How do Wearables Empatica E4 and Polar H10 perform against ECG on Heart Rate Variability?

Harmiena Geertruida Tamsma

S3804178

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: 45

Supervisor: dr. Mark M. Span

Second evaluator: prof. dr. Dick de Waard

In collaboration with: Lisanne Zondag, Nienke Buist, Rover Willemars and Theres Patzelt

June 20th, 2022

Abstract

In this study, wearables Polar H10 and Empatica E4 were compared to a golden standard ECG amplifier, the TMSI Refa. The Polar H10 and Empatica E4 were tested on cross-instrument and test-retest validity for monitoring Heart Rate Variability in a lab experiment. All three devices were tested simultaneously in four conditions, sitting, standing, a normal Stroop task and an emotional Stroop task. The sample consisted of 28 Psychology students. Three hypotheses were tested in this study. Firstly, it was hypothesized that all three devices produced similar Interbeat Intervals during the study over all conditions. Supporting evidence was found that this is true for the Polar H10 and the TMSI Refa, not the Empatica E4. Secondly, it was expected that the standing condition would lead to smaller Interbeat Intervals in comparison with the sitting condition. The TMSI Refa and the Polar H10 found a significant difference between these conditions. Thirdly, it was hypothesized that the normal Stroop task and the emotional Stroop task condition would lead to smaller Interbeat Intervals in comparison to the sitting condition. This effect was not found in the study. A lack of data points in the measurements of the Empatica E4 made a clear conclusion about its results unattainable. The results of the study are promising for the potential to use the Polar H10 as a tool in further research.

Keywords: Heart Rate Variability, Empatica E4, Polar H10, Cross-Instrument Validity, Test-Retest Validity

How do Wearables Empatica E4 and Polar H10 perform against ECG on Heart Rate Variability?

Recent developments in heart monitoring gear have led to more accessible ways to measure heart rate variability (HRV) parameters (Schuurmans et al., 2020), like Heart Rate (HR). Wearables like the Empatica E4 and Polar H10 are cheaper (Schuurmans et al., 2020) and more comfortable (Kunkels et al., 2021) than the golden standard TMSI Refa ECG. These new devices might provide new research opportunities. Besides being less expensive than currently used devices, the Empatica E4 and Polar H10 are wireless and free of electrodes (Schuurmans et al., 2020). These properties make it easier to measure heart rate in an ambulatory setting. Also, it might make measuring HR less uncomfortable for research participants (Kunkels et al., 2021), which hopefully will lead to less dropout among participants. In addition, these wearable, wireless devices are possibly a convenient tool to measure HR while the participant is moving. However, before incorporating the Empatica E4 and the Polar H10 in (experimental) research, it is important to know if these devices perform similarly to currently used devices. This research will examine the reliability of the Empatica E4 and the Polar H10 by comparing their measurements of HR with the TMSI Refa ECG.

At this moment, most psychophysiological research utilizing HRV is done with electrocardiogram (ECG) devices (Schuurmans et al., 2020), like the TMSI REFA ECG used in this study. An ECG device monitors the electrical activity in the heart (Rowlands & Sargent, 2014). This electrical activity usually follows a certain pattern called the PQRST complex (Rowlands & Sargent, 2014). The letters are the names of the tops in the complex (Rowlands & Sargent, 2014). In this research the time in seconds between the R-tops of two separate heartbeats will be used, the interbeat interval (IBI) (Kunkels, 2021). The IBIs are used to assess HRV, smaller IBIs indicate a faster heartbeat and bigger IBIs a slower heartbeat. The variation in IBIs or HR is called Heart Rate Variability (Kunkels, 2021). Heart rate is the number of heartbeats per minute (Rowlands & Sargent, 2014). IBIs can be also calculated using HR.

ECG measurements are reliable and precise, but have their downsides too. Movements can decrease the reliability of the measurements. Some movement is possible with the TMSI REFA ECG, but the size of the device and the fact that the subject is wired to the machine restricts the movement of participants.

Here’s where wearable HR monitors like the Polar H10 and Empatica E4 might come to play. The Empatica E4 is a wristband that derives HR from Blood Volume Pulse measurements. Research by Schuurmans et al. (2020) shows that the Empatica E4 is able to give reliable results for monitoring HRV in comparison to an ECG. Though, in the experiment the participants’ HRV parameters were only measured in minimal movement conditions. In a study by Van Voorhees et al. (2022) the Empatica E4 was worn for 24 hours by participants with PTSD symptoms in an ambulatory setting. The researchers concluded that the Empatica E4 was not suitable for these conditions, due to too much missing data. Possibly the Empatica E4 was not able to pick up the HR because of the participants’ movements (Van Voorhees et al., 2022). The participants in this research did however indicate that the Empatica E4 was only slightly to moderately uncomfortable (Van Voorhees et al., 2022).

