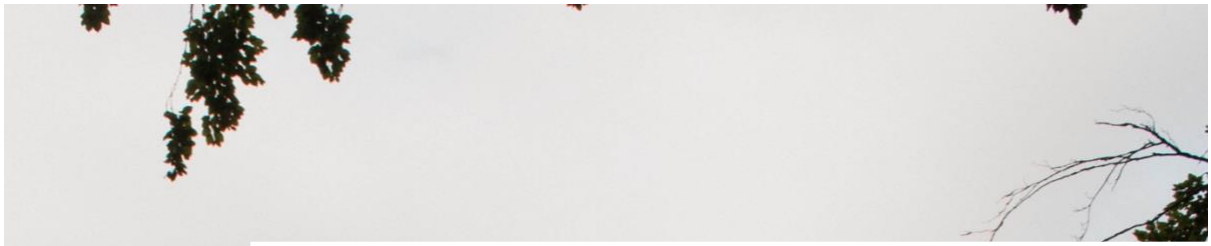

Can Consumer Wearables Provide Information for Stress Measurements?

Emilia Törnell

DIVISION OF ERGONOMICS AND AEROSOL TECHNOLOGY | DEPARTMENT OF DESIGN SCIENCES
FACULTY OF ENGINEERING LTH | LUND UNIVERSITY
2019

MASTER THESIS

Can Consumer Wearables Provide Information for Stress Measurements?

Emilia Törnell

June 12, 2019

LUND
UNIVERSITY

Can Consumer Wearables Provide Information for Stress Measurements?

Published by
Department of Design Sciences
Faculty of Engineering LTH, Lund University
P.O. Box 118, SE-221 00 Lund, Sweden

Subject: Interaction Design (MAMM01)
Division: Division of ergonomics and aerosol technology
Supervisor: Johanna Persson
Co-supervisor: Rikard Lundstedt
Examiner: Mattias Wallergård

Abstract

Psychiatric diagnoses are growing as the most common reason for people in Sweden to leave in sick from work, and within this category, stress-related health problems is the largest. At the same time, a market of wearables is exploding with new ways to keep track of health. What if these consumer products could be used in the work of preventing stress? This study is an investigation in the potential use of two different commercial smart watches for recording stress levels on individual and organizational basis. The test products have gone through three substudies, to see what they can measure, how accurate they are and what experiences they provide the user, and thus find out if they can serve as stress trackers or not. It was found that both devices managed to track changes in heart rate when a person sitting still was exposed to stress. Secondly, a link was found between the amount of deep sleep and a stress value that was presented by the Garmin watch. Single stress events presented in a diagram on the mobile application could also be correlated with moments of high self-perceived stress. A user experience test resulted in overall poor scores for the test products, although Fitbit was significantly better than Garmin in attractiveness and stimulation. Overall, the watches do provide approved accuracy in measuring heart rate, but it takes more than that to track stress in an effective way. It is crucial to also have a feature that makes calculations between heart rate and physical activity. This led to the conclusion that wearables with integrated stress features can be a useful tool to monitor stress, as long as the information is presented in an appropriate way for the user - a combination of technology and good design.

Key words: Wearables; Heart rate; Accuracy; Feature; Effective; User experience.

Sammanfattning

Psykiska diagnoser ökar och är den vanligaste orsaken till sjukskrivning i Sverige. Inom denna kategori är stress den mest förekommande. Samtidigt finns en växande marknad av wearables med nya sätt att hålla koll på hälsan. Tänk om dessa kunde komma till användning vid förebyggandet av stress. Denna studie är en undersökning av två kommersiella smarta klockor och deras potential att användas som stressmätare, både på individuell och organisatorisk nivå. Testprodukterna har genomgått tre delstudier där mätnoggrannhet, utbudet av funktioner och användarupplevelser har utvärderats. Resultatet visade att klockorna lyckades mäta signifikanta skillnader i puls när en stillasittande person blev utsatt för stress. Det hittades också en länk mellan mängden djupsömn och en stressfunktion som fanns att tillgå i Garmin-klockan. Enskilda tidpunkter med högt stressvärde kunde också kopplas till stunder då den självupplevda stressen var hög. Ett användarupplevelsetest resulterade i tämligen svaga poäng för de båda testprodukterna, men i två av kategorierna, attractiveness och stimulation, visade det sig emellertid att Fitbit-klockan var signifikant bättre än Garmin-klockan. Sammanfattningsvis får klockorna godkänt i noggrannhet vid mätning av hjärtfrekvens, men det krävs mer än så för att spåra och mäta stress på ett effektivt sätt. För att göra det krävs också en funktion som gör beräkningar mellan hjärtrytm och fysisk aktivitet. Detta ledde till slutsatsen att smarta klockor med en inbyggd stressfunktion kan vara ett hjälpsamt verktyg vid stressmätning, så länge informationen presenteras på ett lämpligt sätt för användaren - en kombination mellan teknik och bra design.

Acknowledgements

I would like to thank Johanna Persson for guidance and great support throughout the whole work.

A big thanks also to all test participants who agreed to go through the stress test och wear the watches for a day in order to help the research.

Lastly, I would also like to thank Rikard Lundstedt for providing introduction in the VR lab and test setup and Peter Jönsson for guidance in the data processing programs.

Contents

List of Abbreviations

Abbreviation	**Meaning**
CSV	Comma-separated Values
HPA	Hypothalamic-pituitary-adrenal
HRV	Heart Rate Variability
KSQ	Karolinska Sleepiness Questionnaire
KSS	Karolinska Sleepiness Scale
TSST	Trier Social Stress Test
VR	Virual Reality
UX	User Experience

1 Introduction

In Sweden, the increase of psychiatric diagnoses increased with 129 % from 2011 to 2017 [1]. Today they stand for 48 % of the sick leave and around half of these are due to stress reactions [2], [3]. At the same time, today's market of wearables claiming not only to count people's steps, but to keep track of their overall health including physical activity and stress is growing more than ever. People want objective measurements of how well their health is, and since stress is something that we do not always notice at the time, it could be useful having a tool telling us when we cross the limit of what is good for us. Our own estimation might not be good enough to state how stressed we are. Still, one should know that the testing performed on these wearables is limited, thus we do not know yet if they are reliable enough to be used in formal contexts [4].

This study is an investigation in the potential use of two different activity bracelets for recording stress levels on individual and organizational basis. If stress can be monitored through a simple heart rate recording wearable, correlations between certain situations and environments and people's state of health could be identified and investigated. This could be useful when evaluating workplaces and could serve as a precious tool in stress preventive work. But can simple activity bracelets deliver accurate data and how is it presented for the user? Also, what experiences and opportunities do they serve? And from a user experience (UX) perspective, are they easy and effective to use?

It is also worth questioning if heart rate as the only marker is enough to achieve a complete image of a person's stress level, since it is far from the only parameter that gets affected by it. There are also other factors than stress that can result in changes in heart rate, like for example physical exercise or attraction [5], [6]. In a practical example, how do one know that a certain change in heart rate is caused by stress? Imagine on a workplace, where one task might be situated on second floor, meaning when people go there their heart rate increase due to walking in stairs. If there is a difference in the increase depending on whether the reason is physical activity or mental stress, it might be easy to separate the two types of changes, but if not, will there be a problem when trying to identify particular stressful situations within an organization.

1.1 Purpose

The aim of the study is to reach understanding of potential use of activity bracelets, with a certain focus on stress. Further, the goal is to bring us one step closer to the introduction of simple and concrete techniques to create safer and more convenient workplaces.

1.2 Research questions

The questions which the three parts of the study are suppose to provide answers to are the following:

1. Substudy one
    - In general, how accurate is the heart rate recording sensor in commercial activity bracelets?
    - How the accuracy vary with pulse?
2. Substudy two
    - How well do the test products operate in real-life situations?
    - What features do they have (and how can they be useful from a stress measuring point of view)?
3. Substudy three
    - What user experience do the test products provide?
4. Overall
    - Are the devices effective from a stress measuring point of view?

1.3 Limitations

The project presented in this report prolonged for about six months and with the three substudies performed, not so much time was put on each one of them, meaning the number of test products and test participants had to be kept low. A low budget added to this made it even more difficult to freely spend money on more test products and test participants.

1.4 Report structure

The next chapter, Background, will provide a brief introduction to the subjects stress and design theory, as well as information about the test products. From that, the continuation of the report will present the overall working process and the three substudies in detail, including methods and results. After that comes the chapter Analysis, where observations in the results are interpreted, followed by Discussion and Conclusions.

2 Background

2.1 Stress - what is it and how can we measure it?

When a person is exposed to a situation where something is perceived as a threat or danger or if there is a lack of control, the body goes into a state preparing itself for either attack or running away, known as the fight or flight response [8]. There is also a third reaction called freeze, which means that the individual goes into a state of being unable to move [9]. What happens inside the body when a human is exposed to stress, both physiological and physical, is that the sympathetic nervous system is activated, making us prepared to deal with the external threat [10]. This causes acute secretion of the stress hormones adrenaline and norepinephrine, causing increase in heart rate, redirection of blood flow and sweating [11]. The liver releases energy in shape of glucose, blood flows out to skeletal muscles and brain, while the blood flow towards skin and gastrointestinal tract is decreased [12]. There is also an activation of the immune system as well as the hypothalamic-pituitary-adrenal (HPA) axis. Stimulation of the HPA axis causes secretion of cortisol [10].

The stress mechanism and thus our ability to quickly adjust to different situations has historically saved humans from life-, and other mammals, from danger and is one crucial factor to why we have become such a successful species [10]. However a stress reaction takes much energy and, as a consequence of that, rest and recovery must follow. If not, the stress reactions turns from something positive to affecting us in a rather negative way. The high levels of cortisol will lead to increased signals of hunger and ability to store more energy, meaning an increased risk of obesity [10]. Other common effects from long-term negative stress are insomnia and depression as well as deterioration in performance and immune response [15]. Today, the stress related harm on mental health is a large problem on a societal level and from both an individual and organizational viewpoint [16]. Previous studies have shown that one thing that can reduce the negative consequences of stress is workout [13], [14]. These findings are not to be confused with lack of physical stress response, but could be of interest when comparing the self-perceived stress and the one physiologically measured.

2.1.1 Inducing stress

Since stress affects all people, sometimes causing mental and physical harm, it is also a common subject of interest in research. The Trier Social Stress Test (TSST) is a tool developed to understand physiological changes due to a stressful situation constructed in a laboratory environment. In this study, the TSST became interesting because it could help to see what the test products show in a situation where the use is stressed. In the test, the participants are asked to hold a speech and perform mental arithmetic in front of an audience. This stressful situation has been proved to cause changes in terms of increased stress hormone levels as well as a higher heart rate [17]. In 2011, a virtual

version of the TSST was developed at Lund University [18]. This does not require actors or a physical environment. Therefore, the virtual version was considered more suitable for this specific study, performed by one person.

2.1.2 Measuring stress

The objective way of measuring stress, by for example recording heart rate and levels of stress hormones, is one way to go, but there are other alternatives as well. One could also use methods based of self-perceived mental condition by asking people to fill in forms. That would be more of a subjective method. An example, developed by the organization *Sunt arbetsliv* (Healthy working life) is the stress test form called "Stress och balans", which directly translated means stress and balance [19]. It consists of nine self assessment questions regarding ability to concentrate, recovery, sleep, memory, etc. If this could be done at workplaces and on an organizational level, which it already has, it is possible that it would be a certainly effective way of identifying stressful situations and keep track of people's mental health. However, the point with monitoring stress using activity bracelets is that if it really works, it does not take anything from the person in question. They would not have to evaluate how they are feeling, but instead that information could be found from concrete, accurate measurements of the physiological process of stress.

