Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

Article

About the Accuracy and Problems of Consumer Devices in the Assessment of Sleep

Mohamed S. Ameen [1], Lok Man Cheung [1,2], Theresa Hauser [1], Michael Hahn [1], Manuel Schabus [1, 3, *]

[1] University of Salzburg, Department of Psychology, Laboratory for Sleep, Cognition and Consciousness Research, Hellbrunner Strasse 34, 5020 Salzburg.

[2] School of Psychology, University of Surrey, Surrey, United Kingdom, Stag Hill, Guildford GU2 7XH, UK.

[3] Center for Cognitive Neuroscience Salzburg (CCNS), Hellbrunner Strasse 34, 5020 Salzburg.

* Correspondence: manuel.schabus@sbg.ac.at; Tel: (+43) 66280445113

Abstract: Commercial sleep devices and mobile-phone applications for scoring sleep are gaining ground. In order to provide reliable information about the quantity and/or quality of sleep, their performance needs to be assessed against the current gold-standard, i.e. polysomnography (PSG; measuring brain, eye and muscle activity). We here assessed some commercially available sleep trackers, namely; a commercial activity tracker: Mi band (Xiaomi, BJ, CHN), a scientific actigraph: Motionwatch 8 (CamNTech, CB, UK), and a much used sleep application: Sleep Cycle (Northcube, GOT, SE). We recorded 27 nights in healthy sleepers using PSG and these devices. Surprisingly, all devices had very poor agreement with the gold standard. Sleep parameter comparisons revealed that specifically the Mi band and the sleep cycle application had difficulties in detecting wake periods which negatively affected the total sleep time and sleep efficiency estimations. However, all 3 devices were good in detecting the most basic parameter, the actual time in bed. In summary, our results suggest that, to-date; available sleep trackers do not provide meaningful sleep analysis but may be interesting for simply tracking times in bed. A much closer interaction with the scientific field seems necessary if reliable information shall be derived from such devices in the future.

Keywords: Wrist-worn devices; Sleep trackers; Activity trackers; Sleep classification; Polysomnography.

1. Introduction

Our knowledge about the structure and function of sleep is derived mainly from recordings that are done in sleep laboratories. In these recordings physiological activity is measured using polysomnography (PSG), which requires a combination of electroencephalography (EEG), electrooculography (EOG) and electromyography (EMG) data. Although these recordings contribute widely to our constantly expanding knowledge about sleep, their major drawback is that they do not mimic the habitual sleeping environment at home. The effect of the laboratory setup on sleep has been repeatedly addressed and several studies have highlighted differences in sleep parameters between at-home and laboratory sleep recordings. For instance, Portier and colleagues [1] compared several sleep parameters between a night of sleep in the laboratory versus a night of sleep at-home. Specifically, the authors reported a significant reduction in the amount of total sleep

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

time (TST) and time in bed (TiB) as well as the deterioration of the subjective experience of sleepers on the quality of their sleep in the laboratory as compared to at-home sleep. In a similar vein, home-based PSG recording showed higher sleep efficiency (SE) values than hospital-based PSG for the same participants [2]. These findings suggest that in-laboratory sleep does not accurately reflect habitual sleep (at-home sleep) and consequently might introduce some bias in the diagnosis of some sleep disorders. Indeed, it has been shown that home-based PSG recordings show better dissociation between healthy sleepers and insomnia sufferers than in laboratory settings [3]. Therefore, it is a priority in the field of sleep research and sleep medicine to develop better tools that can accurately and reliably measure sleep at home and in a vast amount of the general population.

Already in the last years we have witnessed a vast increase in the available consumer devices, i.e. sleep devices and mobile phone applications, which aim to assess and ultimately improve sleep. These devices might be of potential help to overcome the bias induced by the laboratory setting as they are supposed to assess sleep outside the laboratory with minimal effort for the end-user. However, it is essential to scientifically test and compare these devices against the "gold-standard" to ensure that such devices and applications do not provide random feedback to the naïve end-user and potentially backfire with "unintended effects on sleep beliefs and behaviors" [4]. Only the adherence of such devices and applications to the gold standard ensures reliability and validity and ethically justifies these new methods advertised by the industry.