The Polar H10 is a band worn around the chest, measuring HR via a sensor that touches the skin. A study by Correia et al. (2020) showed that the Polar H10 was able to pick up changes in participants’ HRV during a Stroop task. Research by Speer et al. (2020) showed reliable results for the Polar H10 on HRV. The researchers tested this device on children while they were at school for multiple days. The Polar H10 also showed reliable results during sports, like running and cycling, according to Budig et al. (2021).

Research on HRV indicates that it might be a useful predictor in the field of psychology. For example, studies have shown that HRV can predict both Major Depressive Disorder (Koch et al., 2019) and disordered eating (Watford et al., 2020). Also, research by Pendleton et al. (2016) has shown a correlation between HRV and mental engagement during cognitive and psychomotor tasks. In this study, the participants performed both a cognitive and a psychomotor task, whereby either participants got assigned Executive Control tasks or Non-Executive Control tasks. All conditions showed a decline in HRV during the tasks, whereas the Non-Executive Control tasks lead to the biggest change (Pendleton et al., 2016).

Having wearable HR monitors available for research, might open up new research possibilities for psychologists. Also, in the long run it may lead to new clinical interventions based on HR measurements. Yet, at this moment a limited amount of research is done on the reliability of the Empatica E4 and the Polar H10. Studies comparing the Empatica E4 and ECG devices find mixed results on the reliability of measurements of HRV parameters (Schuurmans et al., 2020; Van Voorhees et al., 2022).

The aim of this study is to further examine the usability of the Polar H10 and Empatica E4, by comparing the HRV results of these devices to the results of the TMSI Refa ECG. These measurements will be done in different conditions. The participants' HR will be monitored in a physical task, with different conditions: sitting and standing. Earlier studies have shown that standing leads to an increase in HR (Riese et al., 2004; Vrijkotte et al., 2000). Furthermore, participants will do a classic Stroop Task and an Emotional Stroop Task. According to earlier research, a Stroop Task can lead to differences in HR and stress (Correia et al., 2020; Renaud & Blondin, 1997). To increase the mental effort during the task, the participants have to perform a psychomotor vigilance task at the same moment.

First, our hypothesis is that the Heart Rate measurements obtained from either the TMSI Refa, Polar H10 or the Empatica E4 can be interchangeably used in research for deriving Heart Rate

Variability. It is expected that the IBIs derived from the Polar H10 and the Empatica E4 measurements do not differ significantly from the IBIs derived from the TMSI Refa. Secondly, the IBIs are expected to be shorter in the standing condition compared to the sitting condition. Thirdly, we expect shorter IBIs during both the normal and the emotional Stroop tasks in comparison with the sitting condition.

Method

Participants

A total of 44 students participated in this study. During data processing, 16 participants got excluded due to unusable data. Of the remaining 28 participants, 20 were women and 8 were men. The participants consisted of students following a first-year research course at the University of Groningen. Students that take this course are placed in a participants pool, through which they can sign up for studies to obtain SONA-credits. These credits are mandatory for passing the course. However, students are free to choose which studies they sign up for.

Prescreen

While signing up for the experiment, participants filled out a questionnaire. Participants indicated if they had normal vision either naturally or corrected (normal vision = 21, contact lenses = 5, glasses = 2). Also, they declared whether or not they had a driver’s license (license = 11, no license = 16, decline to answer = 1). Finally, participants were asked what their dominant hand was (lefthanded = 2, righthanded = 26).

Design

In this experiment, we looked at the cross instrumental validity and reliability of the Polar H10 and Empatica E4 to monitor HRV. To compare devices, all participants wore three heart rate monitors simultaneously during the experiment, a regular ECG, a Polar H10, and an Empatica E4.

Before starting the experiment, the participants had the opportunity to read information about the study and were asked to sign an informed consent form. A verbal explanation was given on top of the written instructions and information. The verbal instructions were given in either English, Dutch, or German, depending on the preferences of the participant. The on-screen instructions and the Stroop task were in English.

