Cross-instrument validity and reproducibility of heart rate measurements devices

Lisanne Zondag

S3940764

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number 45

Supervisor: Dr. Marcus Span

In collaboration with: Nienke Buist, Theres Patzelt, Harmien Tamsma, Rover Willemars.

Month 06, 2022

Abstract

This study tries to understand the consensus between modern alternative heart rate monitors and our golden standard TMSI REFA. Demonstrating if the alternative monitors are able to measure valid Heart rate (HR) and Heart rate variability (HRV), also referred to as inter-beat intervals (IBI). Wired ambulatory devices measuring with the electrocardiogram (ECG), like the TMSI REFA, are traditionally used in research settings or clinical practice. This study investigates if these wireless devices are reliable to use in these settings. We made use of the Polar H10, a chest-worn ECG monitor, and the Empatica E4, a wrist-worn watch that measures through a photoplethysmography (PPG) sensor. A sample of 28 first-year psychology students participated in an approximately 30-minute laboratory experiment. The participants were instructed to complete a variety of mental tasks (normal color Stroop task, emotional Stroop task) and physical tasks (sitting, standing). The tasks intend to elicit psychological and physical arousal, which is supposed to be detected by our devices. We found a positive correlation only between the Polar H10 and the TMSI (r = 0.94). We concluded that only the Polar H10 monitor is feasible to use in research settings since it was able to obtain the same mean IBI values as the TMSI REFA. Due to vulnerability to movement artefacts by the Empatica E4, there were too many data inconsistencies, making it unreliable to use. Further research is needed to find reproducibility between the EE4 and TMSI REFA.

Keywords: heart rate, heart rate variability, wireless monitors, Polar H10, Empatica E4

Cross-instrument validity and reproducibility of heart rate measurements devices

Are alternative heart rate monitors able to compare with the traditional gold standard ECG device TMSI REFA? Lately there are many new and affordable heart rate monitors available to measure heart rate (HR) and heart rate variability (HRV). These can also be used in the application of measuring physical-/ mental arousal, as these are said to be reflected in the HR and HRV. However, their effectiveness for research or clinical practice is still uncertain.

Traditionally, Heart rate (HR) and Heart rate variability (HRV) in laboratory settings are measured with ECG (electrocardiography), employing devices like the TMSI. Despite the accurate measurements of the TMSI, its usage is rather inconvenient as it is time-consuming and expensive (Goodie et al., 2000). Therefore, research can potentially benefit from more easy and affordable measurement devices. In support of this, there is evidence that these new wearable monitors are able to measure HR and HRV more conveniently. On the contrary, even though previous research shows promising results, there are still some limitations and differences between measurement devices. Especially, the sensitivity to motion of some devices can have large drawbacks for clinical assessment.

In the current study, we test efficient and accessible alternatives for use in laboratory/ research settings. To investigate the different monitors, we elicit arousal with stress-inducing mental-/ and physical tasks. This study makes use of a normal Stroop task, an emotional Stroop task, and physical task conditions (standing, sitting).

HR, HRV, and ECG analysis

Heart rate (HR) describes heartbeats per minute (BPM). ECG analysis determines the Heart rate variability (HRV). HRV is a series of RR-intervals, also called inter-beat intervals (IBI), that represent the time interval between each R wave, also known as R-tops (Feng et al., 2021). This means the IBI describes the time between two successive heartbeats. If HR increases, it reduces the time between two successive beats. Thus, higher HR results in lower IBIs (Shaffer & Ginsberg, 2017). HRV is considered to be a biomarker of the automatic nervous system (ANS). This incorporates the sympathetic nervous system (SNS), known as the fight or flight response, which is especially interesting for our study (Shaikh al arab et al., 2012). HR evaluates our sympathetic arousal, which is measured with our monitors. ECG signals are made up from recurrent sequences of P-, QRS-, and T-tops, within each heartbeat. As visualized in Figure 1, P-top is the first positive deviation on the ECG, caused by atrial depolarization (Mehta et al., 2009). The QRS complex is the highest amplitude in the cardiac cycle, which is due to the depolarization of the ventricles. Hence why the ECG normally starts with the detection of this QRS complex (Ledezma & Altuve, 2019). At last, there is the T-top, representing the ventricular repolarization of the heart. It is more difficult to detect P- and T-tops compared to detection of QRS complex, due to QRS having a higher signal-to-noise ratio. P-top is more likely to be absent for some ECG measures, as a consequence of low amplitudes and low signal-to-noise ratio (Mehta et al., 2009).