The aim of this study was therefore to assess the performance of some of these readily used consumer devices which claim to monitor sleep and provide reliable information about sleep quality and sleep architecture night-by-night. Specifically we assessed sleep data from 2 devices: 1) a commercial activity tracker, the Mi band (v2, Xiaomi, BJ, CHN), 2) a scientific actigraph watch, Motionwatch 8 (CamNTech, CB, UK) as well as one readily used mobile phone application: the Sleep Cycle App (v3.0.1.2511-release; Northcube, GOT, SE). We compared the full set of sleep measures from these 3 platforms against our PSG gold standard and relied on semi-automatic sleep staging using the SOMNOlyzer 24X7 solution [5,6].

2. Materials and Methods

Study sample: For the study we recruited 19 healthy participants (13 females, mean age: 29±13. Range: 19-64 years). Participants arrived at the sleep laboratory of the University of Salzburg at 9pm. They were instructed about the procedure and the purpose of the experiment. After signing the consent forms they were given the Mi band (MB) and the MotionWatch (MW). After we confirmed that both the MB and the MW were recording we started with the PSG preparation. Before turning lights off and starting the PSG recording we started the Sleep Cycle application (SC) and placed the device next to the subject. Participants went to bed at around 11 pm and stayed in bed (TIB, time in bed) for approximately 8h (452.29 ± 81.78 min). Two of these participants had to be excluded from our analyses due to technical problems with the PSG recordings. Additionally, we recorded 8 ambulatory "home" PSG nights using an ambulatory EEG device together with the MB device and the SC application after the participants visited the lab for electrode placement; therefore, we had a total of 27 nights of PSG recordings. Due to other problems with some of the devices and applications we finally analyzed 21 nights for the MB, and 12 nights for the MW and the SC.

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

EEG Data Acquisition: For the nights spent in the laboratory, brain activity was recorded using high-density-EEG with a 256-electrode GSN HydroCel Geodesic Sensor Net (Electrical 478 Geodesics Inc., Eugene, Oregon, USA) and a Net Amps 400 amplifier. Additionally, we recorded electrocardiography (ECG), electromyography (EMG) and electrooculography (EOG) using bipolar electrodes. Ambulatory PSG was recorded using a 16-channel EEG, bipolar EMG and EOG using the AlphaEEG amplifier and NeuroSpeed software (Alpha Trace Medical Systems, Vienna, Austria).

Sleep Scoring: Our PSG was analyzed for sleep stages using the computer-assisted sleep classification system Somnolyer 24x7 as developed by the SIESTA group (The SIESTA Group Schlafanalyse GmbH., Vienna, Austria; [5,6]) and was following the revised standard criteria described by the American Association for Sleep Medicine (AASM, [7]). The derived sleep features and sleep stages serve as gold-standard for the rest of the analyses. Sleep staging for the SC application was realized via a simple image processing of the figures generated by the application; basically we discretized the SC illustrations into 3 sleep-wake states as suggested by the application in wake, light sleep and deep sleep (cf. Suppl. Material and Supplementary Figure S1 for more details).

Statistical analysis: The following five main sleep parameters were evaluated: i) sleep onset latency (SOL), ii) sleep efficiency (SE), iii) wake after sleep onset (WASO), iv) total sleep time (TST), and v) time in bed (TiB). SOL was defined as the difference between the start of the recording and the time when the participant actually fell asleep (i.e. the 1st N1 or "light sleep" epoch). SE (%) was defined as: (TST / TIB)*100. Importantly, measurements from all the devices were accurately synchronized to the start of the PSG recording. Correlations were computed non-parametrically using spearman correlations. For the MB device measurements we needed to calculate SOL, SE, and TST (as described above) manually and used WASO values as provided by the device. For the SC, we manually calculated SOL, SE, TST, and WASO (as described above) and used the time points provided by the application to calculate TiB. For the MW measurements, the sensitivity threshold was set to 20 activity counts and adjusted the lights off and lights on times according to sleep diaries as usually done for scientific actigraph measurements. SOL, SE, TST, WASO and TiB values are then used as provided by the MotionWare software (v1.1.20, empire Software GmbH, Cologne, Germany).