Procedure

The experiment consisted of the following tasks participants had to carry out, while they wore all three heart rate monitors simultaneously. All the instructions for the task showed up on a computer screen. First, participants sat down (1 minute) as a baseline measurement. Next, we measured heart rate while the participants were standing up (1 minute). The sitting and standing tasks were followed by two Stroop Tasks, which also appeared on the monitor. Participants either started with a regular or an emotional Stroop Task, in random order. Participants indicated the colour of the words with a button box, made for this experiment. Both Stroop Tasks combined lasted around 20 minutes. Any of the aforementioned tasks could be interrupted by the alarm of the psychomotor vigilance task, which the participant then had to turn off again on a separate laptop. The alarm went off every 3 to 5 minutes.

Apparatus

For this study, we made use of three different heart rate monitors. We used a TMSI REFA amplifier as a golden standard to compare with the Polar H10 and the Empatica E4.

Polar H10

The Polar H10 is a band worn around the chest. In the band, there is a sensor located that measures ECG.

Empatica E4

The Empatica E4 is a wristband with several functions, in our study we only looked at the capability of the band to monitor HR. The device makes use of a photoplethysmography

(PPG) sensor (Empatica, 2020). This sensor does not directly measure the heart rate, instead it monitors differences in blood volume pulse (BVP) (Empatica, 2020). The HR is derived from these differences in BVP (Empatica, 2020). The interbeat intervals (IBIs) are determined based on the HR (Empatica, 2020). For this study, we used the IBI measurements to infer the HRV.

Button box

The button box used for the Stroop task was made by the research support department at the University of Groningen. On this box were four buttons, lighting up in different colours in a set order: red, blue, green, and yellow. These buttons were used to indicate the colour of the words in the Stroop Tasks.

Materials

Stroop Tasks

Both Stroop Tasks were performed on a computer screen with an attached button box. A staircase procedure was programmed into the task, gradually increasing the difficulty of the task. Each time a participant gave two correct answers in a row, the available time to react got shorter. If the participant made a mistake or ran out of time, the available reaction time got longer. Also, participants were presented with on-screen feedback. After a correct response ‘Good job!’ showed up on the monitor in green font. If the participant answered incorrectly or ran out of time, the feedback ‘FALSE! Try harder!’ showed up in red font.

For the original Stroop Task, participants began with congruent trials (e.g. RED) and moved on to incongruent trials (e.g. RED), and ended with mixed trials. Before every new set of trials, the participants were able to practice the task. Each set of trials consisted of four words, repeated 16 times.

For the Emotional Stroop Task, participants started with ‘positive emotion’ words (e.g. JOY or ALLY) depicted in either red, blue, green, or yellow. Next, the participants were shown ‘negative emotion’ words (e.g. War or Anxiety). Afterwards, words with positive and negative valence were mixed. All three conditions consisted of a sequence of 20 trials. Both the positive and the negative condition contained 16 words.

Psychomotor Vigilance Task

For the Psychomotor Vigilance Task, we used an alarm that went off at random times during the experiment. The alarms went off at an interval varying from 3 to 5 minutes. Participants had to turn off the alarm by pressing the space button on the laptop, which was placed on their right.

Results

Data Processing

Before comparing the data, the data were processed using a custom-made script in Matlab, called Alakazam. All ECG data was inspected prior to the data processing, to check for abnormalities. Additionally, a Poincare plot was made to assess the data. Not all measurements proved to be suitable for analysis. Firstly, the data from 16 participants have been found to be unusable. Secondly, the Empatica E4 monitored no valid data points for participant 4.

Alakazam was used to calculate the R-tops and IBI based on the HR as monitored by the devices. Next, the IBI for the different monitors were lined up. Lastly, the program was used to match the IBIs for the TMSI Refa and the Polar H10. Pairs with missing values were deleted. Due to a difference in the amount of valid data points measured by the Empatica E4 compared to the TMSI Refa, the IBIs of both devices were not paired. As a result, fewer statistical analyses have been performed on the data output of the Empatica E4. Before the data got analyzed, a value of 2ms was determined as an acceptable deviation of calculated IBIs for the TMSI Refa.

Descriptive Statistics

Table 1 shows the descriptive statistics of the IBI based on all three devices. One thing that catches the eye, is the amount of valid data points the Empatica E4 has in comparison to the TMSI Refa and Polar H10. The difference in data points made it unattainable to compare the IBI of all three devices pairwise.

Besides a difference in the amount of valid IBIs, the Empatica E4 also shows a higher mean and a bigger standard deviation than the TMSI Refa. The Polar H10 on the other hand, gives exactly the same mean and standard deviation as the TMSI Refa. There is however a noticeable difference between the reported maximum IBI for the Polar H10, the Empatica E4 and the TMSI Refa.

