Validity and Reproducibility of the Empatica E4 and Polar H10 to Measure Heart Rate

Nienke C. Buist

S3390365

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: 2122_2a_45

Supervisor: dr. Mark Span

Second evaluator: prof. dr. Dick de Waard

In collaboration with: Theres Patzelt, Harmien Tamsma, Rover Willemars, Lisanne Zondag.

June 20, 2022

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Abstract

Wireless and valid ambulatory monitoring of heart rate, and heart rate variability would be advantageous in (non-)clinical and research settings because the wired laboratory electrocardiogram (ECG) is limited in its usability. Widely accessible and cheaper devices as the Empatica E4 watch, and the Polar H10 waistband were tested on their validity and reproducibility, comparing the commercial devices with the TMSi REFA amplifier. In a laboratory setting, effort-requiring cognitive and physical tests were performed, to be precise the Colour and Emotional Stroop task and the Psychomotor Vigilance task. R-tops and inter beat intervals (IBI) were measured by each device and were tested on their reproducibility. There proved to be a large correlation between the Polar H10 and the TMSi and therefore this waistband can be deemed as an outstanding alternative. The Empatica was very sensitive to movement and was therefore not able to measure enough IBI's. The findings concerning the Polarband are promising enough to perform long-term research with cardiovascular data. However, more research is needed in the context of generalisation, and to see whether it is possible to rule out the limitations of motion artefacts.

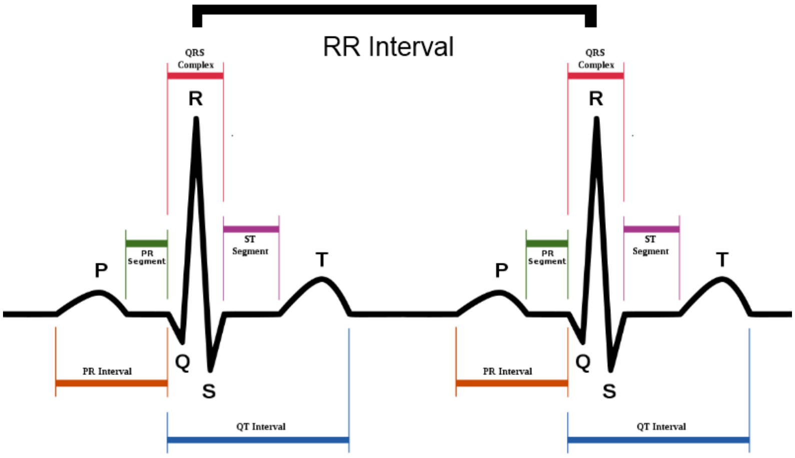
Keywords: electrocardiogram, Empatica E4, heart rate variability, inter beat interval, Polar H10

Validity and reproducibility of the Empatica E4 and Polar H10 to measure heart rate

The laboratory electrocardiogram (ECG) monitor is a well-known and reliable heart rate monitor, used to measure heart rate (HR) or heart rate variability (HRV). However, there are limits in its ease of use, for example skin irritation caused by the electrodes, the wires interfering motion, the costs, and the occurring problems in non-clinical setting where invasive HR measurement with an ECG monitor is not possible (Kunkels et al., 2019). To illustrate this, advances in ECG devices can give a better insight in the relationship between heart rate variability and mental health resilience (Perna et al., 2020). In short, the demand for more widely HR(V) measurements is growing and device innovations can be advantageous in (non-) clinical practice and scientific research.

Figure 1

The illustration of an ECG.



Note. This figure demonstrates cardiac electrical activity leading to an action potential with notable peaks of activity (the R-top in the QRS-complex). Retrieved from *International Conference on Advanced Computer Science and Information Systems* by Tawakal, I., Suryana, E., et al., 2012, Research Gate (https://www.researchgate.net/).