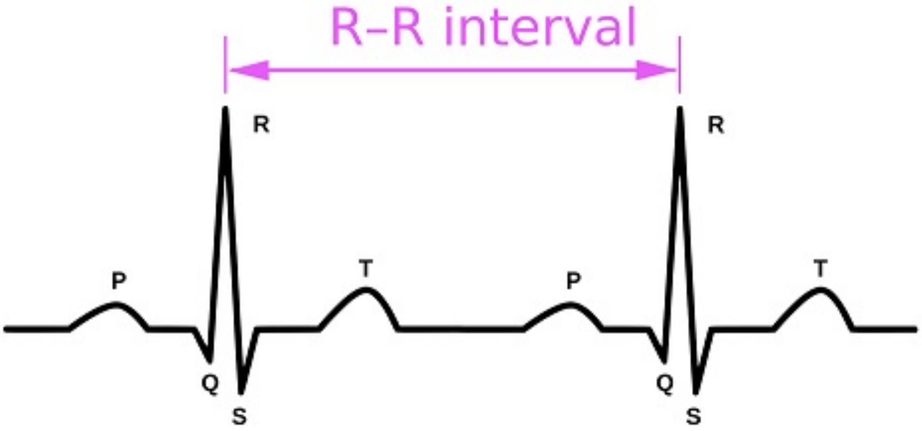


Figure 1 ECG signal showing RR-interval

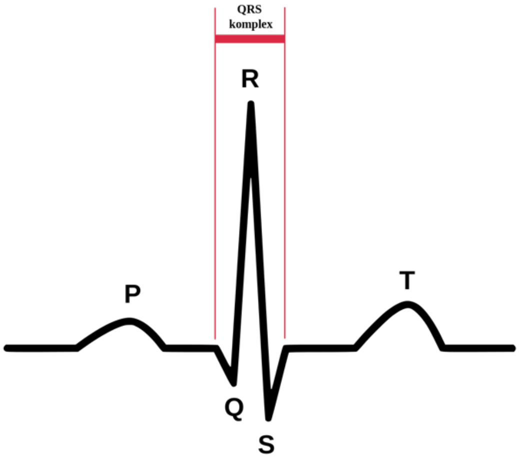


Figure 2 ECG signal showing QRS complex

Monitor

To determine the accuracy of our alternative measurement devices, their data is compared to our gold standard, TMSI REFA, measuring ECG signals. The ECG is a diagram mapping of atrium and ventricles beating, based on bioelectricity changes. Research revealed its power of distinguishing stress detection. The process that is considered stress, or subjective arousal, entails the feeling of perceived nervousness, which changes the rhythm of the heartbeat. The TMSI can detect these apparent changes in HR. Researchers mainly indicate these variations by describing RR intervals, or IBIs (Feng et al., 2021). However, for clinical research, the usage of the TMSI can have limitations. Mainly the restrictions in motion due to being wired to the device with multiple electrodes.

One alternative device our study makes use of is the Polar H10. This monitor measures HR with ECG, which is worn around the chest. This makes the usage and wearability easy for research participants. Additionally, it is rather inexpensive, which makes it accessible for various studies. Previous Research states valid measures of HRV obtained by the Polar monitor, that were comparable to the values of the TMSI. HRV relies on the measurement of the heart and its electrical activity to be able to determine IBIs (Richard et al., 2021). Research shows that the Polar H10 can monitor HR values for both mental and physical tasks. Moreover, it is advertised as being able to withstand a variety of physical activities and movements, which is especially interesting for our physical task (Goodie et al., 2000).

The other measurement device used in this study is the Empatica E4. This device is a wristband, which makes it comfortable and easy to wear for our participants. Furthermore, besides HR and HRV, the EE4 is capable of measuring skin conductance, acceleration, and temperature. Additionally, it can deliver raw data. HR and HRV are measured using Photoplethysmography (PPG). This method uses an optical technique, entailing the analyzation of blood volume pulse (BVP) variations, according to the quantity of light found in blood vessels. BVP signals are used to obtain inter-beat intervals (IBIs). A body of research shows that the EE4 is reliable for measuring HR values, both in static and dynamic conditions. However, it was less accurate for HRV values, especially under dynamic conditions including hand movements. The EE4 was unable to detect the pulse peaks in the BVP signal, which resulted in faulty data. Meaning HRV measures were only satisfactory in the absence of movement (Menghini et al., 2018).

Tasks

The Stroop task is hypothesized to elicit physiological reactions, that can be interpreted as arousal or stress. We will call it arousal for the rest of the paper. Our study uses three different components: congruent words, non-congruent words, and mixed congruent/ non-congruent words. Previous research shows that the noncongruent words elicit greater psychological arousal compared to the congruent words. Additionally, greater error rates regarding incongruent words were also evident. The reason for this is expected to be the conflict between the word and its color. For example, we expect participants to be faster in correctly determining the color green if the word is also “Green”. Hence, when the word is “Red”, participants take longer to answer or make an error. Furthermore, previous research revealed that the color-word interference of the Stroop task affected HR levels. However, the reason for heightened HR levels is still unclear. Potential reasons could be the order, where the HR is heightened in the first sequence of trials but not in the following ones. Another reason could be the larger demand for effort during incongruent color-word interference (Renaud & Blondin, 1997).

The emotional Stroop task contains three conditions: only positive/neutral words, only negative words, and both positive-negative words mixed. Participants are asked to indicate the correct color, disregarding the word and its meaning. This study is interested in the cognitive processes. Particularly, the elicited psychological arousal, together with resulting engagement in the task, in response to positive and negative emotional words. Previous research shows that the emotional Stroop task only detected interference for depressed participants. This could mean that the task might not affect neurotypical participants (Dell`Acqua et al., 2020).

The physical task consists of two conditions: sitting down and standing up. The goal is to detect a difference in HR and HRV during the resting- and movement conditions. In addition, it is highly interesting to explore if the measurement devices show valid assessments under all conditions. Previous research has shown that the new measurement devices can be sensitive to movement, resulting in missing and faulty data (Milstein & Gordon, 2020). Especially, HRV measures show high discrepancies in valid values under dynamic conditions. Furthermore, research indicates that the heart rate will increase when engaging in activities or movement. Hence, HR tends to be lower during resting conditions (Menghini et al., 2018). Other research shows promising results, that almost all devices can accurately determine HR/HRV when the participant is not engaging in movements (Milstein & Gordon, 2020).

Our study additionally made use of a Psychomotor Vigilance Task (PVT), intending to employ it as a startle response. An Alarm goes off to elicit heightened arousal in the participant. PVT is traditionally used to measure visual attention. Therefore, if the initial task is interrupted by an unpleasant noise, it disrupts the visual attention of the participants. This in turn can elicit changes in HR. The influence on HR can be estimated by HRV, which can be done by examining ECG. Previous results showed that ECG entails information on an individual´s vigilance state. Thus, HRV can predict if PVT caused changes in participants' visual attention, in addition to a heightened risk of attentional failure (Chua et al., 2012). However, the information on PVT is simply added for reproducibility purposes, its effects will not be discussed in this paper.