Table 1

Interbeat Intervals (in seconds) derived from the TMSI Refa, Empatica E4 & Polar H10

	IBI [s]		
	TMSI Refa	Empatica E4	Polar H10
Valid	44854	7758	44854
Missing	0	0	0
Mean	0.700	0.728	0.700
SD	0.107	0.135	0.107
Minimum	0.330	0.406	0.327
Maximum	2.419	1.844	1.704

Table 2 shows the descriptive statistics of the IBIs derived from the TMSI Refa measurements. Noteworthy is participant 27, as it is the only participant with an IBI longer than 1,5 seconds in this table.

Table 2

Descriptive statistics based on the Interbeat Intervals per participant (P) as measured by the TMSI Refa

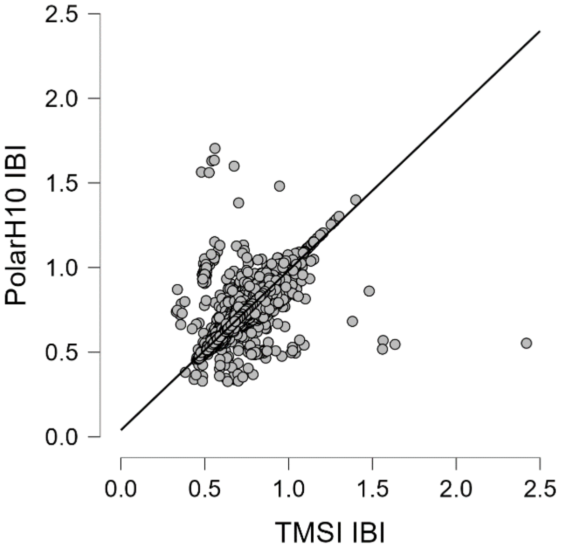
	IBI [s]				
	Valid	Mean	SD	Minimum	Maximum
P001	1250	0.903	0.095	0.653	1.401
P004	1308	0.819	0.105	0.553	1.078
P005	1665	0.669	0.054	0.330	0.911
P006	1566	0.661	0.059	0.518	0.961
P008	1579	0.736	0.068	0.336	1.108
P014	1642	0.654	0.056	0.487	0.896
P016	1438	0.840	0.102	0.613	1.154
P017	1525	0.720	0.056	0.571	1.206
P018	1597	0.708	0.080	0.554	1.189
P019	1494	0.711	0.089	0.488	1.149
P020	1601	0.624	0.038	0.523	0.824
P021	1657	0.666	0.052	0.535	0.899
P025	1489	0.762	0.061	0.593	1.044
P026	1523	0.750	0.084	0.531	1.151
P027	1886	0.573	0.076	0.450	2.419
P028	1790	0.641	0.047	0.511	0.858
P029	1957	0.570	0.040	0.459	1.024
P032	1588	0.708	0.072	0.501	0.964
P034	1668	0.628	0.049	0.496	0.978
P035	1611	0.711	0.080	0.511	1.481
P036	1651	0.691	0.039	0.535	0.851
P037	1518	0.861	0.068	0.337	1.071
P038	2002	0.610	0.050	0.498	0.932
P039	1255	0.850	0.095	0.486	1.150
P040	1505	0.720	0.050	0.425	0.868
P041	1895	0.610	0.043	0.464	0.837
P042	1491	0.748	0.068	0.580	1.126
P043	1703	0.696	0.062	0.515	1.088

Polar H10

The correlation between all the IBIs over all conditions of the TMSI Refa and the Polar H10 was calculated, as is shown in figure 1. The association between the IBIs of the two devices was significant.

Figure 1

Interbeat Intervals in seconds as measured by the TMSI Refa and the Polar H10



Note. Pearson´s r = .942 (p<.001)

In table 3 the descriptive statistics of the IBIs based on the Polar H10 measurements are displayed. None of the IBI means per participant differ more than 2ms in comparison to the participants’ means of the TMSI Refa. When comparing the standard deviations of the golden standard to the Polar H10, participants 5, 8, 27 and 41 show a SD that differs more than 2ms. The minimum of participants 5, 8, 29, 35, 37, 41 and 43 deviate more than 2ms from the TMSI Refa. The maximum of participants 5, 8, 18, 27 and 41 also showed a difference larger than 2ms compared to the TMSI Refa.

