was to do the mindful attention task

Please list any other experiences you had during the mindful attention task