Heart rate measurement

A laboratory ECG monitor, the golden standard in measuring HR, measures heart rate through cardiac electrical activity. With each heartbeat, the concentration of the potassium, sodium, and calcium content changes through the flow of these molecules into, or out of the cardiac cells of the atria and ventricles. The electrical voltage across the membranes of the cells, the membrane potential, will change during this transport, and when high enough, will cause an action potential (Surawicz, 1987). This is illustrated by a peak in activity in the ECG, as shown in Figure 1 (Tawakal et al., 2012). In this action potential, depolarization and repolarization of the cardiac tissue produce waves that can be observed with an ECG. This depolarization phase is also known as the QRS complex (Ledezma & Altuve, 2019), in this paper mostly referred to as R-top. The repolarization phase is the return to the resting potential. The time between two R-tops (the RR-interval) is known as the inter-beat interval (IBI). Heart rate variability is a physiological marker that detects the difference in the ECG from beat-to-beat.

In a laboratory setting, the validity and reproducibility of two commercially available devices is tested, comparing their test results with an ECG monitor. The devices chosen to compare with are the Empatica E4 wristband and the Polar H10 waistband. These monitors should give access to unprocessed IBI-data. The Empatica E4 wristband has different sensors to measure sympathetic nervous system (SNS-) activity and HR (Empatica Inc., 2020), and should demonstrate, according to their website, remarkable results in measuring physiological signals. In this experiment we are interested in the photoplethysmography (PPG) sensor, a light sensor that measures the blood volume pulse (BVP) to determine the HRV or IBI's. According to research, the Empatica E4 performed excellent in assessing HR and that its test results are highly comparable to the VU University Ambulatory Monitoring System (VU-AMS), which is considered golden standard (Schuurmans et al., 2020). What should be noted here is that Empatica's goal is to detect beats of which it is certain, which can result in failed detections of IBI's, nevertheless this was not an issue in the latter study.

The Polar H10 Heart Rate Sensor monitors HR by capturing the full ECG trace (Polar Electro, 2022). Polar Electro (2022) considers this device as golden standard in wireless HR-monitoring. We will test whether these devices give the same results as our golden standard, because it is uncertain how exactly the R-tops in the commercial devices are measured and if the given measurements are reliable. The golden standard used in this experiment is the TMSi REFA amplifier. This device gives access to the raw ECG by electrodes attached to the body (TMSi, 2022). In contrast to the golden standard, the commercial monitors have a wider usability. Because they are both wireless, they can connect with other devices through Bluetooth, and thus can easily store data and send their measurements, and they are less likely to cause skin irritations (Schuurmans et al., 2020). Besides this, both devices are less expensive than the ECG monitor is, and therefore can be relevant in comparing the outcomes.

Experimental tests

During this experiment a set of stress-inducing and effort-requiring tasks are to be performed by the participants, while HR(V)- or IBI-data will be recorded. In response to the tasks, the SNS will be activated and elicits a physiological stress reaction, namely an increased heart rate. Three tasks are utilized to measure heart rate variability: the Psychomotor Vigilance Test (PVT), the Stroop task, and a physical task. The PVT is an attention task in which the speed of a response to a visual stimulus, in this study an alarm, will be measured. Research has found that during this task the ECG can detect changes in heart rate to estimate sleepiness (Chua et al., 2012). Researchers have also found that HRV is highly sensitive to cognitive processing and sustained attention, processes people may encounter during this task (Luque-Casado et al., 2016). The effects of executive demands in cognitive tasks reduce the HRV, and the time between two R-tops decreases as the heart beats faster. Taken all these factors into consideration, the PVT proves to be a suitable task to compare the monitors in this research with each other.

The Colour Stroop task is frequently used in testing psychophysical responses to stress among humans (Bali & Jaggi, 2015). During this task, the participant is asked to identify the colour of the word that is displayed, rather than the word itself that is named after a colour. For example, if the word ‘green’ is written in ‘blue’, identify it as ‘blue’. A significant cardiovascular response can be recognised by this traditional Stroop task with an ECG monitor (Boutcher & Boutcher, 2006). On the contrary, research found that the Stroop task was not able to elicit the expected HRV effect, suggesting that the setup was insufficient (Kunkels et al., 2019). Other research did find a significant response in the Colour Stroop task, but in a mental arithmetic challenge the response was higher (Brugnera et al., 2018).