Table 3

Descriptive statistics based on the Interbeat Intervals per participant (P) as measured by the Polar H10

	IBI				
	N	Mean	SD	Minimum	Maximum
P001	1250	0.903	0.095	0.653	1.399
P004	1308	0.819	0.105	0.553	1.078
P005	1665	0.671	0.057	0.437	1.599
P006	1566	0.661	0.059	0.518	0.960
P008	1579	0.736	0.080	0.330	1.704

Table 3

Descriptive statistics based on the Interbeat Intervals per participant (P) as measured by the Polar H10

	IBI				
	N	Mean	SD	Minimum	Maximum
P014	1642	0.654	0.056	0.486	0.895
P016	1438	0.840	0.102	0.614	1.156
P017	1525	0.720	0.056	0.572	1.205
P018	1597	0.708	0.080	0.554	1.192
P019	1494	0.711	0.089	0.489	1.151
P020	1601	0.624	0.038	0.523	0.825
P021	1657	0.666	0.052	0.535	0.899
P025	1489	0.762	0.061	0.593	1.045
P026	1522	0.750	0.084	0.533	1.151
P027	1886	0.572	0.063	0.449	1.634
P028	1790	0.641	0.047	0.512	0.859
P029	1957	0.569	0.041	0.358	1.026
P032	1588	0.708	0.072	0.501	0.964
P034	1668	0.628	0.049	0.495	0.979
P035	1611	0.710	0.080	0.327	1.481
P036	1651	0.691	0.039	0.536	0.852
P037	1518	0.861	0.068	0.329	1.070
P038	2002	0.610	0.050	0.499	0.932
P039	1255	0.850	0.095	0.485	1.152
P040	1505	0.720	0.050	0.423	0.868
P041	1895	0.610	0.046	0.359	1.131
P042	1491	0.748	0.068	0.579	1.127
P043	1704	0.695	0.063	0.353	1.087

A paired two-sample t-test was performed, to test for equivalence between the IBI means per participant of the TMSI Refa and the Polar H10. The results of the test are listed in table 4. Based on the equivalence test, no evidence was found that the means of the IBIs in this sample differ significantly.

Table 4

Equivalence Paired Sample t-test

			t	df	p
TMSI/Refa	-	Polar H10	-0.015	44853	0.988
	Upper bound		-145.645	44853	<.001
	Lower bound		145.645	44853	<.001

Experimental Conditions

Table 5 shows the descriptive statistics of the TMSI Refa and the Polar H10 in four different conditions, namely sitting, standing, a normal Stroop task and an emotional Stroop task. When comparing the values in the table of the TMSI Refa and Polar H10 based on condition, it is noteworthy that none of the means or standard deviations of the Polar differ more than 2ms from the TMSI. All minimum and maximum IBIs from the Polar H10 deviate from the 2ms limit, except for the maximum value in the standing condition.

The mean of the IBI in the sitting condition is larger than the means in the standing, the normal Stroop task and the emotional Stroop task. This is true for both the TMSI Refa and the Polar H10. The difference in means is the biggest between the sitting and the standing condition.

Table 5

Descriptive Statistics of the TMSI Refa and Polar H10 in four different conditions

	Sitting		**Standing**		**Normal**		**Emotional**	
	TMSI	**Polar**	**TMSI**	**Polar**	**TMSI**	**Polar**	**TMSI**	**Polar**
N	2270	2276	2610	2610	10832	10837	26908	26901
Mean	0.738	0.739	0.639	0.639	0.705	0.705	0.700	0.700
SD	0.136	0.137	0.115	0.115	0.111	0.110	0.097	0.097
Minimum	0.492	0.327	0.450	0.358	0.489	0.437	0.336	0.329
Maximum	1.401	1.599	1.381	1.382	2.419	1.634	1.189	1.704

Table 6 displays the result of three, left-tailed independent sample t-tests, each comparing the means of the sitting conditions with the other conditions. After testing the sitting condition versus the standing condition, a significant p-value was found. This means that the means found in the standing condition are significantly smaller than the means in the sitting condition. Additionally, the t-value of the Polar H10 deviates from the t-value found for the TMSI Refa.

When comparing the sitting condition versus the normal Stroop task condition, a difference was found between the group means for both the TMSI Refa and the Polar H10. This difference was not significant for either of the devices. The t-value for the Polar H10 differed from the t-value of the TMSI Refa.

Lastly, when the t-test compared the sitting and the emotional Stroop task condition, a significant p-value was found. This indicates that the means in the emotional Stroop task condition were significantly smaller than the means in the sitting condition. This effect was found for both the TMSI Refa and the Polar H10. Both devices had different t-values however.