2.2 Test products

The two consumer products tested in this study are Garmin Vivosport and Fitbit Charge 3, see figure 1a and 1b. The first one keeps track on heart rate and also claims to provide all-day stress tracking. The heart rate sensor has a sampling frequency of 1 Hz and is optical, making use of a technique called photoplethysmography (PPG) where LED light is emitted to the body and measures the degree of reflection which reveals amount of blood in the vessels [20],[21]. The Fitbit Charge 3 continuously measures heart rate and also this product's sensor is based on the optical principle. In Fitbit's own product specification it is mentioned that the sampling frequency differs depending on physiology, location of device and movements, but exactly how much these factors affect the measurement does not appear [22]. However, from a request to Fitbit's support center, it was found that the Fitbit watch used in this study, over a seven day period, sampled at a frequency from 0,1 to 0,2 Hz. The sampling frequency is not referring to the blood volume but actual values of heart rate frequency. How often the blood volume is sampled does not appear, and questions regarding that was not answered from the companies, but to get an accurate value of the heart rate it has to be, according to the Nyqvist sampling theorem, at least the double of the heart rate [23].

(a) Fitbit Charge 3. **(b)** Garmin Vivosport.

Figure 1: The two test products, with start screens that are both settable.

The heart rate sampling frequency of the test products can indeed be considered low, namely up to 5000 times lower than the ECG (electrocardiogram) recording Bittum Faros 360 which was used as reference system in substudy one. This fact has previously proven to affect the accuracy when Fitbit was tested in a study where the participants got to ride a bike for ten minutes while their heart rate was recorded [24]. However the limit of what is considered high enough sampling frequency is not all clear. Another study showed that a decrease from 5 kHz to 50 Hz did not lead to considerable errors in heart rate variability (HRV) data, meaning that the sampling frequency rate required for approved result might sometimes be lower than expected [25]. Still it should be kept in mind that the test products versus the reference system measures different things - heart rate depending on blood volume in the vessels and ECG based on potentials respectively. To gather accurate heart rate information the demands on sampling frequency does not have to be the same as for ECG.

In both test products, recorded pulse data is presented in an app and data from both devices can also be exported respectively, the Fitbit data as CSV (comma-seperated values) files and the Garmin in a format called fit. However, in the virtual TSST, what is analyzed is an average of the heart rate over every five or ten minutes [18]. Therefore, the data of interest even in this study is such an average to be compared to the result from the reference. To obtain a five minute average, the data exportation is not needed, but instead this data is found in the apps associated to the test products respectively.

There are three ways to read data from the Garmin vivosport - directly from the watch, on the mobile application and from a user account available on the website Garmin connect. The Garmin vivosport watch has a feature presenting

the current level of stress. On the user account there is also a presentation of the overall stress level of the day. Also the Fitbit Charge 3 provides information through the watch itself, the mobile application and from the user account on their website. In total there is one feature regarding the subject stress. This is found in the watch and is called Relax, offering guided breathing sessions, based on the user's heart rate, for two or five minutes. On their own website they express that the goal with this feature is to make it easier to manage stress, referring to research which have shown that guided breathing can be a way to lower stress levels [26].

The motivation for the two products chosen to be included in the study is based on four arguments. First, they both have a reasonable price of about 1500 SEK which makes this study applicable for a wide range of potential users. Secondly, they have a slim design which should make them comfortable to wear all day. The third argument is the fact that they both use the same heart rate measuring technology, which makes it interesting to see how much they differ, not only from the reference system device, but also from each other. Lastly, since one of the products states that it has a stress feature while the other one does not, it is interesting to see if the stress feature works well and also of it is possible to manually measure stress without the feature.

2.3 Design theory

To ensure a product will succeed in making users satisfied, the design of it has to be well thought out and that from a user perspective. If not, it does not matter how cool and high tech features there is, if no one understands how to utilize it. The design of a product includes both hardware and software. The hardware should be *e.g.* attractive, especially when it is something visual, like for example a fitness watch on your arm. Other aspects like ergonomics and durability as well as range of features are also important. When it comes to software, this has to have a logic system for how to navigate and also the visual appearance of it is important. This is often referred to as interaction design, which means that the product is in line with how people interact in their everyday lives, and thus provide user experiences that improve and simplifies peoples' way of living their lives [27].

In design it is important to know that people are different and thus that all potential users of a product are not the same. Therefore, the product should not be designed as if it was for one specific person, because then most people will be disappointed. Instead, one idea is to do activity-centered design [28]. People can be different from each other, but with one specific product they might perform the same activities and are similar in that sense. This means that through keeping the activity and not the individual person in focus, it is possible to include more people as potential satisfied users.

Yet is to find out whether the test products' design provide good user experience

or not. The concept user experience covers questions about how a product behaves and how people feel about using them [27]. In the specific case of activity bracelets it might be that they suit well for someone who is personally interested in fitness, sleep and stress tracking. These people might have similar expectations and approach to activity bracelets. But if the products are to be used on workplaces, the group of users might include a wider range of opinions and standpoints, which put even higher demands on an inclusive way to design.

3 Process

Answering the research questions arisen in the introduction would deliver a complete picture of how the test products work in particular and how similar products could be used in general, which also could mean better health at workplaces, in society and for single individuals. In order to do that, the two products of two different brands have been evaluated in three different aspects, including heart rate data accuracy, range of features and user experience, through examination in three substudies, with three different subjects of interest, figure 2. First, a laboratory test where the heart rate measuring accuracy was determined in a situation were the participants were exposed to social stress in a test. The data was evaluated by comparison with a reference system, the ECG recording Bittium Faros 360. The second substudy was a real-life test where I tested the products myself for one week, where the purpose was to find out what features are available and how well the data from the devices matches how the user is feeling. The third and last substudy was an examination of the user experience.

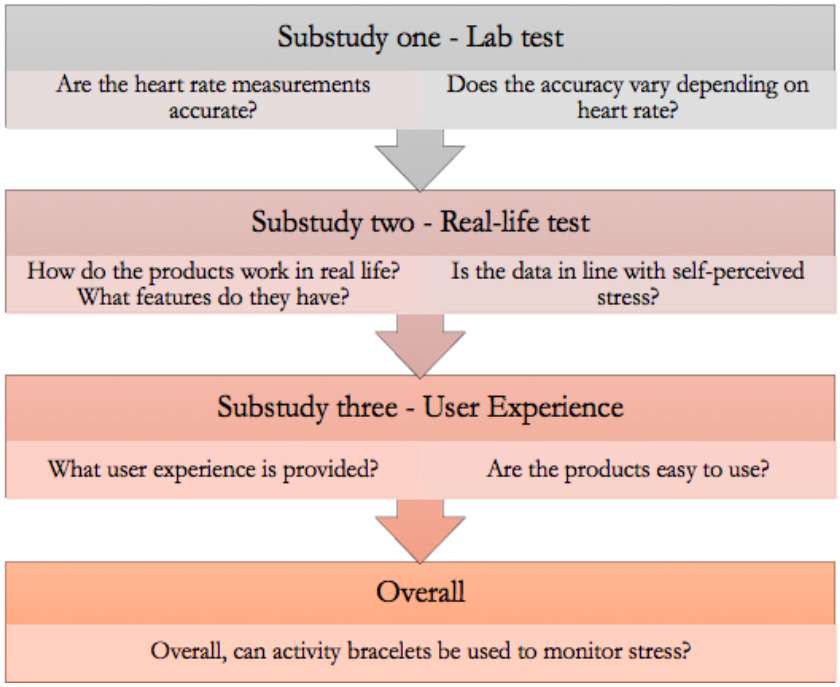

Figure 2: An illustration of the working process throughout the whole study.

In the first part of the study, a virtual reality version of the Trier Social Stress Test, developed at Lund University was used to evaluate the technical performance of activity bracelets compared to a reference system. The virtual test has

gone through a pilot study which showed significant effects on heart rate and T-wave amplitude, where T-wave represents the ventricle repolarization [18]. By making use of a stress test that has proven to induce stress in humans, the testing of the bracelets becomes concrete since potential changes in heart rate can be correlated to the stress. The parameter of interest in this study was narrowed down to solely heart rate. Previous studies have also shown that a large HRV means that the person suffer less from stress than a person with a low HRV [29]. This is because a large degree of variation means the body is smoothly adapting to the external circumstances. However, in the pilot test of the virtual TSST, no significant changes in HRV were noticed and therefore the parameter considered most important when testing the test products' accuracy is the mean heart rate, although the percentual change from one average to another will also be analyzed [18].

Another aspect of interest, which is investigated in the second part of the study, is how activity bracelets work practically in people's everyday life and what experiences are brought to the users. If activity bracelets are to be used to record and monitor stress levels at organization level, they have to fit any potential user, not only people who themselves have a big interest in technology. In this study the two test products went through a real-life test where several aspects were brought to the table. These aspects have been of interest in previous studies and in this project questions will be obtained from such that are already formulated in questionnaires and forms used for research at Karolinska institute.

How well the devices perform technically and what the user can have them for are both important factors to a successful product, but in addition to that the user experience that is provided is just as crucial to be able to apply them in real practical life. Therefore, this is what the third substudy is about.

4 Substudy one - Lab test

4.1 Method

4.1.1 Material

Reference measurement
The device used as reference system was the ECG recording Bittium Faros 360, figure 3 which is the device that has been used in previous studies that included the virtual TSST. The device records ECG with a sampling frequency of 1000 Hz, which is a common frequency for ECG measurements and have been considered optimal for accurate HRV analysis [30]. This method as reference is considered truthful since it, unlike the test products, captures every single heartbeat through recording potentials.

Figure 3: Bittium Faros 360, which measures ECG with a sampling frequency of 1 kHz.

4.1.2 Procedure

The test procedure was intended to be same as the one performed in the pilot study of the virtual Trier Social Stress Test [18]. A brief test protocol was provided from PhD student Rikard Lundstedt who have worked with the test several times, see Appendix A. This came to serve as a base, while the details were explained orally. The stress part consists of two tasks. First, the test person is intended to hold a five minute speech about themselves in front of a committee, with the purpose of trying to convince them and get a job they imagine they had applied for. After the speech they are given a new task which is to count down from 1687 down to zero in steps of 13 as fast and accurately as possible. Every time they failed, they had to start all over again. After another five minutes, the stress test is over and the person gets five minutes of rest.

As the person arrived the test was initialized with a period of about five minutes where the they were informed about the upcoming procedure and got to settle down and get familiar with the gear. However, the test participants were only informed about the speech and not the arithmetic part. This was to make sure it would come as a surprise and thus even further increase the stress level. One

test product on each wrist and the reference system under the chest. The VR glasses were put on to present the test environment, which included a waiting room and another room with a committee consisting of three people. After the introduction the test person had five minutes of rest without the VR glasses on. This period is later referred to as the baseline or base. After the initial rest the test person had five minutes to prepare for their speech, followed by five minutes dedicated to the speech itself, five minutes for the arithmetic task and five minutes of subsequent rest. The whole procedure resulted in five times five minute periods, see figure 4.

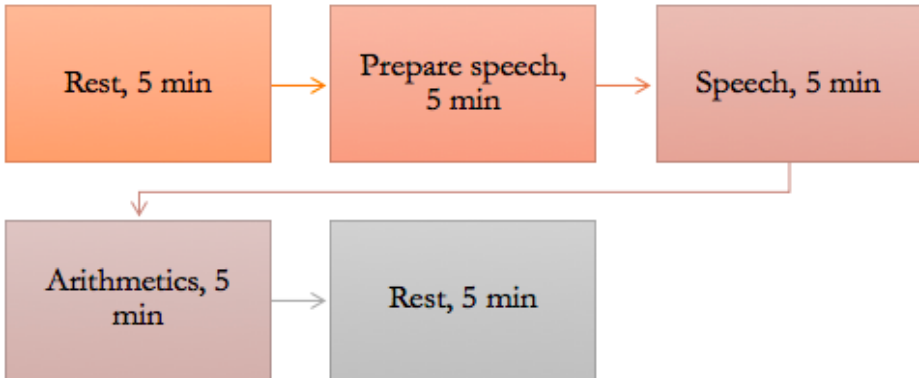

Figure 4: The whole VR TSST procedure, including five times five minute blocks.