Present Study/ Hypotheses

Regarding the main hypothesis of our study, to investigate if modern wireless monitors are reliable alternatives for the TMSI REFA, we expect that the monitors match the performance of our golden standard device. Furthermore, concerning the normal Stroop task and the emotional Stroop task, for our second hypothesis, we expect increased HR and HRV in the non-congruent condition, as well as for the negative-emotional condition. Finally, for the third hypothesis, our study expects that HR and HRV are influenced by the physical tasks. The arousal should be identified by all monitors, however, we expect some movement artefacts in the physical task conditions by the Empatica E4.

Method

Participants

A total of 28 participants (after exclusion), with 20 females and 8 males. The participants consisted of Psychology students from the University of Groningen. The participants represented a variety of native languages (Dutch, German, English). However, all completed the experiment in English. The initial sample was 44 participants, after the exclusion of participants, the final sample yielded 28 participants. Participants were excluded when the devices didn't measure accurately or yielded abnormal data. The participants were recruited via convenience sampling, meaning first year students participate in the study to gain SONA credits. SONA is a system created by the RUG to reward students, who need to get 35 credits to pass the first year. The study was approved by the ethical committee of our university. Before engaging in the study, the participants filled out a prescreening, indicating gender, native language, etc.

Table 1

Participant characteristics after exclusion

	Gender	
	Female	*Male*
Total	20	8
%	71.43	28.57

Note.

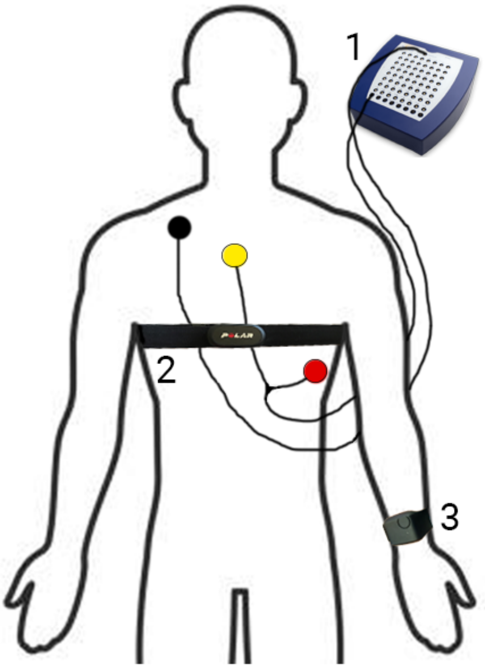
Research Design and Procedure

The study consisted of the administration of multiple mental and physical tasks with the aim to assess HR and HRV of the participants. We designed an experiment with three conditions: a normal Stroop task, an emotional Stroop task, and a physical task. The Procedure starts with describing the experiment to the participant, either orally explaining it or providing the subject with an information sheet. The participants need to sign a consent form after the procedure has been made clear and before the experiment can start. Three different heart rate measurement devices are being put on the participant's body. First the Polar band around the upper part of the torso. After that, three electrodes are placed in a line: one on the lowest rib on the left side, one in the middle of the chest above the heart, and the last one on the right collarbone. Lastly, the Empatica E4 watch is put on the wrist of choice. Once the devices are set, the participant is put in an isolated and soundproof room. The electrodes will be connected to the TMSI device, and the participant will be asked to sit down on a chair in front of the computer. Finally, the participant is left alone in the room and can begin with completing the experiment. The participants start with the physical task, which starts with the sitting condition, followed by the standing condition. This is followed by the Stroop tasks. For the Stroop tasks, the participants were instructed to indicate the correct color of the color-/ or emotional words.

Clear instructions were given on the computer before the start of every task condition, explaining what the subject was supposed to do. In addition, an alarm went off randomly each 1 to 4 minutes on a separate laptop. The alarm needs to be turned off by the subject as fast as possible by pressing the spacebar. After finishing all the tasks, which took approximately 20 minutes, the participant is granted with 2 SONA credits.

Figure 3

Presentation of device placements



Note. 1. TMSI REFA, 2. Polar H10, 3. Empatica E4

Measures

Before completing the computer tasks the participants are instructed to complete a physical task. This starts with 60 seconds of simply sitting down, followed by 60 seconds standing up.

The normal Stroop task consists of three components: congruent, non-congruent, and mixed. The first component includes four colors (red, green, yellow, blue) and the compatible color-word. Meaning, if the word red is also in the color red. For the second component the same colors were used, however with conflicting color-words. E.g., the word green is presented in the color blue.

The emotional Stroop task consisted of the same four colors, but instead with either positive or negative words. The negative-emotional words were related to war, e.g., *death, gun*, etc. The positive-emotional words are the opposite, e.g., *friendship, peace*, etc. The words were evenly distributed: with neg. words = 16 and pos. words = 16. The word list for the emotional Stroop task can be found in the Appendix B. Besides the positive-emotional and negative-emotional word conditions there is also a mixed condition. The mixed condition simply presents the participant both positive and negative words.

The order and color of the words are presented randomly, and each component presents a sequence of trials. The words are presented separately, after a response from the participant or if time has run out, the next one follows. The participants were instructed to identify the color of the word with buttons that light up in the four colors. The button box was especially built for this experiment. It's made up from 4 buttons that are positioned in a linear way, starting with red, then green, blue, and lastly yellow. The goal of the button box was to make it more entertaining for the participant, therefore increasing engagement in the tasks.

Furthermore, the task presents the participant with feedback. For a correct answer a single green dot with a message saying, "Good job!", was presented and a red dot for an incorrect answer. Moreover, the incorrect answer feedback is accompanied with a message: "FALSE!! Try harder!!", to elicit more psychological stress and make them more engaged in the task. Lastly, the Stroop task entailed a staircase feature, which decreased the maximum response time after two consecutive correct answers. Meaning the participants had less time to indicate the color of the word after each correct response. An incorrect answer increases the time again, which results in a longer valid response period

Results

We performed a Pearson correlation analysis to define the strength of the linear relationship between our golden standard TMSI REFA and the Polar H10. The Mean IBIs showed a significant positive correlation (r = 0.94, p < .001) between both devices. A correlation computation between TMSI REFA and the EE4 was not possible because the values could not be paired or matched.