Table 6

Left-tailed independent sample t-test, testing for a difference in means in the sitting condition versus the standing, normal Stroop and emotional Stroop task conditions.

	Independent Sample T-Test					
	Sitting vs Standing		Sitting vs Normal Stroop		Sitting vs Emotional Stroop	
	TMSI	Polar	TMSI	Polar	TMSI	Polar
t	3.629	3.631	1.493	1.489	1.777	1.780
df	54	54	54	54	54	54
p	< .001*	< .001*	0.071	0.071	0.041*	0.040*

Note. *p < .05, for all tests, the alternative hypothesis specifies that group *sit* is greater than the other group.

Empatica E4

Table 7 shows the descriptive statistics of the Empatica E4 per participant. When comparing this table to table 2, which contains the descriptive statics of the TMSI Refa, a few differences can be observed. First, the number of valid data points differs for the two devices. The TMSI Refa has more valid data points per participant than the Empatica E4. When looking at the Empatica, participant 37 has 1081 valid IBIs. For the Empatica, there was not another participant with this many valid IBIs. Participant 8 shows the lowest amount of IBIs as measured by the Empatica E4. This is except for participant 4 which did not have any valid data point for the Empatica. In comparison, the amount of valid IBIs for the TMSI Refa ranged between 1250 (participant 1) and 2002 (participant 38).

When observing the means per participant per device, for the Empatica E4 all means but one deviate more than 2ms from the means for the TMSI Refa. The mean of the Empatica E4 for participant 36 does not differ more than 2ms in comparison to the mean of the TMSI Refa.

When looking at the minimum IBI derived from the Empatica E4 and TMSI Refa measurements, it is noteworthy that most IBIs differ more than 2ms from each other. The minimum IBIs from the Empatica E4 and the TMSI Refa for participants 6, 32, 42 and 43 are within the 2ms limit. None of the maximum IBIs for the Empatica E4 and TMSI Refa are within this range.

Table 7

Descriptive statistics based on the Interbeat Intervals per participant (P) as calculated by the Empatica E4

	Valid	Mean	Std. Deviation	Minimum	Maximum
P001	345	0.940	0.116	0.656	1.391
P005	122	0.685	0.078	0.500	0.922
P006	283	0.697	0.094	0.516	0.984
P008	21	0.766	0.079	0.641	0.953

Table 7

Descriptive statistics based on the Interbeat Intervals per participant (P) as calculated by the Empatica E4

	Valid	Mean	Std. Deviation	Minimum	Maximum
P014	393	0.667	0.077	0.500	1.016
P016	205	0.904	0.127	0.547	1.250
P017	215	0.748	0.104	0.531	1.391
P018	292	0.747	0.095	0.547	0.984
P019	44	0.722	0.110	0.484	0.969
P020	144	0.634	0.049	0.516	0.828
P021	246	0.685	0.077	0.500	1.156
P025	423	0.783	0.086	0.563	1.078
P026	101	0.816	0.160	0.578	1.844
P027	457	0.570	0.061	0.406	0.875
P028	758	0.648	0.061	0.500	0.938
P029	395	0.575	0.044	0.438	0.734
P032	219	0.729	0.097	0.500	1.094
P034	213	0.644	0.070	0.453	0.797
P035	230	0.706	0.091	0.500	1.000
P036	100	0.692	0.051	0.578	0.797
P037	1081	0.871	0.062	0.656	1.203
P038	223	0.654	0.097	0.469	1.266
P039	92	0.857	0.111	0.656	1.078
P040	243	0.761	0.146	0.516	1.391
P041	391	0.628	0.072	0.438	0.875
P042	338	0.759	0.077	0.578	1.000
P043	184	0.738	0.121	0.516	1.500

Note. Empatica E4 had no valid measurements for participant 4.

Table 8 shows the descriptive statistics of the Empatica E4 per condition. The descriptives of TMSI Refa are added as a reference. Firstly, the Empatica E4 has fewer valid data points than the TMSI Refa. Secondly, when comparing the values in the table of the Empatica E4 to the values of the TMSI Refa, not one value stays within the 2ms limit.

The Empatica E4 shows a difference between the sitting and the standing condition, just like the TMSI Refa and the Polar H10. The differences between the sitting condition and the normal Stroop task and emotional Stroop task condition are also visible for the group means of the Empatica E4. Due to the difference in valid data points between the Empatica E4 and the TMSI Refa, the difference between the conditions will not be tested further for the Empatica.