There is a vital difference in aim between this study and the previous use of the virtual TSST. In the pilot test study, the aim was to answer the question about whether the test could induce any stress or not, while in this project, it is already known that the test person will be stressed and instead the interest lies in finding out if the two test products provide the same result as a reliable ECG device when it comes to stressful environments. This main difference led to exclusion of the original prohibition of caffeine, tobacco or food intake two hours before the test [31].

4.1.3 Test participants

The practical testing included eleven participants, ten females and one male, in ages 23 to 27 years old. During one of the tests the Garmin watch did not record data during the three middle five-minute phases in the the test, but only during the first five minutes and the last five minutes, so this test was excluded from the result. Thus the result regarding the technical performance is based on the remaining ten tests, including eight females and one male, ages 23 to 27 years old, (mean = 24 years and SD = 1,4 years).

Test participants where gathered through an announce poster. In order to

make sure that enough test persons would attend, the aim was to gather a quantitative group of test people, rather than a qualitative one, and therefore there were no requirements in who that could apply. The importance in this study was to investigate in whether measurements of stress are even possible to perform using activity bracelets. After this initial step, in further studies, comes higher demands on a more selectively composition of testers. However, the composition of the testing group will still be discussed when looking over the results and somehow the lack of a varying composition of test people could mean that the result is not applicable in any situation.

4.1.4 Data treating

From the test products, data in shape of five times five minute heart rate averages were entailed directly from the user accounts which are accessible in foreseeable versions in the mobile applications and more detailed on both of their websites respectively. The Fitbit user account provides heart rate averages over every five minutes, while the Garmin makes averages after every two minutes. The Faros 360 device provides more detailed information since it measures potentials and present an ECG, where every single heartbeat is counted. By treating the ECG, through digital high-pass filtering and adjustment of the peak detection amplitude, using the software LabChart, disruptions in the signals could be eliminated from all ten records and a truthful representation of the heart beats was obtained. From that, five times five minute averages where produced for comparison with the test products.

The heart data match between the different devices was visualized, using the software Matlab, by plotting the mean averages from the three devices next to each other in on diagram and by calculating and plotting the mean square error (MSE) for the whole procedure. The square errors were also separately analyzed within each five-minute-period of the test, to see where the test products had their best and worst match towards the reference. Also, the significance of the products' deviations, as well as the relative changes within each product, was determined through a z-test and based on an assumption of a normal distribution for each five-minute period respectively. In line with previous use of the virtual TSST the result data was also analyzed through comparison of heart rate from baseline each of the four following periods. This was to see not only how close to the absolute heart rate the test products showed, but also how well the relative changes matches the ones presented by the reference device.

4.2 Results

From each test procedure a dataset of five times five-minute heart rate averages were obtained from all three (Fitbit Charge 3, Garmin vivosport and reference system) devices respectively and these are plotted in figure 5, where the mean increase or decrease in heart rate from baseline to each of the four following test periods differ between the three devices, see table 1. In every period, with the

exception of Fitbit during the preparation phase, the relative changes for the test products were smaller than the ones regarding the reference.

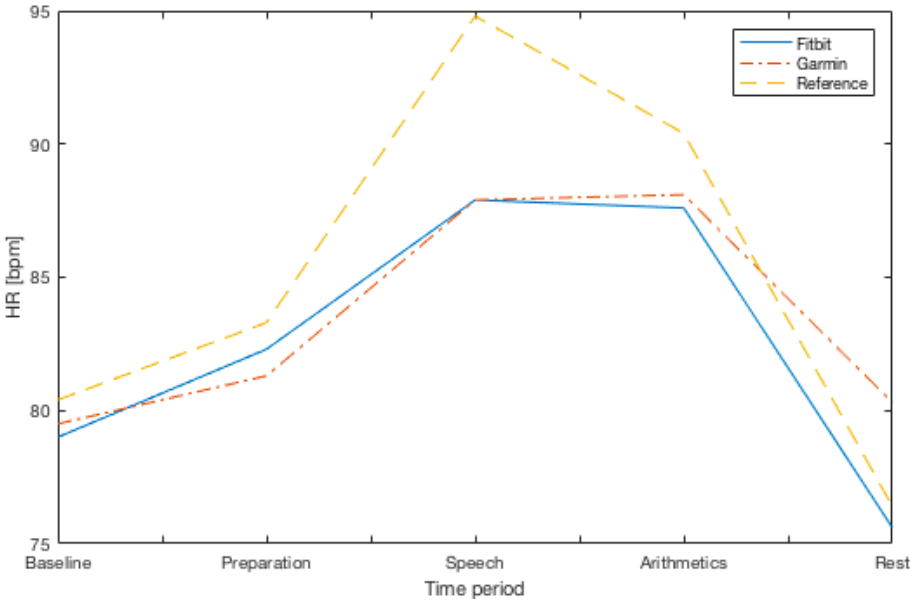

Figure 5: Mean heart rate during the whole test procedure according to the three measurements.

Table 1: The relative increase or decrease in heart rate from each device's baseline respectively.

Device	**Baseline**	**Prepare**	**Speech**	**Arithmetic**	**Rest**
Fitbit	+/- 0	+4.2 %	+11.3 %	+10.9 %	- 4.3 %
Garmin	+/- 0	+2.3 %	+10.6 %	+10.8 %	+1.0 %
Reference	+/- 0	+3.6 %	+17.9 %	+12.4 %	- 5.0 %

The heart rate means presented by three products show that the accuracy of the test products, in this study, deviated from the reference measurement to a certain degree, see figure 5. The error is the largest during the period of the speech, where the highest mean heart rate is found, while seemingly small during the periods with lower heart rate. According to the z-test, the deviations from the reference system are significant to a degree of 95 % confidence, see z-test in equation 1, meaning one can be 95 % sure of that both test products are incapable of showing the right heart rate value after an increase like the one occurred in the stress test. However, also the z-tests for changes within each device shows that the increase in heart rate from baseline to speech also are significant, with a confidence level of 98 % for both the Garmin and the Fitbit watch, see quantiles in equations 2 and 3.

$$\frac{meanFitbitSpeech - meanRefSpeech}{std(RefSpeech)/\sqrt{10}} =$$
$$\frac{meanGarminSpeech - meanRefSpeech}{std(RefSeech)/\sqrt{10}} =$$
$$\frac{87.9 - 94.8}{13.05/\sqrt{10}} = -1.6715 \tag{1}$$

$$\frac{meanFitbitSpeech - meanFitbitBase}{std(FitbitBase)/\sqrt{10}} = 1.98 \tag{2}$$

$$\frac{meanGarminSpeech - meanGarminBase}{std(GarminBase)/\sqrt{10}} = 1.97 \tag{3}$$

Further, the mean squared errors are plotted and presented in figure 6, showing that even though the two test products had the same mean heart rate during the speech period, the Garmin was closer to the reference system in terms of MSE.

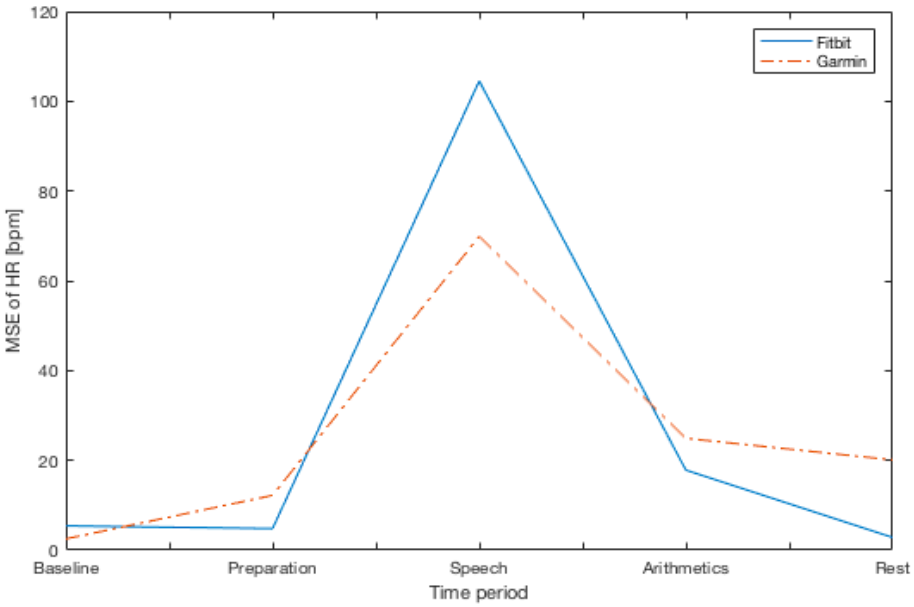

Figure 6: MSE of heart rate data from Fitbit and Garmin assuming the reference system as correct.

5 Substudy two - Real-life test

5.1 Method

Testing the products' technical performance under certain arranged circumstances is one important part of the evaluation of the products. However, information about how they operate in real-life situations and environments is just as essential. To get an idea about the experience for a potential user of the products I decided myself to test both test products in a real-life test consisting of two parts - long-term and detailed. The point was to see both what kind of information that could be gathered from the devices, as well as to see how well that information matched with questions regarding self-perceived stress, sleep, level of activity and mental health. In the first, long-term real-life test, which prolonged for one week, the goal was to draw parallels between the overall self-perceived stress level and data from the test products throughout the whole day. The purpose was not to identify specific moments of stress, which instead is the focus in the later, detailed real-life test, which only lasted for one day.

5.1.1 Test period one - long term

In the first real-life test, the devices were worn 24 hours per day during the whole week, even when sleeping. The week was picked so that as many different kind of days were caught. During the test week I also tried to vary the amount of hours of sleep, which resulted in variations in the time when going to bed and waking up. I wanted to see if this could affect how I felt the day after.

Every of the seven days in test period one, two different forms were filled in, a short free text diary was formulated and the information available from the products was noted. This was done because of the interest of finding out whether it is possible to see correlations between daily heart rate data (including mean heart rate, heart rate graphs and the stress pile available only on the Garmin vivosport), perceived mental and physical health, stress level and sleep and from that find out if the products could deliver consequent information about it or not.

In order to make sure the test would result in useful data the real-life test including question forms was put through a pilot test for one day. From that I realized that one of the questions should be modified to suit the test better. The question was how physically active one was feeling at the moment, but since the question probably never would be asked right in the middle of a workout, it was changed to ask how overall physically active one had felt since the last moment of form filling. This would help to see if parallels could be drawn between a large amount of physical movement and level of stress.

The form filling was performed every morning, midday and evening. The morning form consisted of questions selected from KSQ (Karolinska Sleepiness Ques-

tionnaire), which concerns parameters that are also possible to read from one or both of the test products [32]. The reason to include the KSQ is that sleep is one term strongly associated with stress, where high levels of social stress will result in impaired sleep and in particular a reduction of the deep sleep [33]. The midday and evening form includes Stress-Energi-formuläret, which can be translated to Stress Energy Questionnaire, Karolinska Sleepiness Scale (KSS) and BORG-CR 10 [34], [35], [36]. All forms are found in appendix, both in the Swedish, original version and translated to English. The data obtained from each session of form filling was then compiled to get a general assessment of the whole day. In order to avoid the self assessment being affected from the data from the test products, the test products data was not gathered in connection with the form filling but after the whole test period was finished.

Since the main focus is to draw conclusions between perceived stress and what the devices say the number of questions from the KSQ was reduced from 17 to seven. From the SEF and BORG-CR 10, all questions regarding stress were included. The compilation are presented in appendix B. The exact time for the midday check varied in time, which is a result of the aim to obtain a wide range of physical and mental states to see how well the devices operate depending on the conditions. The test diaries are attached in Appendix B, also including all three forms.

5.1.2 Test period two - detailed

In the second real-life test, which was performed during one day, no forms where used, but instead the point was to write as much detailed free text about mental state as possible, with the aim to in a precise way link certain events to data from the watches. The test was deliberately put on a predominantly calm day, this to make sure there was enough time to keep track on the wearables and make notes every now and then. The day consisted of some traveling by car, meeting with new people, doing some studies and taking it easy in the afternoon. In between these events were some short walks. The afternoon was spent at home and included no physical activity or stressful situations.