In another version, the Emotional Stroop task, the words displayed are either pleasant or unpleasant, while the participant is still asked to identify the colour of the word. This emotional version can also be used to manipulate heart rate by initiating a cognitive process regarding the time necessary to respond to the colour and not the word (Dell’Acqua et al., 2021). Recent findings showed similar effects in the Colour Stroop and the Emotional Stroop task but questioned the close interaction between emotional processing of words and cognitive control of colours (Straub et al., 2021). In other words, they assumed that both tasks rely on different underlying mechanisms. From this follows the expectation that the Emotional Stroop task will lead to a higher HR than the Colour Stroop task, because the task is more emotionally loaded and therefore will lead to higher activation of the SNS. Consequently, taking into consideration these research findings, this study will implement the Colour Stroop task and an Emotional Stroop task, but with extra attention to the task setup to hopefully generate a significant cardiovascular response.

Finally, physical tasks such as sitting, and standing are included. Sitting is the resting position in between task sessions. Standing should give differences in heart rate, when comparing it with sitting. It is not practical to include walking tasks, because of the wires of the ECG that could interfere with moving, due to the limited space in the research room the experiment is taking place. Despite the limitations of our research, research on walking exercises is needed for valid research, because it is not observed enough whether the R-tops interfere with motion artefacts or not (Kunkels et al., 2019).

Validity and reproducibility

The importance of testing the Empatica E4 wristband and the Polarband H10 waist can contribute to make research easier and more comfortable, but the validity and reproducibility of those monitors have not been established yet. Therefore, the aim of this study is to find out whether the Empatica E4 wristband and/or the Polarband H10 waist are valid alternatives for measuring changes in heart rate variability. Based on the theoretical review, it is expected that the chosen tasks will cause a change in heart rate. It is predicted that our golden standard will detect this cardiovascular response and that the Empatica E4 and the Polarband H10 will detect similar changes. The websites of Empatica and Polar claimed promising results in their accuracy, as did some research. As it is unclear how the monitors measure the R-tops, there is a chance that HR(V) in these devices will provide different results than the golden standard, considering the devices will give different results in any case due to normal deviation.

Methods

Participants

A total of 44 adolescents participated in this study, of which 13 males and 31 females. The recruitment was done through a convenience sample: participants were acquired through the SONA research pool of the University of Groningen, which consists of first year psychology students who achieve credits to complete their propaedeutic year. Here, the self-selection sampling method was used by students deciding for themselves which studies they apply for in the research pool. All participants agreed to participate in this study with an informed consent, and there were no inducements other than compensation with SONA-credits. Incomplete data was omitted to minimize pre-processing.

Statistical analysis

Descriptive statistics provided the amount of IBI's measured, the mean, standard deviation, minimum, and maximum. Besides this, correlation tests between the TMSi and Empatica wristband, and the TMSi and Polar waistband were performed to calculate effect size values. After that, the Welch Two Sample t-test was performed to scrutinise the equality of the means. The percentages of differences were also calculated, with the TMSi REFA amplifier as norm value. Because high correlations and normal deviation were expected, a deviation from the REFA less than 0.02sec is seen as valid. One-way Analysis of Variance (ANOVA) tests were used to visualize differences between the Emotional and Colour Stroop task and between the sitting and standing blocks. When homogeneity was violated, the non-parametrical Kruskall-Wallis test had been used as post-hoc test. For processing and visualisation of the data, the programs JASP and R had been used.

Materials and apparatus

For the current study, three heart rate measurements were used to measure heart rate variability: the Empatica E4 wristband, the Polar H10 waistband, and the TMSi REFA amplifier (ECG). The Empatica and the Polarband sent their data through a Bluetooth connection and the ECG had a direct wire connection to the data acquisition computer. The OpenViBE Acquisition Server and OpenViBE Designer (ref) streamed the REFA data to the Lab Streaming Layer's (LSL). Custom build programs were made to stream the Polar and Empatica data (https://github.com/markspan/PolarBand2lsl). MATLAB and the custom-made program Alakazam (https://github.com/markspan/Alakazam) were used to generate a user interface for the ECG analysis.