Table 8

Descriptive Statistics of the means of the TMSI Refa and Empatica E4 in four different conditions

	IBI [s]							
	Sitting		Standing		Normal		Emotional	
	TMSI	E4	TMSI	E4	TMSI	E4	TMSI	E4
Valid	2270	1147	2610	794	10832	1769	26908	3331
Mean	0.738	0.749	0.639	0.656	0.705	0.723	0.700	0.741
SD	0.136	0.132	0.115	0.119	0.111	0.139	0.097	0.132
Minimum	0.492	0.500	0.450	0.438	0.489	0.406	0.336	0.438
Maximum	1.401	1.391	1.381	1.094	2.419	1.844	1.189	1.500

Discussion

The aim of this research was to examine the reliability and validity of the Polar H10 and Empatica E4 for research purposes. This was done by testing these devices and the TMSI Refa in different conditions and comparing the results. The hypothesis was that either of the three HR monitors could be used in research to derive HRV.

Firstly, this study provides supporting evidence that the measurements of Polar H10 can be used instead of the TMSI Refa in the conducted experiment. There was a strong correlation found between the data of the Polar H10 and the TMSI Refa. The IBIs of both the devices did not significantly differ from each other. These findings were consistent with earlier findings by Budig et al. (2021) and Speer et al. (2020).

This study was unable to provide evidence to support the hypothesis that the Empatica E4 can be used instead of the TMSI Refa or the Polar H10. The amount of missing data points for the Empatica E4 made it unattainable to compare the data of this device with the TMSI Refa in a statistical analysis. The Empatica E4 and the TMSI Refa were compared by looking at the mean IBIs per participant and per condition, to check for differences. Most mean IBIs of the Empatica differed from the TMSI Refa. This result does not provide evidence for or against the hypothesis, since the difference in mean IBIs might also be explained by the difference in valid data points between the two devices. The findings in regard to the Empatica E4 were consistent with the findings of Voorhees et al. (2022). The researchers for this study also found that the Empatica E4 was missing data points. According to the researchers, movements of the participants might have been the cause of the missing data. This might also be the case in this study, since participants were not instructed on the movements of their hands. This also provides a possible explanation for the difference in findings with the study of Schuurmans et al. (2020). In this study, the Empatica E4 was only used in minimal movement conditions.

Secondly, it was hypothesized that the IBIs for the standing condition were smaller than the IBIs in the sitting condition. This study found a significant difference between the sitting and the standing conditions, with smaller IBIs in the standing condition. This effect was found for the TMSI Refa and the Polar H10. The Empatica E4 was able to detect a difference in group means, but no statistical analysis was performed to test the significance. The findings of the TMSI Refa and Polar H10 were consistent with earlier research by Riese et al. (2004) and Vrijekotte et al. (2000).

Thirdly, it was hypothesized that the normal Stroop task and emotional Stroop task conditions would lead to smaller IBIs in comparison to the sitting condition. In the normal Stroop task condition, no significant difference was found in comparison to the sitting condition. The Empatica E4 was able to detect a difference in group means, but no statistical analysis was performed to test the significance. Both the TMSI Refa and the Polar H10 found an insignificant p-value. This finding is inconsistent with earlier research by Correia et al.

(2020) and Renaud & Blondin (1997). In those studies, the researchers found differences in HRV during a Stroop task in comparison with a resting condition.

The mean IBIs found in the emotional Stroop condition were found to be significantly different from the sitting condition. This effect was both visible in the measurements of the TMSI Refa and the Polar H10. The Empatica E4 also showed a difference in group means between the sitting and standing condition, but this was not statistically tested. The findings of the TMSI Refa and the Polar H10 supported the hypothesis that the emotional Stroop task condition would lead to smaller IBIs in comparison with the sitting condition.

Limitations

In this study, multiple limitations were encountered. Firstly, during the experiments, electrodes were used for the ECG measurements of the TMSI Refa. However, these electrodes caused problems for our data collection, because they did not always stick. This made some of the data unusable, which led to the exclusion of multiple participants. During the data collection phase the electrodes were swapped with different electrodes, this solved the problem.

Secondly, all the participants were students following a course where participating in research was mandatory to pass. This possibly influenced the effort the participants were willing to put into the tasks. This might be an explanation for the non-significant results for the normal and emotional Stroop task condition. For further research, a non-convenience sample could be used to test for mental stress or mental effort.

Thirdly, both the normal and the emotional Stroop tasks were in English, while most students were Dutch. Possibly, carrying out a Stroop task is less stressful if it is not in your native language. In further research, a Stroop task could be used that is in the participant’s native language.