Every time when I felt nervous or stressed, I wrote this down and noticed the time. This happened two times during the day, first in connection with borrowing a car and then when meeting new people where I had to introduce myself and make an effort to be alert and polite. Apart from these two events there was nothing that really stressed me up.

5.1.3 Data from products

Data from the test products was automatically saved in the user accounts to be obtained from the mobile applications or websites respectively. Sleep data was obtained from both applications respectively, which included information about numbers of hours of sleep and how much of every kind of sleep states, *i.e.* deep,

light, REM and awake.

In both the Garmin app, which is called Connect, and the Fitbit app, curves presenting the heart rate of the day are shown if clicking on the pulse icon on the home screen. The Garmin watch also has a stress pile with values from 0 to 100 and these where noticed everyday after the evening form filling. Values 0-26 means that the user has had an overall resting day, 26-50, 51-75 and 76-100 are results of low, middle and high stress respectively.

Since none of the days in neither the long-term, nor the detailed test period were very stressful, more than a few minutes here and there, I decided to go back to the time when the lab tests where performed and thereby obtain stress Garmin vivosport, when the person wearing it was actually in a very stressful situation. Through that, concrete parallels could be drawn between physical stress reactions and the corresponding data from the test products. This was to investigate in the stress feature which is present in Garmin vivosport, meaning that no data was from Fitbit Charge 3 was included.

5.1.4 Other factors affecting heart rate

As well as social and mental stress, physical exercise is another known cause to increased heart rate [5]. So if a person is active during their workday or in their everyday life it is of interest to see if that would cause an increase in stress. Sometimes, another explanation to increased heart rate is physical attraction and love. This happens due to a stress reaction which helps a person to be alert, pay attention and later on attach to the other person and because of that it is rather seen as something positive and important than something that affects us negatively [6]. These additional factors affecting heart rate were taking into account when analyzing and discussing the results.

5.2 Results

5.2.1 Test period one - long term

The long-term test period prolonged from Tuesday February 12^{th} to Monday February 18^{th} 2019 and the general descriptions of the contents is as follows: Day one included hours of physical activity in the form of alpine skiing followed by a couple of hours studying. In the end of the day the self-perceived stress level was low, which is usual for me when being able to have that combination of both physical activity and mental focus. Too much of either usually makes me feel unbalanced. However, the skiing was quite demanding and included some hikes with high exposed terrain where a fall could be devastating. The day after was similar to the first one, but with a little less intense and heavy physical activity. In the evening my body was sore from the two days and I felt like falling asleep at 7 pm. Day number three consisted of eight hours of traveling by bus, airplane and train. This meant many hours of physical inactivity, but

however also requirements to be on time for every departure, which made me feel stressed at some points. The fourth day was marked by grief and other strong emotions since it included a funeral. No physical activity was done throughout the day and the mental state varied from sad to gratefulness. Afterwards, on the memorial I had to mingle with distant relatives which at some points made me nervous and a bit stressed. The day after the funeral, one of the watches, Garmin vivosport, ran out of battery and since i did not have the charger with me it could not record data for five and a half hours. Therefore this day was excluded from the test period and instead the test got to prolong for one day extra. Day number five included studying and a long walk. Nothing special happened. After that, on the sixth day I was on skis again and felt like there was a good combination between studies and physical activity. In the evening I went out with friends and did not go to bed until three in the morning. On the seventh and last day of the real-life test, came a calm day consisting of walking in town, doing some shopping before lunch and then afternoon studies, with a break for a half hour walk.

Self-perceived stress and sleep

The test period was picked so that many different kinds of days would be included. Some days where active while others not. Also the amount of sleep differed to a certain degree. However, the state of health did not vary as much as planned and in the end I had never answered "stressed" on the question of how I was feeling. Below is a summary of the answers from the form filling that was done three times a day during the whole week. Table 2 presents the result from the morning form, where hours of sleep refers to the time from going to bed to waking up in the morning. Sleep quality is a combination between questions about deep or light sleep, restless sleep or not and how good the overall sleep was. Table 3 presents a summary of the form filled in at midday and in the evening, which shows that I never felt really stressed during the test period. Neither was I on any day tired during daytime. From table 2 one can see that none of the nights were perceived as bad sleep and in all mornings but one I was feeling alert or OK.

Table 2: Conclusion of morning form filling from each of the seven days of the test.

Day	Hours of sleep	Sleep quality	Present feeling
1	8:33	Moderate	Tired
2	8:53	Good	OK
3	7:50	Good	Alert
4	8:32	Very good	Alert
5	9:45	Good	Alert
6	7:58	Good	Alert
7	5:00	Very good	OK

Table 3: Conclusion of midday and evening form filling from each of the seven days of the test.

Day	Mental state	Activeness	Sleepy or alert
1	Relaxed	Very high	Neither
2	Sharp	High	Neither
3	Calm	Very low	Neither
4	Sharp	Very low	Alert
5	Sharp	Low	Alert
6	Energized	Moderate	Alert
7	Relaxed	Low	Neither

Test products

The watches' methods to measure sleep resulted in that Fitbit Charge 3 on every night presented less hours of sleep than Garmin vivosport, while the amount of deep sleep in six out of the seven days was the highest according to the Fitbit, see table 4 för Garmin data and 5 for Fitbit data.

Table 4: Summary of sleep and stress data from Garmin vivosport.

Day	Hours of sleep	Deep sleep	Stress (1-100)
1	8:02	0:49	25
2	8:40	0:28	36
3	7:47	0:56	21
4	8:40	1:07	17
5	9:32	0:41	24
6	9:06	0:57	16
7	5:27	0:51	24

Table 5: Summary of sleep data from Fitbit Charge 3

Day	Hours of sleep	Deep sleep
1	7:24	1:13
2	8:00	1:05
3	6:45	0:45
4	6:52	1:33
5	8:32	1:28
6	7:59	1:44
7	4:50	1:10

Comparison

The day with the largest value on the Garmin stress pile (36) was day number two, which was a day with a high level of physical activity. According to the Garmin watch this was also the night with the least deep sleep, and from the Fitbit results it was the night with the second least, see figure 7.

The two days with the lowest values on the stress pile were day four and six with stress values of 17 and 16 respectively. These were also the days with, according to both test products, the most amount of deep sleep throughout the test period, see tables 4 and 5 and the visualization in figure 7. On both days I felt alert when I woke up, but it was not the nights I had graded with the best sleep quality in the morning form, see table 2. During the days I felt sharp one day and energized on the other.

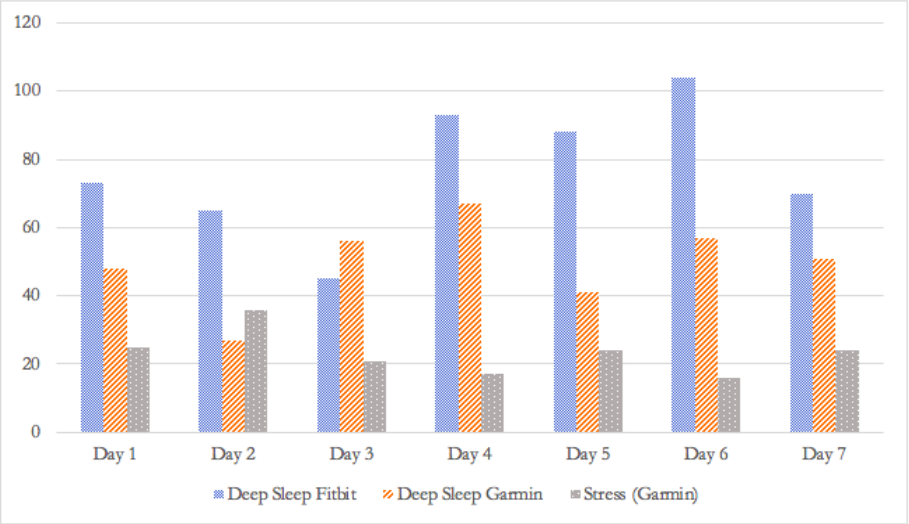

Figure 7: An illustration of the correlation between minutes of deep sleep and stress value according to the Garmin watch, hence mixed units on the y-axis.

In this specific test period, no connection is found between the self-reported mental state and the remaining self-report data, nor between stated mental condition and stress level or sleep data from the test products. The two days that was marked as relaxed (day 1 and 7), considerably differed from each other in activeness, hours of sleep and deep sleep and the same goes for the three days marked with sharp, where also stress level varied widely.

5.2.2 Test period two - detailed

On the day that constituted test period two, the two largest stress values were 76 and 72, see figure 8, and these could both be related to the only two self-assessed stressful situations during the day. The first of the two values occurred in at the time when me and a friend sat down in a car that we had borrowed, and I felt a bit nervous about it. The second highest value happened when introducing myself to new people I had not met before.

The two stressful situations both happened after a period of physical activity, so when looking at the pulse curves from both Garmin and Fitbit, the heart rate during the stress event, was not peaking but just about to go from high to low, see figures 9 to 12. Also the later, smaller stress reactions are difficult to correlate with the heart rate curves.

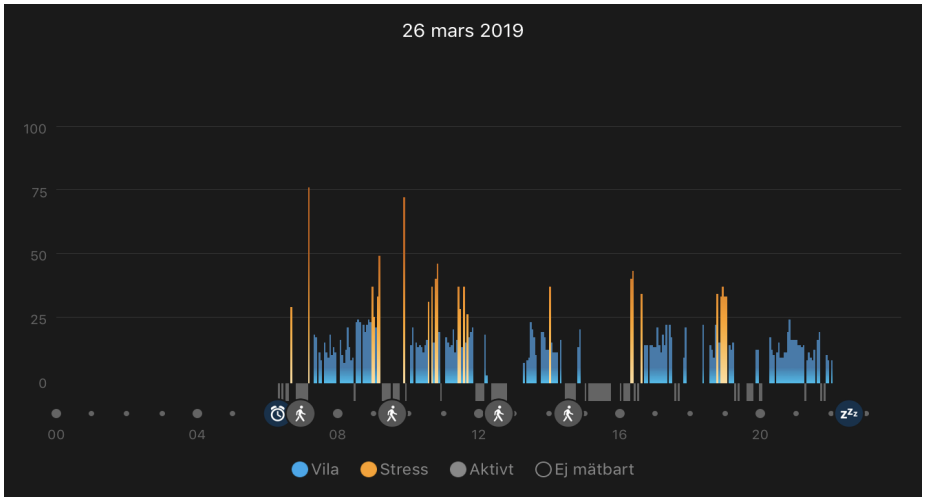

Figure 8: Stress data from the stress feature of Garmin vivosport.

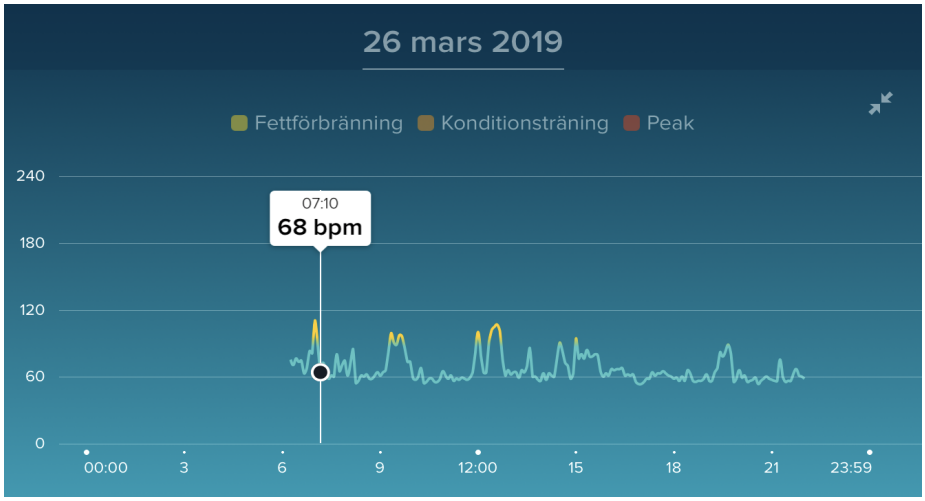

Figure 9: Pulse according to Fitbit Charge 3 at the time with the largest stress pile of the day.