Table 2

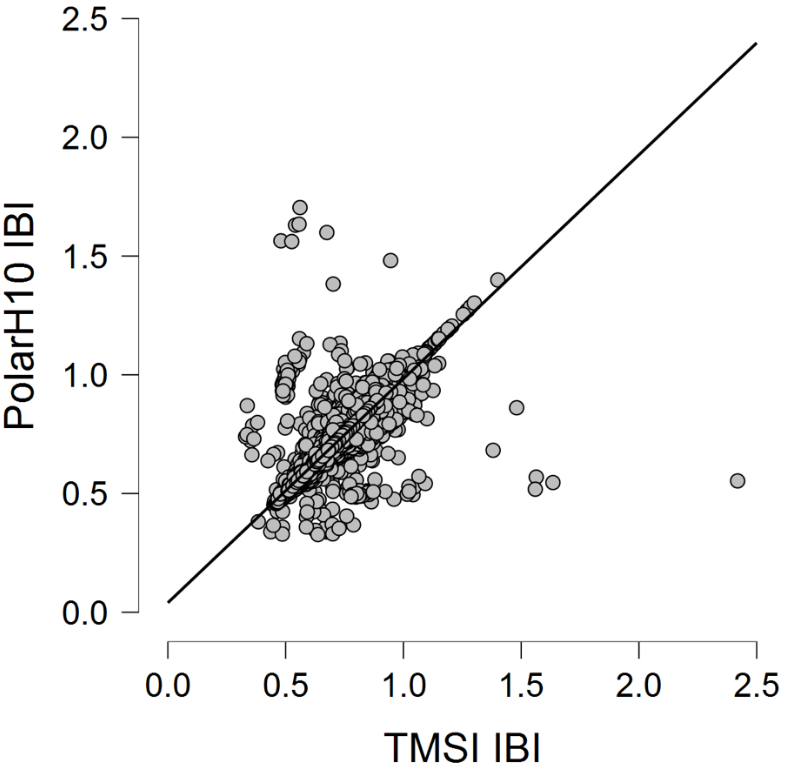
Pearson's Correlations

	Pearson's r	p-value
TMSI IBI - PolarH10 IBI	0.942	< .001

Note. Pearson´s r correlation between Polar H10 and TMSI REFA

Figure 4

Correlation between Polar H10 and TMSI REFA



We computed the Descriptive Statistics, indicating the average IBIs from all three devices (Table 3). Furthermore, we presented the average IBIs for every subject from our experiment, measured by each monitor (This can be found in Appendix A). Regarding Hypothesis 1, the Polar H10 obtained equal amounts of IBIs (TMSI = 44854, Polar = 44854), as well as the same average IBI value (M = 0.700) as the TMSI REFA (M = 0.700). The EE4 deviated from the other two devices. The Empatica obtained notably fewer IBI measures (Valid = 7758). This is only 17.30% of the data obtained by the TMSI REFA. The descriptive statistics also yielded different results for the average IBI value (M = 0.728). Moreover, we conducted a Paired Sample T-test including the Polar band and

TMSI, testing for the difference between both devices. The results were not significant (p = 0.988) with 95% confidence interval of [-0.009, 0.009]. The null hypothesis could not be rejected, meaning there is no statistical significance for an observed difference. However, based on these results we are not able to suggest, just because they are not different, that the devices have a similarity. These findings are in line with the assumption of agreement, however, only for the Polar H10 (Hypothesis 1).

Table 3

Descriptive Statistics of valid IBIs

	IBI		
	TMSI	EE4	Polar H10
Valid	44854	7758	44854
Missing	0	0	0
Mean	0.700	0.728	0.700
Std. Deviation	0.107	0.135	0.107
Minimum	0.330	0.406	0.327
Maximum	2.419	1.844	1.704

Note. All subjects combined

Table 4

Paired Samples T-Test

							95% CI for Cohen's d	
Measure 1		Measure 2	t	df	p	Cohen's d	Lower	Upper
PolarH10 IBI	-	TMSI IBI	0.015	44853	0.988	6.929e-5	-0.009	0.009

Note. t-test of average IBIs from Polar H10 and TMSI REFA

To compare the different task conditions (Hypothesis 2) we computed Descriptive Statistics that showed the average IBIs for each condition of the experiment. We were mainly interested if the non-congruent and negative-emotional words of each Stroop task showed increased arousal. For the normal Stroop task (1 = congruent words, 3 = non-congruent words; in Table 5), participants showed lower Means of IBIs (M = 0.72) in the non-congruent condition compared to the congruent condition (M = 0.74). To estimate the heart beats per minute (BPM), we used the calculation: 60/ Mean of IBIs per condition. Resulting in 83,10 BPM for the non-congruent condition and 81,30 BPM for the congruent condition. This suggested a higher HR for the non-congruent condition. The emotional Stroop task (4 = positive-emotional words, 6 = negative-emotional words; Table 5), showed no big difference in average IBIs between emotional words. The negative-emotional words showed near to equal data (M = 0.724) as the positive-emotional words (M = 0.725). This possibly indicates that the meaning of the words has no effect on the psychological arousal of the participants. These results are not completely in line with the expectations stated in Hypotheses 2, namely that higher psychological arousal would be elicited by the non-congruent/ negative-emotional words. Our analysis merely revealed a small effect (non-congruent words) or no significant variations (negative-emotional words). It can be concluded that the non-congruent words moderately elicited greater arousal than the congruent words, still, we are not able to determine if there is a significant difference between the two conditions. Ultimately, there were no changes in HR or HRV considering higher arousal for participants when they were presented with negative-emotional words.