Despite these limitations, this study found supporting evidence that the Polar H10 is able to produce similar results to those of a golden standard ECG, like the TMSI Refa. This study was not able to provide proof that the Polar H10 can be used for deriving HRV in mental stress conditions. However, the measurements of the Polar H10 during the Stroop tasks were consistent with the measurements of the TMSI Refa. Additionally, the HRV between the sitting and standing condition was visible in the data of the Polar H10 and the TMSI Refa. Due to missing data, no supporting evidence was found for the hypothesis that the Empatica E4 can be used in research instead of the TMSI Refa. In conclusion, the results show that Polar H10 has the potential to be used in research.

References

Budig, M., Keiner, M., Stoohs, R., Hoffmeister, M., & Höltke, V. (2021). Heart Rate and Distance Measurement of Two Multisport Activity Trackers and a Cellphone App in Different Sports: A Cross-Sectional Validation and Comparison Field Study. *Sensors (Basel, Switzerland)*, *22*(1). https://doi.org/10.3390/s22010180

Correia, B., Dias, N., Costa, P., & Pêgo, J. M. (2020). Validation of a Wireless Bluetooth Photoplethysmography Sensor Used on the Earlobe for Monitoring Heart Rate Variability Features during a Stress-Inducing Mental Task in Healthy Individuals. *Sensors (Basel, Switzerland)*, *20*(14). https://doi.org/10.3390/s20143905

Empatica. (2020, January 24). *E4 data - IBI expected signal*. www.empatica.com. Retrieved May 22, 2022, from https://support.empatica.com/hc/en-us/articles/360030058011-E4-data-IBI-expected-signal

Koch, C., Wilhelm, M., Salzmann, S., Rief, W., & Euteneuer, F. (2019). A meta-analysis of heart rate variability in major depression. *Psychological Medicine*, *49*(12), 1948–1957. https://doi.org/10.1017/S0033291719001351

Kunkels, Y. K., van 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), e13898. https://doi.org/10.1111/psyp.13898

Pendleton, D. M., Sakalik, M. L., Moore, M. L., & Tomporowski, P. D. (2016). Mental engagement during cognitive and psychomotor tasks: Effects of task type, processing demands, and practice. International Journal of Psychophysiology, 109, 124–131. https://doi.org/10.1016/j.ijpsycho.2016.08.012

Polar Research and Technology. (2019, November 11). *Polar H10 Heart Rate Sensor System*. www.polar.com. Retrieved May 22, 2022, from https://www.polar.com/en/img/static/whitepapers/pdf/polar-h10-heart-rate-sensor-white-paper.pdf

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/10.1016/S0167-8760(97)00049-4

Riese, H., van Doornen, L. J. P., Houtman, I. L. D., & de Geus, E. J. C. (2004). Job strain in relation to ambulatory blood pressure, heart rate, and heart rate variability among female nurses. *Scandinavian Journal of Work, Environment and Health*, *30*(6), 477–485. https://doi.org/10.5271/sjweh.837

Rowlands, A., & Sargent, A. (2014). The ecg workbook (3rd revised). M & K Update. Retrieved 2022, from https://rug.on.worldcat.org/

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

Speer, K. E., Semple, S., Naumovski, N., & McKune, A. J. (2020). Measuring Heart Rate Variability Using Commercially Available Devices in Healthy Children: A Validity and Reliability Study. *European Journal of Investigation in Health, Psychology and Education*, *10*(1), 390–404. https://doi.org/10.3390/ejihpe10010029

Van Voorhees, E. E., Dennis, P. A., Watkins, L. L., Patel, T. A., Calhoun, P. S., Dennis, M. F., & Beckham, J. C. (2022). Ambulatory Heart Rate Variability Monitoring: Comparisons Between the Empatica E4 Wristband and Holter Electrocardiogram. *Psychosomatic Medicine*, *84*(2), 210–214. https://doi.org/10.1097/PSY.0000000000001010

Vrijkotte, T. G., van Doornen, L. J. P., & de Geus, E. J. C. (2000). Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. *Hypertension*, *35*(4), 880–886. https://doi.org/10.1161/01.hyp.35.4.880

Watford, T. S., Braden, A., & O'Brien, W. H. (2020). Resting state heart rate variability in clinical and subthreshold disordered eating: A meta-analysis. *International Journal of Eating Disorders*, *53*(7), 1021–1033. https://doi.org/10.1002/eat.23287