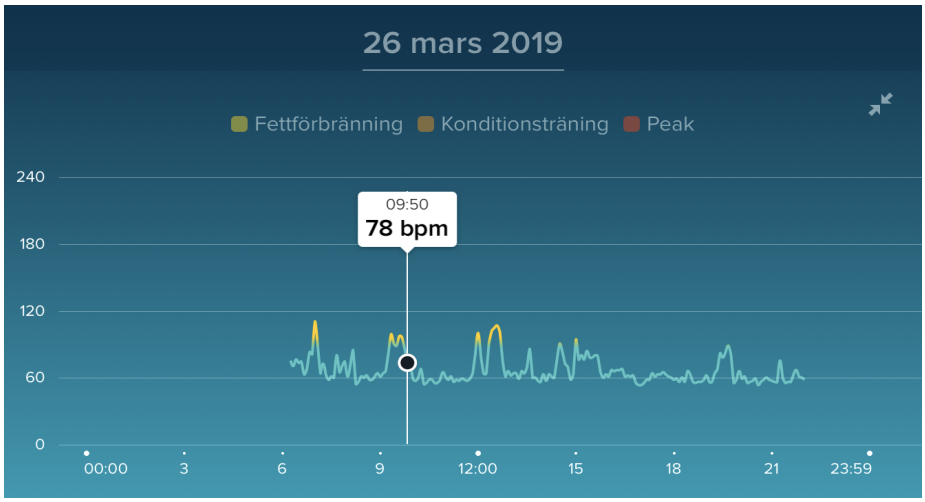

Figure 10: Pulse according to Fitbit Charge 3 at the time with the second largest stress pile of the day.

Figure 11: Pulse according to Garmin vivosport at the time with the largest stress pile of the day.

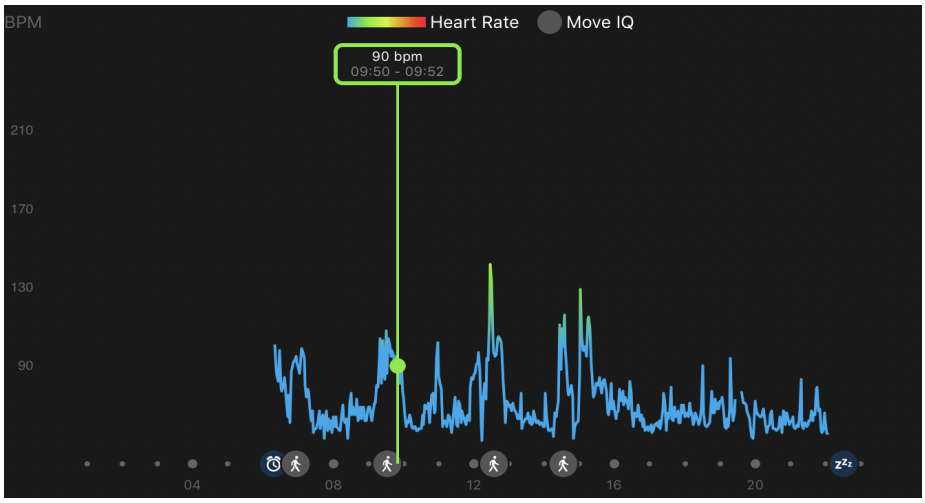

Figure 12: Pulse according to Garmin vivosport at the time with the second largest stress pile of the day.

5.2.3 Additional stress data

Stress data from Garmin vivosport, which is the one of the two test products stating it can measure stress, is presented figure 13. The data is obtained from one of the lab test days which included three tests, with three different participants. Since the stress test has proven to induce stress in the person that is going through it, the piles show that the watch is capable to identify and record stress reactions, at least when the user is sitting still.

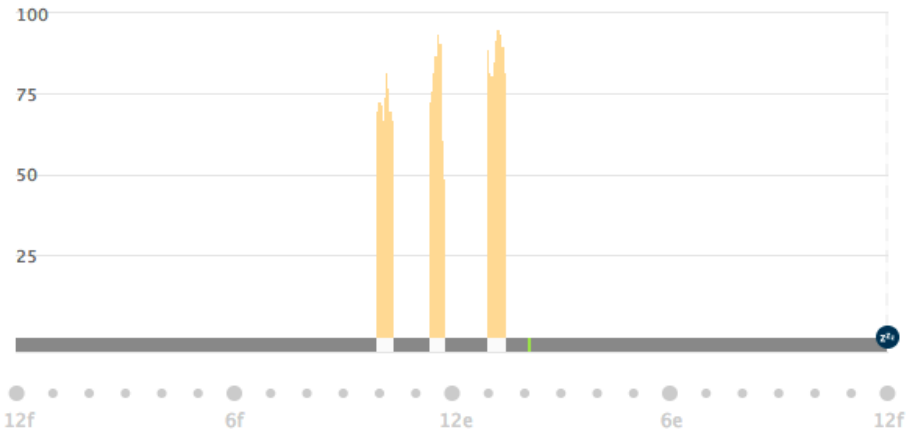

Figure 13: Stress data from one day during the lab test, including three test runs. The data is collected from the user account at Garmin connect.

6 Substudy three - User Experience

The purpose with the third and last substudy was to reach understanding of how people feel about wearbles in general and the test products in particular. If activity bracelets are to be introduced on workplaces, a brief idea of what the employees would think about it is a first step in the right direction to be able to make it happen.

6.1 Method

The user experience was evaluated through a test where the products were worn for one day by six participants. There was an aim to have at least five participants, this due to a previous statement that testing with five users is enough to reveal 85 % of a product's weaknesses, [37], which led to the conclusion that it is most likely also enough to get a fair idea of the user experience (UX) qualities. Keeping the number relatively small also gave more time for thorough analysis of the results.

6.1.1 Procedure

The watches were put on in the evening before the test participant was going to bed and they were taken of in the evening on the day after. The night was included to offer the possibility of checking out how the sleep data was presented, and if that was in line with what they had thought themselves. During the test day, the participant was told to be curious about using the watches and their related mobile applications and explore features to see if these were found useful. In that way it was easier for them to answer questions about the watches after the period was over even though they only had one day to try them out.

6.1.2 Test participants

The test participants were three females and three males in ages 21-52 years old, (mean = 28 years and SD = 11,8 years). None of the participants had attended in the previous lab test that was performed in substudy one, nor had any of the participants prior experience from using any type of smart watch. The reason to this is that there was an aim to generate a result transferable to a situation where smart watches are introduced to completely inexperienced people.

6.1.3 User Experience Questionnaire

The method that was used for examination of the user experience is called UEQ, which stands for User Experience Questionnaire, and is a well established method to evaluate a products user experience [38]. After the test period was finished, the participant was asked to answer a questionnaire, where half of the group got to start with Garmin vivosport and the other half with Fitbit Charge 3. This was to eliminate possible influence due to in what order the products

were evaluated. The questionnaire has 26 items regarding attractiveness, perspicuity, efficiency, dependability, stimulation and novelty, see items in figure 14 and categories in figure 15. Each item is graded with a score from -3 to 3. Efficiency, perspicuity and dependability are so called pragmatic qualities, which means they relate to practical considerations. Efficiency relates to how easy it is to do tasks, *i.e.* if the product is efficient to use, perspicuity is about how easy it is for the user to understand the product and lastly, dependability which has to do with whether the user feels in control or not. Stimulation and novelty are both hedonic qualities, meaning they instead handle questions about sensations and feelings that are brought to the user. Stimulation addresses issues about whether the product makes the user motivated and excited or not, while novelty is more about the degree of innovation and new thinking.

annoying	○	○	○	○	○	○	○	enjoyable
not understandable	○	○	○	○	○	○	○	understandable
creative	○	○	○	○	○	○	○	dull
easy to learn	○	○	○	○	○	○	○	difficult to learn
valuable	○	○	○	○	○	○	○	inferior
boring	○	○	○	○	○	○	○	exciting
not interesting	○	○	○	○	○	○	○	interesting
unpredictable	○	○	○	○	○	○	○	predictable
fast	○	○	○	○	○	○	○	slow
inventive	○	○	○	○	○	○	○	conventional
obstructive	○	○	○	○	○	○	○	supportive
good	○	○	○	○	○	○	○	bad
complicated	○	○	○	○	○	○	○	easy
unlikable	○	○	○	○	○	○	○	pleasing
usual	○	○	○	○	○	○	○	leading edge
unpleasant	○	○	○	○	○	○	○	pleasant
secure	○	○	○	○	○	○	○	not secure
motivating	○	○	○	○	○	○	○	demotivating
meets expectations	○	○	○	○	○	○	○	does not meet expectations
inefficient	○	○	○	○	○	○	○	efficient
clear	○	○	○	○	○	○	○	confusing
impractical	○	○	○	○	○	○	○	practical
organized	○	○	○	○	○	○	○	cluttered
attractive	○	○	○	○	○	○	○	unattractive
friendly	○	○	○	○	○	○	○	unfriendly
conservative	○	○	○	○	○	○	○	innovative

Figure 14: UEQ with 26 items to grade from -3 to 3 [40].

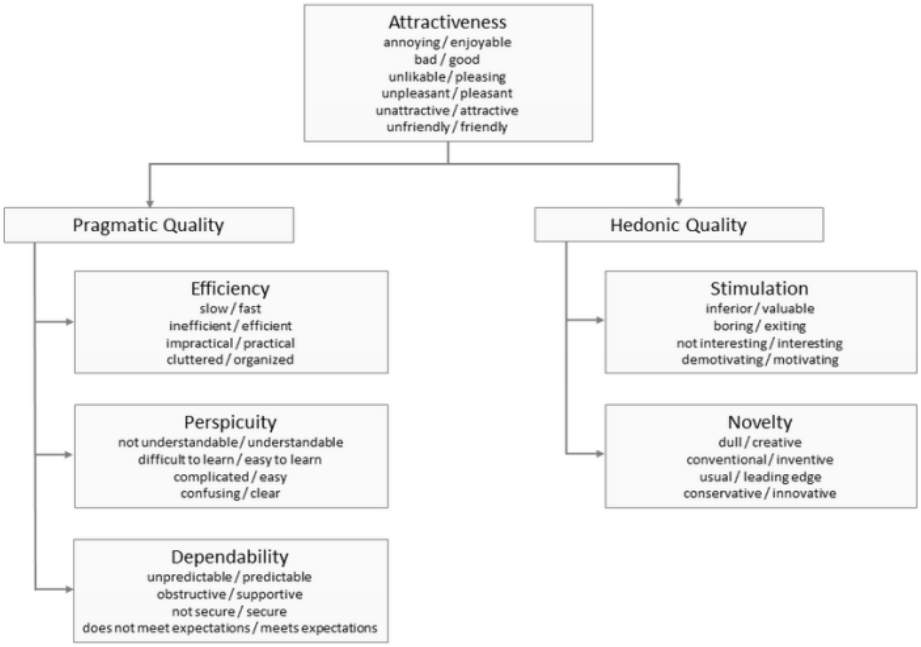

Figure 15: A map over the six factors which the items are divided into [38].

The category scores were compared between the two products and z-tests were performed to find out if one was significantly better than the other in any aspect. The average scores in each of the categories were also compared to a benchmark created from 246 tested products and 9905 test participants [39]. The benchmark consists of five grades with ranges as follows:

- Excellent: The product's score is within the best 10 %.
- Good: 10 % are better and 75 % are worse than the product.
- Above average: 25 % are better and 50 % are worse than the product.
- Below average: 50 % are better and 25 % are worse than the product.
- Bad: The product's score is within the worst 25 %.

In the questionnaire, the items are rated from minus three to three and all items belonging to one factor will contribute to the average grade of that factor. The same form was filled in twice by each test participant, one for Garmin vivosport and one for Fitbit Charge 3. The test participants were told to include both watch and its associated mobile application as one in every question.