Table 5

IBIs per task condition, all devices combined

	IBI								
	1	2	3	4	5	6	7	8	FALSE
Valid	11533	11204	11401	4034	5063	4764	3515	3626	3060
Missing	0	0	0	0	0	0	0	0	0
Mean	0.738	0.734	0.722	0.725	0.741	0.724	0.779	0.671	0.755
Std. Deviation	0.135	0.119	0.146	0.120	0.132	0.121	0.162	0.186	0.365
Minimum	0.336	0.330	0.335	0.344	0.386	0.422	0.491	0.349	0.330
Maximum	1.704	1.746	10.460	1.635	2.419	1.844	1.599	4.249	18.260

Note. Excluded 45505 rows from the analysis that correspond to the missing values of the split-by variable Task Number

The physical task resulted in less average IBIs in the “Standing” condition (M = 0.67) compared to the “Sitting” condition (M = 0.78). This implied increased HR for the “Standing” condition. Again, we calculated the BPM for both conditions, which generated an average of 89,42 BPM during standing and 77,00 BPM while the participants were sitting down. Table 5 represents the average IBI values from all three devices combined. The column “FALSE” denotes the average IBI value measured in between tasks, meaning while participants were doing nothing or simply reading the instructions for the following task. With mean IBI values of (M = 0.75), it resulted in an average BPM of 79,47. Furthermore, we were able to determine a significant difference between the standing and sitting conditions. These effects were apparent in all three devices, with p-values < 0.001 (Table 7). It was also apparent in Table 6 that the EE4 measured more IBIs during the "Sitting" condition (N = 1147) compared to the "Standing" condition (N = 794). Nevertheless, the EE4 was still only capable to achieve 50% of the values obtained by the TMSI REFA. It can be concluded that there is a difference between the physical task conditions, indicating higher HR in the "Standing" condition. This is in line with our prediction that the physical tasks were able to elicit changes in HR/HRV. The expected movement artifacts for the EE4 were also evident (Hypothesis 3).

Table 6

Descriptive statistics of physical tasks conditions per device

	Standing			Sitting		
	REFA	Polar	EE4	REFA	Polar	EE4
N	2610	2610	794	2270	2276	1147
M	.639	0.639	.656	0.738	.739	.749
SD	.115	.115	.119	0.136	.137	.132
IBIs (%)	100.00	100.00	30.42	100.00	100.30	50.53
Min	.330	.358	.438	0.492	.327	.500
Max	2.419	1.382	1.094	1.401	1.599	1.391

Note. N = sample, M = Mean, SD = Standard Deviation, Min = Minimun, Max = Maximum

Table 7

ANOVA comparing sitting condition with the standing condition

Polar	DF	Sum of squares	Mean square	F-Value	P-level
Sitting	1	0.205	0.205	21.13	<.001*
Residuals	74	0.719	0.009		
REFA					
Sitting	1	0.204	0.205	20.91	<.001*
Residuals	73	0.713	0.009		
EE4					
Sitting	1	0.247	0.247	21.95	<.001*
Residuals	66	0.7422	0.011		

Discussion

Our study investigated alternative heart rate monitors (Empatica E4, Polar H10) and compared if their ability to measure HR/HRV is compatible with the golden standard TSMI REFA. We stated hypotheses that the heart rate monitors stay in agreement with the golden standard. Moreover, we hypothesized that non-congruent words and negative-emotional words in both Stroop tasks are expected to trigger greater arousal. Lastly, we stated the hypothesis that both task categories, mental and physical, should be able to elicit changes in HR/HRV, due to physical and psychological arousal. In addition, the alternative devices should be able to measure valid data in both task categories.

The results obtained in our study partially support the first hypothesis. The alternative heart rate monitors were able to achieve average IBIs that corresponds to the TSMI REFA, especially the Polar H10. The Polar H10 data resulted in a correlation close to 1, suggesting a near-perfect consensus between the devices. The EE4 on the contrary acquired less favorable results. Even though it was able to measure accurate HR/IBIs, with apparent R-tops that agreed with the TMSI REFA, the data was not consistent enough. Unfortunately, the number of IBIs were relatively low for all tasks, due to too much movement. Table 5 illustrates that the TMSI and Polar H10 were able to obtain a lot more measurements compared to the Empatica watch. The EE4 merely got 50% of IBI values measured by the TMSI, even less for the dynamic “Standing” condition (30%). The Polar obtained 100% for both “Sitting” and “Standing” conditions. Moreover, we were not able to determine a correlation between the EE4 and the TMSI. The IBIs of both devices could not be matched due to faulty data and missing values. Although the Empatica measured more accurately in the absence of movement, as shown by the “Sitting” condition, the number of missing values was simply too high to conclude reproducibility of the EE4.

We investigated the difference between the various conditions of the normal-/ and emotional Stroop tasks. In particular, if non-congruent words and negative emotional words provoke greater psychological arousal (Hypothesis 2). Our results indicated lower average IBI values, although it is a small difference, suggesting increased HR during the non-congruent task. Since the lower the IBI, the higher the heart rate. This suggests that the participants showed increased arousal for non-congruent words, as opposed to the congruent words. This would be in line with our second Hypothesis. However, we cannot be sure if there is an alternative explanation for this pattern, due to a lack of significant testing. For the influence of negative-emotional words on arousal, results showed conflicting effects. Our study did not find a distinction between positive-emotional and negative-emotional words. This is not in line with our second hypothesis, which assumed increased psychological arousal in response to negative words.

The results obtained in our analysis provided evidence in support of our third hypothesis. The physical tasks had an influence on HR/HRV that were captured by the devices. This is demonstrated firstly by the "FALSE" column, which indicated the mean IBI in between task conditions. It had the second-highest average IBI value, implying one of the lowest HR amid all other mental and physical tasks. This suggests that participants HR decreased when they were not completing a task. The participants showed the highest HR while they were standing up, proposing increased arousal in response to engaging in physical activity. Furthermore, the lowest HR was identified during the "Sitting" condition, possibly due to little physical activity. Lastly, concerning EE4´s sensitivity to movement during physical task conditions, we wanted to examine the devices independently. The Empatica did measure considerably less valid IBIs during the "Standing" condition in contrast to the less dynamic "Sitting" condition. This drawback is especially noticeable if we compared these findings to the other two devices. They paradoxically obtained more IBIs during the active "Standing" condition.