The program OpenSesame version 3.3.11 (Mathôt et al., 2011) was used to manufacture the Stroop and PVT tasks. In the soundproof research room there were two computers, one desktop for the Stroop tasks and one laptop for the PVT. For the Stroop tasks, participants had to use a custom-made button box with four led-coloured buttons (blue, green, red, and yellow) to respond to the words on the screen. There were six blocks, of which three displayed emotional words and three displayed coloured words, starting each with a practice round consisting of 4-8 words. Each trial contained 40-160 words to which the participant had to respond in order to complete the trial. The emotional trials consisted of positive words only (e.g., happy, flower, celebration), negative words only (e.g., pain, bomb, explosion), and a mixed combination of words (a combination of several positive and negative words). The coloured trials consisted of a congruent trial (the colour corresponds to the word), a noncongruent trial (the colour does not correspond to the word), and a mixed trial (a combination of congruent and noncongruent words). A staircase algorithm in the Stroop task would increase and decrease the time-limit to react on the coloured words, depending on the response of the participant. Giving two consecutive correct answers would decrease the time-limit until the participant was not able to respond in the amount of time, what would cause an error response. Giving incorrect answers would in turn increase the amount of time.

For the PVT, several alarms went off during the experiment with random intervals between 60 and 240sec. Soundboxes were used to play the siren sound and the participant had to turn off the alarm as soon as possible by clicking the space bar. The laptop was approximately at arm’s length from the participant, so some movement was necessary to reach the space bar.

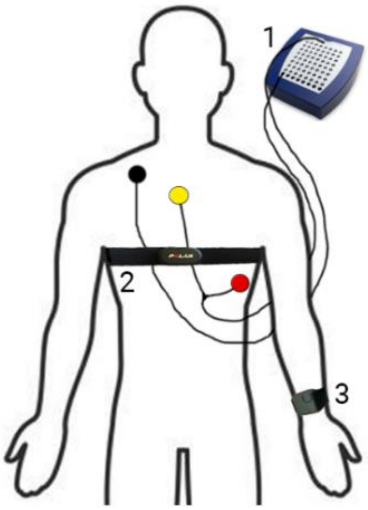
Procedure

Each participant went through the same procedure. In the pre-screen, participants had been asked whether they have a driver’s license, if they need optical help, and what their handedness is. The experiment took place in a research room at the University of Groningen.

The details of the participant will be treated with confidentiality, and it is emphasized that participation is voluntary and that the participant may stop at any time. The participants read the instructions and signed informed consent before the experiment started. After that, the researcher applied three electrodes on cleaned skin and put on the watch and the (slightly wet) waistband, for illustration see Figure 2. Thereafter, the researcher connected the devices to the lab recorder and the participant was connected to the ECG device with wires. When the participant was connected and ready, the researcher started the experiment.

Figure 2

Graphic illustration of the devices connected to the participant



Note. The devices numbered are 1) TMSi REFA amplifier, 2) Polarband H10, and 3) Empatica E4. Retrieved from Rover Willemars.

Heart rate variability was manipulated by the staircase algorithm, that could provoke extra frustration during the performance of the tasks. The alarm could also arouse a startle response. After the completion of the tasks, the researcher disconnected the participant, and the debriefing started to explain the goal of the experiment.

Of the 44 participants who initially participated in this study, 16 were excluded by reason of unusable data. This was mainly due to equipment failures, especially in the wristband, or excessive noise in the ECG trace that could not be removed.

Results

Table 1 shows the details of the sample (n = 28) used in the statistical analysis. The descriptive statistics of the IBI's obtained from each device are shown in Table 2. Remarkable is the amount of missing data points of the Empatica. The valid data measured by the Empatica was only 17.30% in comparison to the TMSi and the Polar H10. Table 3 elaborates the differences between the sitting, standing, and siren phase, indicating the effect of movement. The siren phase had a duration of 12sec, starting 2sec before and ending 10sec after the alarm went off. The Empatica measured a less IBI's than the Polar and TMSi.