6.2 Results

The scores of the UEQ, composed by averages from all six participants, are shown in figure 16. Both Fitbit Charge 3 and Garmin vivosport got positive averages on all six aspects. Garmin vivosport was scored higher than Fitbit Charge three in one category - efficiency. Fitbit was considered better in the remaining five categories - attractiveness, perspicuity dependability, stimulation and novelty. The highest score was Garmin's 1,4 in efficiency.

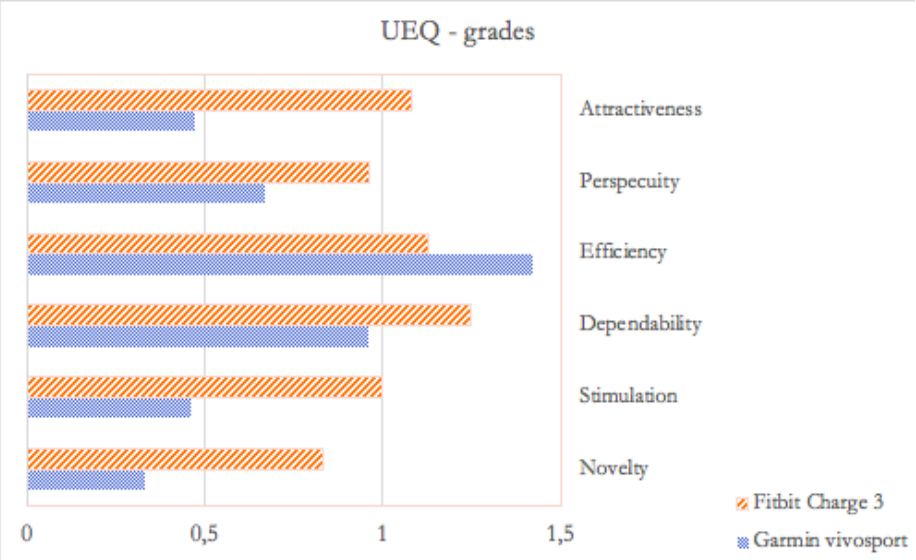

Figure 16: UEQ results, on a scale -3 to 3, showing a tendency that Garmin vivosport is better in perspectuity and efficiency, while Fitbit Charge 3 scored the best in dependability.

The differences in UEQ results are significant with a confidence level of 0.95 or higher in two categories - attractiveness and stimulation, where Fitbit was scored higher than Garmin in both. The z-test quantiles are presented in table 6 and equations (4-9) are found in Appendix D.

Table 6: Quantiles obtained from z-tests comparing the two products in all six categories.

Category	**Z-quantiles**	**Confidence level over 95 %**
Attractiveness	2.02	Yes
Perspicuity	0.62	No
Efficiency	0.71	No
Dependability	0.96	No
Stimulation	1.70	Yes
Novelty	1.02	No

According to a general benchmark for UEQ, the test products qualities are considered as shown in table 7.

Table 7: UEQ grades compared to the benchmark.

Category	**Fitbit Charge 3**	**Garmin vivosport**
Attractiveness	Below average	Bad
Perspicuity	Below average	Bad
Efficiency	Above average	Above average
Dependability	Above average	Below average
Stimulation	Below average	Bad
Novelty	Above average	Below average

After filling in UEQ, the test participants were also allowed to leave other comments and these were the following:

- "The Garmin watch was much more comfortable to wear than the Fitbit watch."
- "I liked the Fitbit watch more but the Garmin app was nicer."
- "The most annoying thing with the Garmin watch was all the reminders that I had to move. It felt almost like a constraint."
- "I liked that the Garmin watch had color screen."
- "I would rather like to have something without a screen, so I couldn't look at them all the time. I think they are stressing."

7 Analysis

7.1 Lab test

From the z-test in equation 1 and the plots in figure 5, it is obvious that the test products did not catch up with the increase in heart rate during the speech period. However, it seems like after a while, when moving on to the arithmetic phase, they actually manage to come up with a more exact result. Thus, it could mean that the test products just need a bit of extra time to read and present changes in heart rate, meaning they could fail when it comes to variations over a short period of time. However, since the increase from baseline to speech is significant to a 98 % confidence, the accuracy could still be considered enough for detecting overall stressful situations and environments, where the main focus is not on exact heart rate or percentual increase, but rather the fact that there has been a change. Also if the products were to be used on a workplace with many employees, it would be even easier to capture general stressors on a significant level.

The MSE calculations reveal that the Fitbit watch stands for the largest deviations from the reference system, and that is found in the speech phase. However, during base the test products are almost the same and in the remaining three periods (prepare, arithmetics and rest), the Fitbit's MSE is slightly lower than the Garmin's. Also when studying the percentage changes in heart rate during the whole test procedure, the Fitbit watch was in all five periods closer to the reference than the Garmin watch was. Since the aim is to detect stress levels through analysis of variation in heart rate rather than to focus on exact heart rate values, this could mean that the Fitbit Charge 3 is slightly more reliable than the Garmin vivosport. However the deviations between test products and reference are still much larger than the deviations in between the two test products, meaning there is not enough evidence to confirm that one product is better than the other one.

7.2 Real-life test

From test period one, it was found that the two nights with, according to the test products, most deep sleep resulted in the two days with the lowest stress value. Also, after the night with least deep sleep according to Garmin and second least according to Fitbit, followed the day with the highest stress value. The reason to this is not clear but from this specific test period it seems like the amount of deep sleep is important to how stressed we are the day after, independent of total amount of sleep or level of physical activity during the day.

The fact that the watches presented different results of hours of sleep, where the biggest difference was as much as 108 minutes awakes the question if the devices can even be used to in an accurate way keep track of ones sleep. However, the differences are quite similar from one night to another, which could mean,

despite the disparities, the devices are approved as long as the data only is compared within each product.

In section 5.2.2 and 5.2.3 evidence is found that the stress feature in Garmin vivosport seems to be approved for tracking stress reactions. On the detailed real-life test, there were two moments when I felt stressed and these two were also the time points with the highest stress score of the day. In the data gathered from one of the lab test days, the stress piles are high during all three tests. Even though it is taken out of its context, where the watch was worn only during the tests and not in between, it still, through categorizing the events as stress and not rest, shows some kind of ability to track stress. Also, during the long-term real-life test, there was one day with very high level of physical activity, and on that day it seems like Garmin's stress feature managed to see that the variations in heart rate did not appear from stress, but namely from being active, see Appendix C. That shows there is no problem to measure and compare stress whether the user is active or not, because the stress reactions are possible to sort out regardless of this.

7.3 User Experience

From the UEQ result bar graph it looks like Fitbit Charge 3 is generally better in UX design than Garmin vivosport. Also, it is significantly better in the two categories attractiveness and stimulation, which means it is considered more successful in terms of aesthetically appealing as well as how excited and interested the product makes the user.

When comparing the scores with the benchmark, the overall scores turned out quite low and in fact most often below average. Garmin vivosport scored above average in only one category and Fitbit Charge 3 in three out of six.

8 Discussion

8.1 The results

Heart rate accuracy
Overall, the test products prove to be accurate enough to track significant changes in heart rate when a person is sitting still. So if that is the case, it is likely that stress reactions can be tracked through analysis of the heart rate curves. Although this a limited way to measure stress even when the user is inactive, since it requires from the user to make notes when going for a coffee or to the bathroom, it is still one step closer to a solution of how to measure stress with simple technique.

What also is proven, is the test products' failure in presenting the exact right value on heart rate, so they should only be used for analysis of relative changes in heart rate and not exact values. About the measurements and the achieved accuracy, since the reference system and the test products make use of different techniques for heart rate recording, it is not fair to compare the rate of electric potentials samples by the reference system and the heart rate values produced by the test products. Since the test products only presents one heart rate value every five to ten seconds there is a possibility that quick changes in heart rate are missed out and this should be taken into account when drawing conclusions from obtained data.

Identifying stress reactions
The correlation between deep sleep and stress, discovered in substudy two, is not excessively clear in the diagram shown in figure 7, which is to a large degree because of difficulty with mixed units. However, the relationship found between deep sleep and stress is not new for this study, but in line with a previous study showing that stress has a large impact on the deep sleep [33]. Maybe this really is something that can be useful in the development of a method to measure stress. Still, if the aim is to work with stress prevention, measuring people's deep sleep is not enough information, since it does not say anything about specific events. However, measuring people's amount of deep sleep could definitely help to see trends and indications on how people are doing.

From the detailed real-life test (period two), diagrams presenting heart rate curves at the two most stressful situations throughout the day indicate that detecting stress is difficult to do by just looking at the heart rate curve. It is known that stress increases heart rate, but what if the stress reaction happens right after physical activity and the heart rate is on its way down? The heart rate peaks occurring due to physical activity were, at least in this study, much larger than the stress related ones. And even if a person is sitting still when having a stress reaction, how can one be sure that the reason is stress and not something else? As mentioned in section 5.1.5, stress is far from the only factor causing increased heart rate. Other reasons could be physical activity or

attraction to another person. Since the Garmin watch uses an algorithm where stress is calculated from a combination between heart rate and footsteps/moving around, the fact that physical activity is a factor to increased heart rate, should not be a problem. When the user is active the increase in heart rate will be defined as training rather than stress. But what if the person is stressed and physically active at the same time? And what about other factors like physical attraction and love? Today, Garmin's own way of calculating stress is not good enough to measure stress while the user is active, or to distinguish attraction and bad stress, but it still has useful qualities.

In contrast to Garmin vivosport, Fitbit Charge 3 is to a high degree inadequate and ineffective when it comes to keeping track on stress levels. Doing that would require thorough examination of the daily heart rate curve, and in combination with physical activity make a compilation to see if the person has been stressed or not. Such a method would require much time and even more importantly, it would not at all be an accurate or scientific method to use. A better method would then be to develop an algorithm similar to the one used by Garmin. Today, the only Fitbit feature regarding stress is the guided breathing sessions called Relax. This feature is meant to help one to stress down a little bit, but still it takes from the user to be aware that the stress level is high and that one needs to calm down for a bit. More optimal would be if the watch could warn the user when stress was too high, and then tell them to take a break with some guided breathing.

Since the devices are still in the initial phase of testing it is difficult to say how well we can trust them or not. Therefore, an alternative could be to use the Garmin vivosport in combination with other, more subjective, stress tests like the one called Stress and balance mentioned in chapter 3 Process. There is still no proof that different people get the same results from the Garmin stress data even if they have the same level of stress. In a first stage of using the activity bracelets for monitoring stress, every person would have to make their own matching between the two. It might be that one Garmin stress value means for one person that they are very stressed, while for someone else everything is fine, and the self-perceived stress is very low.

UEQ scores

The fact that one of the watches (Fitbit) scored significantly better in two categories (attractiveness and stimulation) sure says something, but are these the most important qualities if wearables are to be introduced in organizations? If not, which UX category is then the most important? This is a question that has not been investigated in this study, so there is no clear answer yet. Allowing me to reflect on the question, I think that if people are to accept the introduction of stress tracking wearables, it must be easy for them to understand the device and it has to make them feel secure. This would mean that perspicuity and dependability are the two most important qualities. Also efficiency, including qualities like fast, practical and organized, could be important, depending on

whether the users are to collect the data themselves or if that is supposed to be done by someone else. A high degree of novelty and stimulation might not be desirable at all, but maybe a product that gives impression of being a commonly used consumer product is better. For some users it would also be of importance that the wearble looked good, especially if they had to wear it 24 hour per day, but if it was only to be used at work, it might not be the deciding factor.

The low UEQ scores, which most often were below average according to the benchmark, was somehow disappointing and awakes the question if the products are even considerable to be used at all, or if there might be other, more successful alternatives from a UX design point of view. But is the design in these products really that poor? To remember is that all the test participants had never before used any kind of activity wearbles or fitness trackers and they also had no interest of their own to try it. On a workplace, these people are of course likely to exist, but they would probably not be the whole group. The group of people would probably be a mix including a wide range of attitudes to wearables. Therefore, the UEQ results shows an idea of what people with no interest of fitness or stress tracking, but it might not be translatable to an average workplace where all kinds of opinions and ideas is to be expected.