Limitations and Future Directions

One limitation of our study were the electrodes that were used in the beginning. We used electrodes called "Huggables", which are usually used for children and easy to remove from the skin. Over the course of conducting our study, we had some hot weather days, thus subjects were sweating and occasionally wearing sunscreen. This resulted in some incidences where the electrodes did not stick appropriately, therefore resulting in faulty data. After noticing we changed our electrodes and switched to ones that were more firmly attached to the skin (Kendall Electrodes). However, looking at the data after we finished the experiment, it was noticeable that there were some problems with obtained data. We were not able to use a moderate amount of the subjects due to artefacts caused by the electrodes. Overall, we were able to use the data from 28 of 44 subjects.

About the hypothesis that predicted higher arousal for noncongruent, it is difficult to determine the exact reason for the variation in the HR/HRV. Although it could be the changes of needed effort for incongruent words (vs. congruent), it could equally be stress caused by the staircase feature or the irritation from the negative feedback after making a mistake. However, we are not able to give a clear answer, this would require more additional research that controls for confounding variables, as well as significance testing like an ANOVA analysis. Overall, we cannot conclude causation or difference between both conditions (congruent vs. non-congruent) because we only provide descriptive statistics.

As mentioned previously, especially the EE4 was susceptible to motion artefacts, resulting in a large amount of missing data. Even small amounts of movement, like the slightest hand movements in our experiment, resulted in a significant lack of valid measurements. This consequently also made it impossible to establish a correlation between the EE4 and TMSI REFA. In the future, the EE4 device should be improved to make usage more reliable. Since the wireless monitor would be particularly useful for physical conditions allowing a wide range of movement. It would be necessary to test if these dynamic studies would be possible if the hand wearing the watch is kept steady. Overall, we would need further research to be able to test for correlation between the two devices.

Theoretical and Practical Implications

As far as the theoretical implications are concerned, our study adds confidence to previous research about modern heart rate monitors. Previous findings also showed valid data measured by the Polar H10 that is congruent with the TMSI. It is reported that the Polar monitor reliably tracks HR changes following psychological and physical arousal (Goodie et al., 2000). Menghini et al. (2018), showed findings that the Empatica E4 represented missing RR-interval/IBI detection, which resulted in faulty HRV, due to dynamic conditions and hand movements. This was similarly apparent in our study. Kunkels et al., (2019) proposed EE4´s use of PPG and BVP signals as a contributing factor. These measurements are especially vulnerable to motion artefacts, compared to the robust ECG measurements. However, previous research likewise illustrates that the Empatica can measure HR accurately, especially under static conditions (Menghini et al., 2018), which was also observed by our research with better measurements during the "Sitting" condition. Milstein and Gordon (2020) even found a nearly perfect correlation between the EE4 and its ECG device during static conditions, this could not be replicated by our study. This knowledge about the Empatica E4 could help future research in understanding the limitations of the device, along with exclusively using it for static conditions.

Previous research, about the normal Stroop task including Color-Word Interference, revealed contradicting findings. On the one hand, research suggests no cardiovascular effect elicited by mental tasks (Kunkels et al., 2019). On the other hand, Renaud and Blondin (1997) state that the Stroop task, and its non-congruent word trials, affect HR levels. They found that it demands more effort to indicate the color if there is a color-word interference. This theory could also be a possible explanation for the effects found in our non-congruent condition. Considering the emotional Stroop task, earlier findings argued that the task might have no influence on neurotypical participants (Dell`Acqua et al., 2020). Other researchers primarily looked at the interference effect of emotional material on participants' reaction time and error rates (Dresler et al., 2008). The emotional Stroop task was able to elicit affective interference exclusively for participants with clinical depression or anxiety. These individuals show increased rumination and an inability to disengage from negative-emotional words. Resulting in slower reaction time due to delayed information processing of emotional material (Dell`Acqua et al., 2020). Dresler et al., (2008) reported the same evidence with anxious individuals. Their attentional bias makes it harder to disengage from emotional words. This could be a reason why our research found no evidence for increased HR/HRV because emotional word conditions mainly trigger a prolonged response time (Straub et al., 2021). Hence, future research should additionally focus on the response time for the different emotional Stroop task trials (negative-emotional, positive-emotional), to find an effect. Lastly, these effects are not warranted for neurotypical participants, which our sample primarily consisted of. Thus, this should be considered in future research.

Finally, our study gave supporting findings on the device's ability to measure mental and physical arousal. Considering the Polar H10, previous findings stated that the chest-worn band was able to withstand an array of activities and physical movements, as well as psychological arousal (Goodie et al., 2000). These findings are in line with our study outcomes. For the EE4 opposite findings were found. Our research adds to the existing studies stating the unreliability of the EE4 in dynamic conditions that include physical activity. Future studies should further investigate possible countermeasures to prevent movement artefacts to find out the real potential of the EE4.

Practical implications

The usage of these alternative heart rate monitors could be of interest to researchers that look for more convenient devices. Our study gives insight into how practical and accurate these modern wireless monitors are. Especially for those who are interested in research, field experiments outside the laboratory, or longitudinal studies. The Polar H10 and the EE4 diminish the limitations on movement and could be equipped for field studies, as well as longitudinal studies. These wireless monitors have multiple advantages. It allows for continuous recordings and easy wearability, which is practical for long-term measurements. For these types of studies, participants would be able to comfortably wear the device over a longer period of time. The devices provide online data storage, as well as high-duration batteries without the requirement of replacement. Furthermore, as mentioned previously, the modern monitors have low costs, hence being affordable for many researchers (Kunkels et al. 2019). However, if the study involves excessive body movements, we suggest sticking to measurements using ECG monitors, like the Polar H10. This chest-strapped monitor is robust to movement and assessed valid HR/ HRV measurements. It prevents artefacts caused by movement and results in less faulty data (e.g., no loose electrodes). If studies make use of the EE4, employing PPG technology, we advise limiting the movement of the arm that is wearing the watch. Because our study was in a laboratory setting, we are not able to guarantee accurate data selection in field studies.