Table 1

Characteristics of participants after exclusion

	Gender		**License**			**Vision aids**			**Handedness**	
	Female	*Male*	*Yes*	*No*	*Decline*	*No*	*Contacts*	*Glasses*	*Left*	*Right*
Total	20	8	11	16	1	21	5	2	2	26
%	71.43	28.57	39.29	57.14	3.57	75.00	17.86	7.14	7.14	92.86

Note. Participants were excluded when data was unusable.

Table 2

Descriptive statistics IBI's obtained from TMSi REFA, PolarH10, and Empatica E4

	Valid	**Missing**	**Total %**	**Mean**	**Std. Dev.**	**Minimum**	**Maximum**
TMSi IBI	44854	0	100	0.700	0.107	0.330	2.419
Polar IBI	44854	0	100	0.700	0.107	0.327	1.704
Empatica IBI	7758	37096	17.30	0.728	0.135	0.406	1.844

Note. Data measured with the Empatica is only 17.30% when TMSi (and Polar) is 100%.

Table 3

Descriptive statistics of the IBI's measured during the sitting, standing, and siren

		Valid	**%**	**Mean**	**Std. Dev.**	**Minimum**	**Maximum**
Sitting	TMSi	2270	100	0.738	0.136	0.492	1.401
	Polar	2276	100.26	0.739	0.137	0.327	1.599
	Empatica	1147	50.53	0.749	0.132	0.500	1.391
Standing	TMSi	2610	100	0.639	0.115	0.450	1.381
	Polar	2610	100	0.639	0.115	0.358	1.382
	Empatica	794	30.42	0.656	0.119	0.438	1.094
Siren	TMSi	2658	100	0.698	0.106	0.330	1.560
	Polar	2667	100.34	0.697	0.107	0.339	1.561

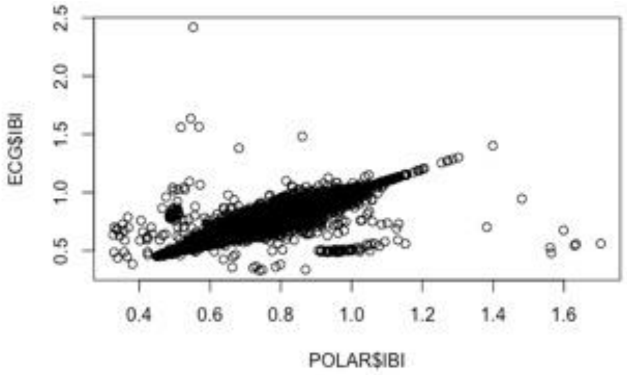
Empatica	502	18.89	0.741	0.138	0.500	1.281

Note. The Polar estimating a higher amount of IBI's than the TMSi may explained by a delay in the Polarband sending the data to the acquisition computer.

In compliance with the expectation, the golden standard and the Polarband showed a large correlation (*r*(df) = .942, *p* < .001) with a 95% confidence interval [.941, .943] on grounds of Pearson's correlation coefficient. For an illustration of this relation, see Figure 3. However, the correlation between the TMSi and the Empatica was small (*r*(df) = .114, *p* < .001), along with the correlation between the Polarband and the Empatica (*r*(df) = .111, *p* < .001). The results obtained from the Welch Two Sample t-test to compare the TMSi and Polarband displayed *t*(89706) = .004, *p* = .9972 with a 95% confidence interval [-.0014, .0014]. This is a nonsignificant result; however, the Welch test was for testing differences, stating a p-value equal to 0 as difference and equal to 1 as no difference.

Figure 3

Scatter plot showing the relation between the TMSi REFA amplifier and the Polar H10.



Note. Data derived from the PolarH10 is displayed on the X-axis and the TMSi on the Y-axis.