Other UX aspects
Several test participants mentioned that the Garmin watch was very comfortable to wear while Fitbit was slightly uncomfortable against the wrist. A comfortable wearable is of course of high importance if people are to wear the product during a long period of time. If the device is insufficient in comfort it might lead to some users wanting to take them off for a while, which in turn would mean loss of important data recordings. Another comment that ought to be discussed is the one about having a screen or not. The person felt stressed about the screen and would rather prefer something without a screen. If one was interested in the data this could be obtained from the mobile app. This should definitely be taken into account when planning for a future implementation of wearables in organizations. If the wearables themselves act as stressors, then the whole idea of using them in prevention of stress is a failure.

8.2 Methodology

The combination of methods constituting this study is for sure worth discussing. It is questionable if this was the best way to go or if it would be better to focus on solely one of the methods, on a more detailed level. Still, the idea of using smart watches as stress trackers indeed includes a wide range of issues. If each of them was to be sorted out one at a time, regardless of the other, maybe it would be difficult to reach full understanding of how all different challenges have to be solved together as one.

8.2.1 Lab test

The virtual TSST was an appropriate way to get an idea of the test products' accuracy in heart rate recording, especially since it caught moments when a person goes from calm to stressed and back to calm. About the practical parts of the lab test, since the procedure includes many tasks for the test leader, manually directing the VR jury, keeping track on time, keeping an eye on the live update of ECG, the details of the execution may have varied to a small degree from one test run to another, which might have caused different outcomes from the tests. On the other hand, since the purpose of this study is not to investigate if and to what specific degree the virtual TSST can induce stress in humans, but how well the recorded data from ordinary activity bracelets with low sample frequency match a ECG recording device that samples once every millisecond.

Another aspect to discuss when comparing the two test products is the fact that the Garmin watch during one test lost its connection and did not record any data for 15 minutes. The reason for this is not clear, but what is sure is that the watch was put on the same way as in every other test session and the test participant was sitting still during the whole test. It is possible that the interruption happened only by chance, but it is not a good thing if the watch is sensitive and easily looses connection with the skin causing data to be missed out, especially not when the person is not even moving around.

8.2.2 Real-life test

At first, the real-life test was meant to only consist of test period one, but after that was finished, it became clear that the overall evaluations did not bring any information about whether the test products managed to catch shorter events of feelings that happened during the day. Instead, the self reports reflected the general feeling throughout the day. The long-term test period was still useful though, revealing the connection between deep sleep and stress level in the following day. But it was in combination with the detailed test period that the real-life test became the most valuable. It provided an idea of how exact the features are, as well as what patterns that can be identified when looking at the data over a longer time period.

8.2.3 User Experience

The added comments from the test participants revealed some insufficiencies in the method UEQ. The questionnaire brings up overall covering questions about the interface, but misses out the physical product. This is today the most common case for UX evaluation methods, since many of them are developed for mobile applications and websites. But what is needed is an updated method for products which are rather a combination of digital and physical design, which are becoming more and more common.

Another part of the method that is worth discussing is the fact that all test

participants worn both test products during the same day, and then evaluate both of them directly after each other. There is a risk that this makes it difficult to evaluate each test product separately and not relative to each other. This could mean that if I think one product is much more *e.g.* interesting than the other, it will entail a positive grade for the product considered as the better one, and a negative grade for the worse, even if the second product maybe was not that bad after all. This possible risk was something I thought of before doing the UX test. What led to the conclusion that it probably was better to let all test participant wear both products instead of just one, is that the amount of participants was too few to be able to eliminate possible differences between them in their evaluations.

8.3 From an ethical point of view

So far, when discussing whether the test products are good enough to be used at *e.g.* workplaces, the focus has been on the watches themselves, but to remember is that the people who are involved, and the ethical issues that it implies, are even more important. This includes both the people who have participated in this study and even more importantly in a future where the watches are implemented to collect data about people within an organization.

In total, substudy one and three had 17 test participants. In many cases the participants were friends with each other, since all of them were friends of the author, and that fact made it more difficult than usual to keep everyone anonymous. Holding the test data closed and never talk about the participants (as persons or their results) with other people, and especially not with other test participants, is therefore kept important in all future ahead. There is also an ethical aspect on substudy two, where the author herself revealed detailed information, including both physical and emotional circumstances, from my days. It was sometimes difficult to decide whether an event or emotion could be shared or if it was considered too private. It was a known fact that the more details noted, the more data would be gathered to analyze and draw conclusions from.

There are also several ethical issues to discuss before a possible organizational implementation of stress tracking wearables. Are people prepared to be recorded on a daily basis? Who would handle the data and who would have access to it? If the idea of using devices like the test products was to be implemented within an organization such as a workplace, the first step would be to ask everyone involved if they would accept the company management to collect data about them. What if only some or no one would go along with it? Is it worth doing if not everyone agree? To make sure as many people as possible would agree to share their data there would need to be a well formulated plan about how to treat the data. One idea is to make a result every week and delete all individual data from the latest week. The result should then be analyzed to see trends of the average stress level over time. But how is this actually going to work? How often would the wearables be on and for how long? Would it be required

that the watch was worn even at night or could it be stationed at work so that people took them on every day when arriving? If one wants to find critical tasks or situations at a workplace, more detailed information about what the employees are doing during their day needs to be added. But this would mean that the individuals were even more exposed. What if one person in a group turns out to be more sensitive and less stress resistant than the others? Can the information be used against the individual? This would probably have a negative effect rather than positive, making people more insecure and stressed, when the main purpose was to actually create better and safer environments.

A solution to all these challenges requires its own investigation and it is important to remember that what comes out from this specific study is not a complete tool ready to be implemented, but an examination of only the watches performance in different aspects. My hope is that someone will continue working with the project, taking it into the next phase – the phase of planning and execution of implementation. The findings that have been obtained in this study make a first step, allowing the next person to move on with knowledge of that an initiative like this could be successful.

9 Conclusions

Stress is complex and it is difficult for people even themselves to know if they are stressed or not. There are many things other than negative stress that can make our heart rush. Also, the border between good and bad stress is diffuse. On a moderate level, it helps us to perform better at work, and also to be alert when we are near someone we see as a potential partner. But too much stress is dangerous and today it poses one of the biggest threats to health.

In this study it has been found that the two wearables that have been tested, both have qualities that would be valuable in a future use for prevention of stress. Although they differ a bit from the correct absolute value of heart rate, they are both capable of showing significant raises when the user is exposed to stress. It has also been found that the stress feature available in Garmin vivosport seems to be functioning and good at identifying stressful situation, meaning this product is effective as a stress tracker. Without a feature like this, either something similar has to be coded by the buyer themselves, or the ability of tracking stress is limited to solely periods where the user is sitting completely still. The user experience test resulted in overall mediocre opinions from the test participants, but it can be questioned if this result is useful, or if the deficiencies in method and execution were too crucial.

Through the three substudies, which allowed investigation of three different aspects of commercial wearables in use, a new level of understanding of how they can be used in new ways is reached. In other words, we have come one step closer to the implementation of objective measurement methods to create healthier social environments.

References

[1] Försäkringskassan. (2017). *Psykisk ohälsa bakom nästan hälften av alla pågående sjukskrivningar*. Press release 10th of October 2017.

[2] Försäkringskassan. (2019). *Pågående sjukfall efter diagnos*. Excel file available at Statistik om sjukpenning och rehabiliteringspenning, Försäkringskassan.

[3] Lidwall, U and Olsson-Bohlin, C. (2016). *Sjukskrivning för reaktioner på svår stress ökar mest*. Försäkringskassan korta analyser, vol. 2.

[4] Ghose, T. (2015). *New Trackers Claim to Measure Your Stress, But Do They Work?*. Life Science, Tech.

[5] Micus, C.R, Earnest, C.P, Blair, S.N and Church, T.S. (2009). *Heart rate and exercise intensity during training: Observations from the DREW study*. Journal of Sports Medicine, vol. 43, pp. 750-755.

[6] Seshadri, K.G. (2016). *The neuroendocrinology of love*. Indian Journal of Endocrinology and Metabolism, vol. 20, issue 4, pp. 558-563.

[7] Seshadri, K.G. (2016). *The neuroendocrinology of love*. Indian Journal of Endocrinology and Metabolism, vol. 20, issue 4, pp. 558-563.

[8] Kindermann, P. (2013). *What is stress?*. BBC Science, 19th of April 2013.

[9] Schmidt, N.B, Richey, J.A, Zvolensky, M.J and Maner, J.K. (2008). *Exploring Human Freeze Responses to a Threat Stressor*. Journal of Behaviour Therapy and Experimental Psychatry, vol. 39, issue 3, pp. 292-304.

[10] Harvard Health Publishing. (2018). *Understanding the stress response*. Harvard Health Publishing, Harvard medical school, 1st of May 2018.

[11] Nordqvist, C. (2017). *Why stress happens and how to manage it*. Medical News Today

[12] Widmaier, E.P, Raff, H, Strang, K.T. (2008). *Vander's Human Physiology, The Mechanisms of Body Function. Eleventh Edition*. McGraw-Hill, New York, USA.

[13] Childs, E and de Wit, H. (2014). *Regular exercise is associated with emotional resilience to acute stress in healthy adults*. Frontiers in Physiology, vol. 5.

[14] Clark, P.J, Amat, J, McConnell, S.O, Ghasem, P.R, Greenwood, B.N, Maier, S.F and Fleshner, M. (2015). *Running Reduces Uncontrollable Stress-Evoked Serotonin and Potentiaties Stress-Evoked Dopamine Concentratinons in the Rat Dorsal Striatum*. PLoS One, vol. 10.

[15] Bergamini, G, Mechtersheimer, J, Azzinnari, D, Sigrist, H, Buerge, M, Dallmann, R, Freije, R, Kouraki, A, Opacka-Juffry, J, Seifritz, E, Ferger, B, Suter, T and Pryce, C.R. (2018). *Chronic social stress induces peripheral and central immune activation, blunted mesolimbic dopamine function, and reduced reward-directed behaviour in mice*. Neurobiology of Stress, vol. 8, pp. 42-56.

[16] de Castro, U.R, Palha, A.J.P, de Olivera, N.R, Martins, J.C.A. (2014). *Stress in Community Health Agents: a Bioethics Protection Perspektive*. European Researcher, vol. 83, pp. 1707-1717.

[17] Kirkbaum, C, Pirke, K.M and Hellhammer, D.H. (1993). *The 'Trier Social Stress Test' - A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting*. Neurophycobiology, vol. 28, pp. 76-81.

[18] Wallergård, M, Jönsson, P, Österberg, K, Johansson, G and Karlson, B. (2011). *A Virtual Reality Version of the Trier Social Stress Test: A Pilot Study*. Precense, vol. 20 (4), pp. 325-336.

[19] Neij, J. (2019). *Mäter stressnivå för att motverka utmattning*. Sunt arbetsliv, 11th of February 2019.

[20] How the Optical Heart Rate Sensor Works on Wearable Device. Garmin, Support Center.
https://support.garmin.com/en-US/?faq=CFrxExLN717PAWGd0bnRn5
Retrieved 6th of November 2018.

[21] Allen, J. (2007). *Photoplethysmography and its application in clinical physiological measurement*. Physiological Measurement, vol. 28(3), pp. 1-39 (?)

[22] Fitbit Charge 3 Advanced Health and Fitness Tracker. Fitbit Products.
https://www.fitbit.com/eu/shop/charge3
Retrieved 6th of November 2018.

[23] Weik M.H. (2000). *Nyquist theorem. In: Computer Science and Communications Dictionary*. Springer, Boston, MA

[24] Benedetto, S, Caldato, C, Bazzan, E, Greenwood, D.C, Pensabene, V and Actis, P. (2018). *Assessment of the Fitbit Charge 2 for monitoring heart rate*. PLoS One, vol. 13(2).

[25] Mahdiani, S, Jeyhani, V, Peltokangas, M and Vehkaoja, A. (2015). *Is 50 Hz high enough ECG sampling frequency for accurate HRV analysis*. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[26] Harvard Medical School. (2015). *Relaxation techniques: Breath control helps quell errant stress response*. Harvard Health Publishing, January 2015, updated 13th of April 2018.