Nevertheless, field studies encourage the usage of these wireless monitors, simply because of the lack of restriction to electrodes, hence making physical movement possible.

Conclusion

Altogether, these results provide support for the Polar H10, and its reproducibility of data obtained by our golden standard TMSI REFA. The Polar H10 was able to detect arousal for mental- and physical tasks in almost perfect accordance with the TMSI. Although there is evidence for accurate detection of HR/IBIs by the Empatica E4, the movement artefacts resulted in too many missing values. Therefore, correlation and agreement between EE4 and TMSI REFA were not established. Although results suggest better measurements during the static “Sitting” condition, the usage of the EE4 should be carefully considered. The Polar H10 can be reliably recommended for research settings. Additionally, we found a significant difference between the “Sitting” and “Standing” conditions in all three devices. Indicating that the devices were able to measure changes in HR and HRV while the participants engaged in physical activity. Our study was not able to find a significant result in the normal-/ and the emotional Stroop task, demonstrating that the material used in the mental tasks did not elicit higher psychological arousal.

References

Chua, E. C.-P., Tan, W.-Q., Yeo, S.-C., Lau, P., Lee, I., Mien, I. H., Puvanendran, K., & Gooley, J. J. (2012). Heart rate variability can be used to estimate sleepiness-related decrements in psychomotor vigilance during total sleep deprivation. *Sleep: Journal of Sleep and Sleep Disorders Research*, *35*(3), 325–334.

Dell'Acqua, C., Dal Bò, E., Benvenuti, S. M., Ambrosini, E., Vallesi, A., & Palomba, D. (2021). Depressed mood, brooding rumination and affective interference: The moderating role of heart rate variability. *International Journal of Psychophysiology*, *165*, 47–55. https://doi-org.proxy-ub.rug.nl/10.1016/j.ijpsycho.2021.03.011

Dresler, T., Mériau, K., Heekeren, H. R., & van der Meer, E. (2009). Emotional Stroop task: Effect of word arousal and subject anxiety on emotional interference. *Psychological Research*, *73*(3), 364–371. https://doi-org.proxy-ub.rug.nl/10.1007/s00426-008-0154-6

Feng, Z., Li, N., Feng, L., Chen, D., & Zhu, C. (2021). Leveraging ECG signals and social media for stress detection. *Behaviour & Information Technology*, *40*(2), 116–133. https://doi-org.proxy-ub.rug.nl/10.1080/0144929X.2019.1673820

Goodie, J. L., Larkin, K. T., & Schauss, S. (2000). Validation of Polar heart rate monitor for assessing heart rate during physical and mental stress. *Journal of Psychophysiology*, *14*(3), 159–164. https://doi-org.proxy-ub.rug.nl/10.1027//0269-8803.14.3.159

Kunkels, Y. K., Roon, A. M., Wichers, M., & Riese, H. (2021). Cross-instrument feasibility, validity, and reproducibility of wireless heart rate monitors: Novel opportunities for extended daily life monitoring. *Psychophysiology*, *58*(10). https://doi-org.proxy-ub.rug.nl/10.1111/psyp.13898

Ledezma, C. A., & Altuve, M. (2019). Optimal data fusion for the improvement of QRS complex detection in multi-channel ECG recordings. *Med Biol Eng Comput*, *57*, 1673–1681. https://doi.org/10.1007/s11517-019-01990-3

Mehta, S., Lingayat, N., & Sanghvi, S. (2009). Detection and delineation of P and T waves in 12-lead electrocardiograms. *Expert Systems: International Journal of Knowledge Engineering and Neural Networks*, *26*(1), 125–143. https://doi-org.proxy-ub.rug.nl/10.1111/j.1468-0394.2008.00486.x

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

Milstein, N., & Gordon, I. (2020). Validating measures of electrodermal activity and heart rate variability derived from the empatica E4 utilized in research settings that involve interactive Dyadic States. *Frontiers in Behavioral Neuroscience*, *14*. https://doi-org.proxy-ub.rug.nl/10.3389/fnbeh.2020.00148

Renaud, P., & Blondin, J.-P. (1997). The stress of Stroop performance: Physiological and emotional responses to color–word interference, task pacing, and pacing speed. *International Journal of Psychophysiology*, *27*(2), 87–97. https://doi-org.proxy-ub.rug.nl/10.1016/S0167-8760(97)00049-4

Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health*, *5*(258). https://doi.org/10.3389/fpubh.2017.00258

Shaikh al arab, A., Guédon-Moreau, L., Ducrocq, F., Molenda, S., Duhem, S., Salleron, J., Chaudieu, I., Bert, D., Libersa, C., & Vaiva, G. (2012). Temporal analysis of heart rate variability as a predictor of post traumatic stress disorder in road traffic accidents survivors. *Journal of Psychiatric Research*, *46*(6), 790–796. https://doi-org.proxy-ub.rug.nl/10.1016/j.jpsychires.2012.02.006

Straub, E. R., Schmidts, C., Kunde, W., Zhang, J., Kiesel, A., & Dignath, D. (2022). Limitations of cognitive control on emotional distraction – congruency in the Color Stroop task does not modulate the Emotional Stroop effect. *Cognitive, Affective & Behavioral Neuroscience*, *22*(1), 21–41. https://doi-org.proxy-ub.rug.nl/10.3758/s13415-021-00935-4

Figure 1: https://commons.wikimedia.org/wiki/File:ECG-RRinterval.svg

Figure 2: https://www.kindpng.com/imgv/hhoioib_diagram-ecg-heart-diagram-diagram-schematic-circuit-ecg/