The IBI means found for the colour conditions were *M* = .718 for the REFA, *M* = .718 for the Polar, and *M* = .741 for the Empatica. For the emotional conditions the IBI means were *M* = .709 for the REFA, *M* = .708 for the Polar and *M* = .732 for the Empatica. No significant difference is found between the conditions for neither the TMSi (*F*(1,291) = .768; *p* = .382), the Polarband (*F*(1,291) = .903; *p* = .343), and the Empatica (*F*(1,210) = .266; *p* = .607).

The IBI means in the sitting phase in the REFA (*M* = .646), the Polar (*M* = .646), and the Empatica (*M* = .640) differed from the IBI means in the standing phase in REFA (*M* = .751), the Polar (*M* = .750) and the Empatica (*M* = .761). This difference between the standing and sitting phase, had found to be a significant result for the TMSi (*F*(1,74) = 20.91; *p* < .001), the Polarband (*F*(1,73) = 21.13; *p* < .001), and the Empatica (*F*(1, 66) = 21.95; *p* < .001). The Kruskal-Wallis test revealed a significance difference between standing and sitting in the TMSi (*H*(1) = 195.14; *p* < .001), the Polarband (*H*(1) = 195.14; *p* < .001), and the Empatica (*H*(1) = 18.429; *p* < .001).

Table 4 displays the descriptive statistics obtained per participant for elaborative details of the measurements of the devices. It is noteworthy that the Empatica measured few IBI's compared to the other devices and for P016 no IBI's were even registered. The testing outcomes of the Polar led to almost similar values as for the TMSi.

Table 4

Descriptive statistics obtained from each device per participant

	Device	Valid	Mean	Std. Dev.	Minimum	Maximum
P001	TMSi	1250	0.903	0.095	0.653	1.401
	Polar	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
	Polar	1308	0.819	0.105	0.553	1.078
	Empatica	0	-	-	-	-
P005	TMSi	1665	0.669	0.054	0.330	0.911
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	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
	Polar	1704	0.695	0.063	0.353	1.087
	Empatica	184	0.738	0.121	1.516	1.500

Note. For P004 the Empatica did not measure any data.

Discussion

This research attempts to contribute to the process of making heart rate measurement devices more accessible, user friendly and affordable. This is done by investigating the validity and reproducibility of two commercial devices in comparison to the ECG-equipment used in clinical settings.

The results indicate outstanding differences in accuracy between the Empatica E4 and the Polar H10, where the Polarband scored near the golden standard and the Empatica simply did not. Particularly, the correlation between the Empatica and TMSi REFA turned out to be small. The Empatica missed a large amount of measuring points, and besides this, during movement, it was sending hardly any data to the lab streamer. Moreover, in the case of one participant, the watch was not able to measure any IBI at all during the entire procedure. Furthermore, the Empatica measures different and generally higher means than the TMSi and Polarband.

The differences of amount of IBI's found in the standing and sitting conditions in the Empatica are supported by research. It concluded that the Empatica E4 was a practical and valid tool to measure heart rate, however, they noted that a large amount of IBI's (40%) were missing even during meditation and relaxation techniques that required little motion (Schuurmans et al., 2020). In the standing condition the Empatica was only able to measure

794 data points, 30.42% of what both the REFA and Polarband measured. However, in the sitting condition the Empatica was able to detect 1147 data points, 50.53% of the total from the golden standard and the Polar H10. This percentual difference implies that the Empatica watch is probably not resistant to movement. As a matter of fact, Empatica (2021) mentioned factors that may influence the EE4 PPG sensor, including movement and how the watch is placed on the wrist. Participants had to react as fast as possible by pressing the button box and could use both of their hands. Next to that, the participants could not rest their arms on the table in the standing condition, but during the sitting condition they could. These movements may have affected the amount of IBI measure points modulated by the watch. Besides that, the watch might have been worn too tight, or too loose and was therefore insufficient to detect the blood volume pulse and therefore the IBI's. Still, it was noteworthy that the watch could only measure 50% of the IBI's during the sitting phase, in which movement hardly occurred.