[27] J, Preece, H, Sharp and Y, Rogers. (2015). *Interaction design - beyond human-computer interaction. 4th edition.* John Wiley & Sons Ltd. Chichester, United Kingdom.

[28] Norman, D. (2013) *The design of everyday things. Revised & expanded edition.* Basic Books. New York, United States of America.

[29] Lischke, A, Jacksteit, R, Mau-Moeller, A, Pahnke, R, Hamm, A.O and Weippert, M. (2018). *Heart rate variability is associated with psychosocial stress in distinct social domains.* Journal of Psychosomatic Research, vol. 106, pp. 56-61.

[30] Hejjel, L and Rooth, E. (2004). *What is the adequate sampling interval of ECG signal for heart rate variability analysis in time domain?* Physiological Measurements, vol. 25, 1405-1411.

[31] Jönsson, P, Wallergård, M, Österberg, K, Hansen, Å.M, Johansson, G and Karlson, B. (2010). *Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the Trier Social Stress Test: A pilot study.* Psychoneuroendocrinology, vol. 35, pp. 1397-1403.

[32] Ingre et al., 2000; Kecklund and Åkerstedt, 1992. (1992). *Karolinska Sleepiness Questionnaire.* Karolinska institutet, Stockholm, Sweden.

[33] Åkerstedt, T. (2006). *Psychosocial stress and impaired sleep.* Scand J Work Environ Health, vol. 32, pp. 493-501.

[34] Hadzibajramovic, E, Ahlborg Jr, G, Grimby-Ekman, A and Lundgren-Nilsson, Å. (2015). *Internal construct validity of the stress-energy questionnaire in a working population, a cohort study.* BMC Public Health, vol. 15, issue 180.

[35] Åkerstedt, T. (1990). *Karolinska Sleepiness Scale* Karolinska institutet. Stockholm, Sweden.

[36] Borg G. (1970). *Perceived exertion as an indicator of somatic stress.* Scandinavian Journal of Rehabilitation Medicine, vol. 2, pp. 92-98.

[37] Nielsen, J. (2000). *Why You Only Need to Test with 5 Users.* Nielsen Norman Group, March 2000.

[38] Laugwitz, B, Held, T and Schrepp, M. (2008). *Construction and evaluation of a user experience questionnaire.* In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.

[39] Schrepp, M, Hinderks, A and ThomaschewskiLaugwitz, J. (2017). *Construction of a Benchmark for the User Experience Questionnaire (UEQ).* International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, pp. 40-44.

[40] Questionnaire and subject map are downloaded from https://www.ueq-online.org/

Appendices

Appendix A - Lab test protocol in Swedish

Praktiska erfarenheter av TSST Man har använt en manual och ett manuskript med mycket få kommentarer för kommittén att följa. Detta motsvarar dokument tillgängligt på Clemens Kirschbaums hemsida, och som ISM har översatt. http://www.macses.ucsf.edu/Research/Allostatic/notebook/challenge.html

Del 1 - anställningsintervju

Sätt på kamera och klocka.

-Välkommen. Du kan börja. Varsågod.

Fp pratar på en stund, i snitt 1 min 15 sek. Endast en person har hittills pratat i 5 minuter. Efter en stund brukar en pinsam tystnad uppstå. Efter 20 sek sammanhängande tystnad säger mittfiguren i kommittén (de andra två säger aldrig något):

- Du har fortfarande tid kvar.

Ofta följer då en stunds prat ytterligare. Efter 20 sek tystnad igen upprepas samma mening. Om inget mer händer eller efter 10 sek ställer mittfiguren några ledande frågor. Dessa kan variera något beroende på vad för typ av jobb personen söker, men i grunden är det fyra frågor som används på lämpligt sätt:

- Varför tror du att just du är lämplig för detta arbete?
- Har du (fler) erfarenheter inom (detta) området?
- Har du sökt liknandejobb någon annanstans?
- Vad gör du om du inte får detta jobbet?
Man kan också ibland på lämpligt ställe lägga in frågan: *- Varför?*

OBS!! Det får aldrig bli mer än 20 sek tyst, sedan måste någon replik läggas ino annars blir det töjligt och för pinsamt. Om fp ställer frågor till kommittén ger man egentligen ingen speciell respons på det utan använder strängt taget någon av standardreplikerna, t ex "du har fortfarande tid kvar". Ibland frãgar fp hur mycket tid de har kvar. De får inget svar på den frägan utan endast responsen:

- Du har tid har. Jag säger till när tiden har gått.
Alternativt endast: *- Du har tid kvar.*

Då tiden (5 minuter) har gått avbryter man det hela, vilket kan bli mycket abrupt.

Del 2 - aritmetiskt test.

Det aritmetiska testet påbörjas. Repliker kring detta har jag inte, men det finns förmodligen på den aktuella hemsidan och är sparsamma liksom i del 1. Då även denna del är klar säger man helt enkelt:

- Tack, då var det klart. Varsågod att gå ut.

Appendix B - Self report forms and diary

Morning form (KSQ)

Feeling when rising from bed?

- 1, very energized
- 2
- 3, energized
- 4
- 5, neither energized, nor tired
- 6
- 7, tired, not demanding to stay awake
- 8
- 9, very tired, hard to stay awake

Was it difficult to fall asleep?

- 1, not at all
- 2, a little
- 3, moderate
- 4, pretty difficult
- 5, very difficult

When did you go to bed?

When did you wake up?

How did you sleep?

- 1, very good
- 2, good
- 3, neither good, nor bad
- 4, bad
- 5, very bad

Restless sleep?

- 1, not at all
- 2, a little
- 3, moderate
- 4, quite
- 5, very much

Did you sleep deep or light?

- 1, very light
- 2, pretty light
- 3, neither light, nor deep
- 4, pretty deep
- 5, very deep

Midday/evening form

How are you feeling right now? (SEF)

- Relaxed
- Active
- Tense
- Loose
- Stressed
- Energetic
- Ineffective
- Sharp
- Pressured
- Passive
- Calm

How energized or tired are you feeling right now? (KSS)

- 1, Very energized
- 2
- 3, Energized
- 4
- 5, Neither energized, nor tired
- 6
- 7, tired, not demanding to stay awake
- 8
- 9, very tired, demanding to stay awake

How physically exhausting have you felt today, until now? (BORG, modified)

- 0
- 0,5, Extremely little
- 1, Very little
- 2, Slightly
- 3, Moderate
- 4
- 5, Heavily
- 6
- 7, Very heavily
- 8
- 9, Extremely heavily

How mentally exhausted are you feeling?

- 1, Very fresh, on top
- 2
- 3, Fresh
- 4
- 5, Neither fresh, nor tired

- 6
- 7, Tired
- 8
- 9, Very tired, can not bother to do any work

Diary, test period one

Day 1: Intense alpine climbing and skiing. Got a bit of a nervous flash at some moments. Felt really good being outside for six hours and found it easy to focus during my study hours afterwards.
Day 2: Alpine skiing and some hiking in the morning. Afternoon was spent studying, shopping and preparing for the upcoming travel day.
Day 3: Traveled from 7:30 am to 7 pm. Everything went well but had to rush in between two flights and on that I was picked for random check of luggage. However, the stress I felt was limited to a short period of time.
Day 4: After a good night of sleep, I woke up to go with my family to the funeral of my grandfather. It was all very beautiful, and I was feeling both grateful and sad. Afterwards, there was a memorial where I was placed next to long distance relatives which I do not know so well. We asked each other what we do, and I told them about my studies and so on. That kind of mingle situations are good examples of what can make me stressed. I do not like getting too many questions and then make the effort of asking interesting questions back. The physical activity level was at minimum.
Day 5: After a long night's sleep, I had a calm day including studying and a long walk. When studying I felt sharp and alert without feeling pressed or stressed.
Day 6: Day included easy skiing. In the evening I went out with friends and in order to include a short night of sleep I tried to stay up as long as I could.
Day 7: After a short night of sleep I felt pretty tired the whole day. I took a slow morning and then went to town, which included walking and some shopping for a couple of hours. Lunch was late. In the afternoon I was stayed inside the whole time, doing studies, except from a 30 minute walk.

Diary, test period two

I woke up at 6:15 and felt a bit tired. After a short morning walk, I went to travel by car for two hours, not driving. Right when I had finished the walk and we were about to take off, around 7:10, I felt a bit nervous about since the car was borrowed from a friend, and I really did not want anything to go wrong. After arriving there was a short walk again. I arrived to my destination at around 9:50. Now I felt nervous again and the reason was that I had to introduce myself to new people that I had not met before. From 10 am to 1 pm I sat down and did studies. At 10:55 a pause to stretch my legs. Between 11:55 and 12:46 I took an easy walk, with a bit of rest in the middle of it. After

that I went by car again for about 1,5 hours. Then, around 14:15 I went to the supermarket, and then home by bus and a short walk. From 15:03 I did not really activate myself much physically or mentally. I had a calm afternoon watching series and hanging out with close friends.

Appendix C - Stress data at very high physical activity

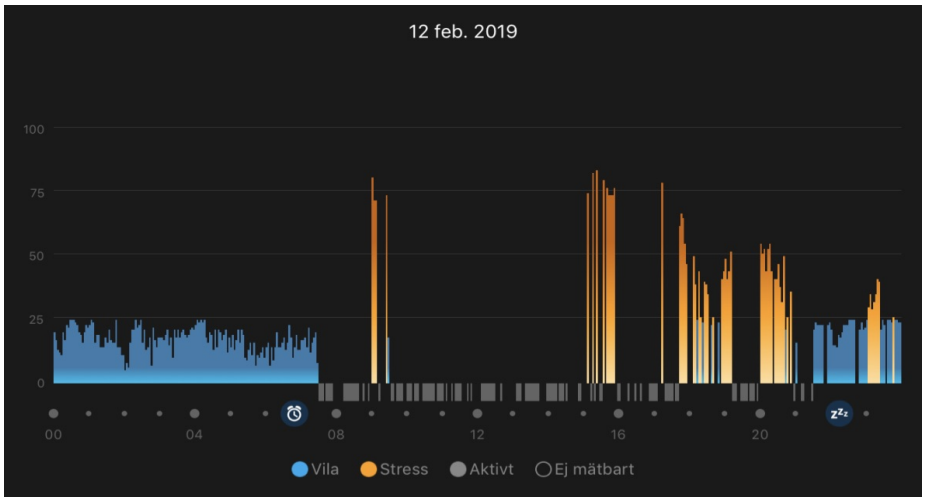

Figure 17: Stress data from Garmin vivosport from day one, which was marked as very high physical activity, during long-term real-life test.

Appendix D - Significance UEQ, z-tests

$$\frac{meanFitbitAttractive - meanGarminAttractive}{\sqrt{\frac{std(FitbitAttractive)^2}{36} + \frac{std(GarminAttractive)^2}{36}}} = 2.02 \tag{4}$$

$$\frac{meanFitbitPerspicuity - meanGarminPerspicuity}{\sqrt{\frac{std(FitbitPerspicuity)^2}{24} + \frac{std(GarminPerspicuity)^2}{24}}} = 0.62 \tag{5}$$

$$\frac{meanGarminEfficiency - meanFitbitEfficiency}{\sqrt{\frac{std(FitbitEfficiency)^2}{24} + \frac{std(GarminEfficiencye)^2}{24}}} = 0.71 \tag{6}$$

$$\frac{meanFitbitDependability - meanGarminDependability}{\sqrt{\frac{std(FitbitDependability)^2}{24} + \frac{std(GarminDependability)^2}{24}}} = 0.96 \tag{7}$$

$$\frac{meanFitbitStimulation - meanGarminStimulation}{\sqrt{\frac{std(FitbitStimulation)^2}{24} + \frac{std(GarminStimulation)^2}{24}}} = 1.56 \tag{8}$$

$$\frac{meanFitbitNovelty - meanGarminNovelty}{\sqrt{\frac{std(FitbitNovelty)^2}{24} + \frac{std(GarminNovelty)^2}{24}}} = 1.02 \tag{9}$$