Appendix A

Table 8

Descriptive statistics of IBIs per participant for all devices independently

	Device	Valid	Mean	Std. Dev.	Minimum	Maximum
P001	TMSi	1250	0.903	0.095	0.653	1.401
	PolarH10	1250	0.903	0.095	0.653	1.399
	Empatica	345	0.940	0.116	0.656	1.391
P004	TMSi	1308	0.819	0.105	0.553	1.078
	PolarH10	1308	0.819	0.105	0.553	1.078
	Empatica	0	-	-	-	-
P005	TMSi	1665	0.669	0.054	0.330	0.911
	PolarH10	1665	0.671	0.057	0.437	1.599
	Empatica	122	0.685	0.078	0.500	0.922
P006	TMSi	1566	0.661	0.059	0.518	0.961
	PolarH10	1566	0.661	0.059	0.518	0.960
	Empatica	283	0.697	0.094	0.516	0.984
P008	TMSi	1579	0.736	0.068	0.336	1.108
	PolarH10	1579	0.736	0.080	0.330	1.704
	Empatica	21	0.766	0.079	0.641	0.953
P014	TMSi	1642	0.654	0.056	0.487	0.896
	PolarH10	1642	0.654	0.056	0.486	0.895
	Empatica	393	0.667	0.077	0.500	1.016
P016	TMSi	1438	0.840	0.102	0.613	1.154
	PolarH10	1438	0.840	0.102	0.614	1.156
	Empatica	205	0.904	0.127	0.547	1.250
P017	TMSi	1525	0.720	0.056	0.571	1.206
	PolarH10	1525	0.720	0.056	0.572	1.205
	Empatica	215	0.748	0.104	0.531	1.391
P018	TMSi	1597	0.708	0.080	0.554	1.189
	PolarH10	1597	0.708	0.080	0.554	1.192
	Empatica	292	0.747	0.095	0.547	0.984
P019	TMSi	1494	0.711	0.089	0.488	1.149
	PolarH10	1494	0.711	0.089	0.489	1.151
	Empatica	44	0.722	0.110	0.484	0.969
P020	TMSi	1601	0.624	0.038	0.523	0.824
	PolarH10	1601	0.624	0.038	0.523	0.825
	Empatica	144	0.634	0.049	0.516	0.828
P021	TMSi	1657	0.666	0.052	0.535	0.899
	PolarH10	1657	0.666	0.052	0.535	0.899
	Empatica	246	0.685	0.077	0.500	1.156
P025	TMSi	1489	0.762	0.061	0.593	1.044
	PolarH10	1489	0.762	0.061	0.593	1.045
	Empatica	423	0.783	0.086	0.563	1.078
P026	TMSi	1523	0.750	0.084	0.531	1.151
	PolarH10	1522	0.750	0.084	0.533	1.151
	Empatica	101	0.816	0.160	0.578	1.844
P027	TMSi	1886	0.573	0.076	0.450	2.419

	PolarH10	1886	0.572	0.063	0.449	1.634
	Empatica	457	0.570	0.061	0.406	0.875
P028	TMSi	1790	0.641	0.047	0.511	0.858
	PolarH10	1790	0.641	0.047	0.512	0.859
	Empatica	758	0.648	0.061	0.500	0.938
P029	TMSi	1957	0.570	0.040	0.459	1.024
	PolarH10	1957	0.569	0.041	0.358	1.026
	Empatica	395	0.575	0.044	0.438	0.734
P032	TMSi	1588	0.708	0.072	0.501	0.964
	PolarH10	1588	0.708	0.072	0.501	0.964
	Empatica	219	0.729	0.097	0.500	1.094
P034	TMSi	1668	0.628	0.049	0.496	0.978
	PolarH10	1668	0.628	0.049	0.495	0.979
	Empatica	213	0.644	0.070	0.453	0.797
P035	TMSi	1611	0.711	0.080	0.511	1.481
	PolarH10	1611	0.710	0.080	0.327	1.481
	Empatica	230	0.706	0.091	0.500	1.000
P036	TMSi	1651	0.691	0.039	0.535	0.851
	PolarH10	1651	0.691	0.039	0.536	0.852
	Empatica	100	0.692	0.051	0.578	0.797
P037	TMSi	1518	0.861	0.068	0.337	1.071
	PolarH10	1518	0.861	0.068	0.329	1.070
	Empatica	1081	0.871	0.062	0.656	1.203
P038	TMSi	2002	0.610	0.050	0.498	0.932
	PolarH10	2002	0.610	0.050	0.499	0.932
	Empatica	223	0.654	0.097	0.469	1.266
P039	TMSi	1255	0.850	0.095	0.486	1.150
	PolarH10	1255	0.850	0.095	0.485	1.152
	Empatica	92	0.857	0.111	0.656	1.078
P040	TMSi	1505	0.720	0.050	0.425	0.868
	PolarH10	1505	0.720	0.050	0.423	0.868
	Empatica	243	0.761	0.146	0.516	1.391
P041	TMSi	1895	0.610	0.043	0.464	0.837
	PolarH10	1895	0.610	0.046	0.359	1.131
	Empatica	391	0.628	0.072	0.438	0.875
P042	TMSi	1491	0.748	0.068	0.580	1.126
	PolarH10	1491	0.748	0.068	0.579	1.127
	Empatica	338	0.759	0.077	0.578	1.000
P043	TMSi	1703	0.696	0.062	0.515	1.088
	PolarH10	1704	0.695	0.063	0.353	1.087
	Empatica	184	0.738	0.121	1.516	1.500

Note. Empatica did not measure any data for subject 4.

Appendix B: Word list of emotional Stroop task

Negative words	Positive words
War	Peace
Attack	Hug
Pain	Happy
Blood	Healthy
Bomb	Flower
Kill	Cure
Gun	Toy
Death	Life
Anger	Friendly
Nuclear	Sunshine
Panic	Calmness
Anxiety	Excited
Combat	Ally
Enemy	Friends
Explosion	Celebration
Fighting	Love