The Empatica did not detect the same amount of IBI's as the golden standard did, however, it was able to detect the significant difference between the standing and sitting condition. This finding suggests that the EE4 can, in fact, measure variability in heart rate, although it only measured half of the data.

Other research reported that short stressors, such as the startle response aroused by the alarm, could not be validly detected by the Empatica (van Lier et al., 2020), suggesting that the setup of this part of the experiment was not sufficient to detect similar changes in the different devices. In our research, the Empatica was able to detect 18.89% of the IBI's when the siren went off. This was a disappointing finding compared to the TMSi and Polarband. Again, this low percentage could be due to the participant moving while turning it off. We chose a wide duration (12sec) in which the participant had more than enough time to turn off the siren, however, this may not have been long enough. Another explanation could be that the Empatica was lacking in detecting startle responses.

Another problem in acquiring the Empatica data was that the IBI's could not be matched to the IBI's measured by the golden standard due to missing values. This was reinforced by the software we used limited us in compensating for this problem, as we were not expecting that the Empatica would send this minimal amount of data. As a solution for matching the IBI's the program KUBIOS could be used (Empatica Inc, 2020), but it was beyond our time capacity to experiment with this program. All together, these were problematical findings, especially when considering the Empatica was designed for doing research and the cost of it was approximately €1600 euro.

On the contrary, the Polarband had a much lower price, namely €90 euro, and was of higher robustness to movement in measuring IBI's. The differences in minimum and maximum values between the TMSi and the Polarband were likely due to outliers or motion artefacts and should be eliminated from the data set. However, unfortunately there was not enough time to do so. However, the differences in means were negligibly small, considering that these outliers did not cause a problem. As the Welch test was measuring differences, the results should be interpreted otherwise when interested in the equality of the data. With 95% confidence we can conclude that the Polarband significantly estimates the same values as the golden standard. Perhaps eliminating outliers will give even higher correlations, but this is still an essential outcome because it demonstrates that the Polar is an appealing alternative.

Another interesting finding was that we found no significant differences between the Stroop tasks in any of the devices. This indicates that the emotionally loaded words have no effect on heart rate in comparison with the coloured words. However, another explanation might be that the emotional manipulation was not strong enough. Thus, no conclusion can be drawn from the effects of emotional processing or cognitive control in colour recognition on cardiovascular responses based on our findings.

Limitations and recommendations

Some limitations occurred in the beginning of the experiment, regarding the use of the ‘huggable’ electrodes. We noticed that they were not continuously capable to adhere to the participant’s body during the entire experiment. We chose to use different electrodes to prevent more electrodes to fall off, as this could interfere with our data collection. The first subject to perform the experiment with the other electrodes was P016, and henceforth fewer participants were excluded from the data set. As mentioned before, at high levels of movement artefacts the devices could fail to measure data. This did not uniquely appear in the Empatica watch, but also in the Polarband and in the TMSi. Several participants were removed from the dataset before the analysis was performed, because that data was unusable. This could be due to wrongly placed electrodes, or to the attachment of the watch or waistband were attached to the body. It would be ideal if these problems will not occur in the future, thus giving special attention to the correctness of placement of the devices will be more beneficial.

For further research, it might be enriching to include stretching or squatting tasks, that are more doable in a research room with limited freedom of movement than tasks as walking or running. As squatting requires some effort, the participant’s heart rate should increase and beside that, it is possible to perform this while staying connected with wires and to remain in the same place. Next to that, it might be interesting to explore if the Polar is also resistant to high intensity movements. For the emotional processing part, it might be interesting to test different emotional tasks. Since we could not find an effect of the emotional manipulation in the Stroop tasks, it still might be possible while using stronger manipulation to find a significant effect.

We only conducted our research among psychology students approximately between 17 and 25 years old. We did not ask the age of the participants, but as we weren’t searching for heart rate variabilities in specific age groups but for the reproducibility of devices, this did not limit our findings. For generalisation of our findings, further research can implement different conditions, for example people of different age, health conditions or settings.

Conclusion

According to the data collected during this research, the Polar H10 is seen as a remarkable replacement for the golden standard, as it is a cheap, reliable, and an accessible heart rate measurement device. Besides that, it can be used in different (non-)clinical settings and research settings. The Empatica E4 however, should only be used in settings where there are low chances of movement artefacts, otherwise it is likely one loses a large amount of data. These conclusions are promising for measuring long-term cardiovascular data, nevertheless more research is needed in the context of high intensity tasks and for proper generalisation.

References

Bali, A., & Jaggi, A. S. (2015). Clinical experimental stress studies: Methods and assessment. *Reviews in the Neurosciences*, *26*(5), 555–579. https://doi-org.proxy-ub.rug.nl/10.1515/revneuro-2015-0004

Boutcher, Y. N., & Boutcher, S. H. (2006). Cardiovascular response to Stroop: Effect of verbal response and task difficulty. *Biological Psychology*, *73*(3), 235–241. https://doi-org.proxy-ub.rug.nl/10.1016/j.biopsycho.2006.04.005

Brugnera, A., Zarbo, C., Tarvainen, M. P., Marchettini, P., Adorni, R., & Compare, A. (2018). Heart rate variability during acute psychosocial stress: A randomized cross-over trial of verbal and non-verbal laboratory stressors. *International Journal of Psychophysiology*, *127*, 17–25. https://doi-org.proxy-ub.rug.nl/10.1016/j.ijpsycho.2018.02.016

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. https://doi.org/10.5665/sleep.1688

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

Empatica Inc. (2020). *Empatica E4 wristband.* Retrieved May 27, 2022, from https://www.empatica.com/en-eu/research/e4/

Empatica Inc. (2021). *Decoding wearable sensor signals: what to expect from your E4 Data.*

Retrieved June 17, 2022, from https://www.empatica.com/blog/decoding-wearable-sensor-signals-what-to-expect-from-your-e4-data.html

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. *Medical & Biological Engineering & Computing*, *57*(8), 1673–1681. https://doi-org.proxy-ub.rug.nl/10.1007/s11517-019-01990-3

van Lier, H. G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H. A., Vollenbroek-Hutten, M. M. R., Schraagen, J. M., & Noordzij, M. L. (2020). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. *Behavior Research Methods*, *52*(2), 607–629. https://doi-org.proxy-ub.rug.nl/10.3758/s13428-019-01263-9

Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology*, *113*, 83–90. https://doi-org.proxy-ub.rug.nl/10.1016/j.biopsycho.2015.11.013

Mathôt, S., Schreij, D., & Theeuws, J. (2011). Opensesame: An open-source, graphical experiment builder for social sciences. https://doi.org/https://doi.org/10.3758/s13428-011-0168-7

Perna, G., Riva, A., Defillo, A., Sangiorgio, E., Nobile, M., & Caldirola, D. (2020). Heart rate variability: Can it serve as a marker of mental health resilience? *Journal of*

Affective Disorders, *263*, 754–761. https://doi.org/10.1016/j.jad.2019.10.017

Polar Electro (2022). *Polar H10 Heart Rate Sensor.* Retrieved May 27, 2022, from https://www.polar.com/en/sensors/h10-heart-rate-sensor/

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), 1–11. https://doi-org.proxy-ub.rug.nl/10.1007/s10916-020-01648-w

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

Surawicz, B. (1987). Contributions of cellular electrophysiology to the understanding of the electrocardiogram. *Experientia*, *43*(10), 1061–1068. https://doi-org.proxy-ub.rug.nl/10.1007/BF01956040

Tawakal, I., Suryana, E., Noviyanto, A., Satwika, I. P., Alvissalim, S., Hermawan, I., M. Isa, S. & Jatmiko, W. (2012). Analysis of multi codebook GLVQ versus standard GLVQ in discriminating sleep stages. *International Conference on Advanced Computer Science and Information Systems*, 197-202. Retrieved May 25, 2022, from https://www.researchgate.net/figure/The-illustration-of-QRS-complex_fig1_260157313

TMSi (2022). *REFA amplifier.* Retrieved May 27, 2022, from https://www.tmsi.com/products/refa/