Bland-Altman plots were used to quantify the agreement between the PSG gold standard and the three consumer devices. The measured bias is defined as the mean of the difference between the two-paired measurements. That is, the further this value is from zero, i.e. the line of equality (difference = 0), the higher the error in the measurement. Spearman-correlations are used to illustrate systematic linear biases of the devices, and are reported at $p<.01$ (corrected for the 5 dependent sleep variables analyzed).

Epoch-wise comparison of sleep stages: For analyzing the epoch-by-epoch agreement of the gold-standard with the three consumer devices we always synchronized the recording start with the start of the PSG recording. In case one device started recording after the other (for example, PSG after SC or vice versa) we simply discarded the earlier epochs and started the analysis from the first epoch which was scored by both. As mentioned above we used the graphs provided by the MB

device and the SC application in order to divide sleep/waking into awake, light and deep sleep and in 30s epochs. For the PSG gold standard light sleep was defined as stages N1 and N2 while Deep sleep is defined as N3 stages. Importantly, we excluded PSG epochs which were scored as stage REM according to the AASM from the analysis as all 3 devices and applications provide no information about REM (or "dreaming") sleep. We report two main parameters for the epoch-wise agreement; sensitivity and positive predictive value (PPV). Sensitivity (in %) estimates the epoch-by-epoch agreement between the MB and SC with the gold standard by measuring the % of correct classifications (according to the PSG standard) per sleep stage (that is, for example labelling 79% of all light sleep detections by the PSG as "light sleep"). The positive predictive value (PPV), on the other hand, is the probability that the assigned state (by the device or app) is indeed that specific state in the gold standard (that is, for example only 41% of assigned "light sleep" epochs are actually light sleep epochs and no other sleep states). Cohen's Kappa (K) was used to assess the pairwise agreement between the devices. Epoch-wise analysis was computed in SPSS software (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.). Cohen's Kappa scores between 0.21–0.40 is often considered as fair, 0.41–0.60 as moderate, and 0.61–0.80 as substantial agreement according to Landis and Koch (1977, [7]).

3. Results

The mean values of the key features of sleep across all participants according to the PSG gold standard were 434.58 ± 95.83 minutes for the TiB, 370.12 ± 104.43 minutes for the TST, 84.08 ± 13.22% for the SE, 25.98 ± 19.35 minutes for the SOL and 39.08 ± 38.43 minutes for WASO. As a first analysis we simply checked whether the mean sleep values per participant and night correlate between the gold standard and the devices. For TiB we found good agreement, that is significant positive associations, of the gold standard values with the 3 consumer devices (MB: r= 0.72, p=0.0002; SC: r= 0.67, p=0.02; MW: r= 0.77, p=0.03). For TST we only found one moderately positive association for the MB device (r= 0.49, p= 0.02), while MW was the only device that showed a significant positive correlation for WASO time (r= 0.78, p= 0.02) (see supplementary figures S2-5). This low agreement is already surprising given that these are the simple associations of the mean values per subject, e.g., whether people taking longer to fall asleep in the case of SOL measurements (according to the PSG gold standard) also tend to fall asleep later according to the output of one of the consumer devices.

3.1. Bland Altman plots:

We used the Bland and Altman analysis to visualize the degree of agreement between the PSG gold standard and each of the 3 aforementioned devices/applications (cf. Figures 1-3). The most global key features of sleep, namely TIB and TST are depicted in Figure 1. Looking at the mean difference, there is only a slight bias towards over- or underestimating TIB (MB: 11.1±59.96 min, MW: 11.96 ± 38.48 min & SC: -5.17 ± 57.57min). However, the 95% confidence interval also indicates that for single cases the devices may still over- or underestimate TIB by an hour or more (cf. Figure 1.A). Mean TST is systematically overestimated by the MB by more than an hour (69.64 ± 67.43 min), and underestimated by the SC application by more than two hours on average (-139.67 ± 85.87 min) (Figure 1.B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2019 doi:10.20944/preprints201907.0303.v1

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

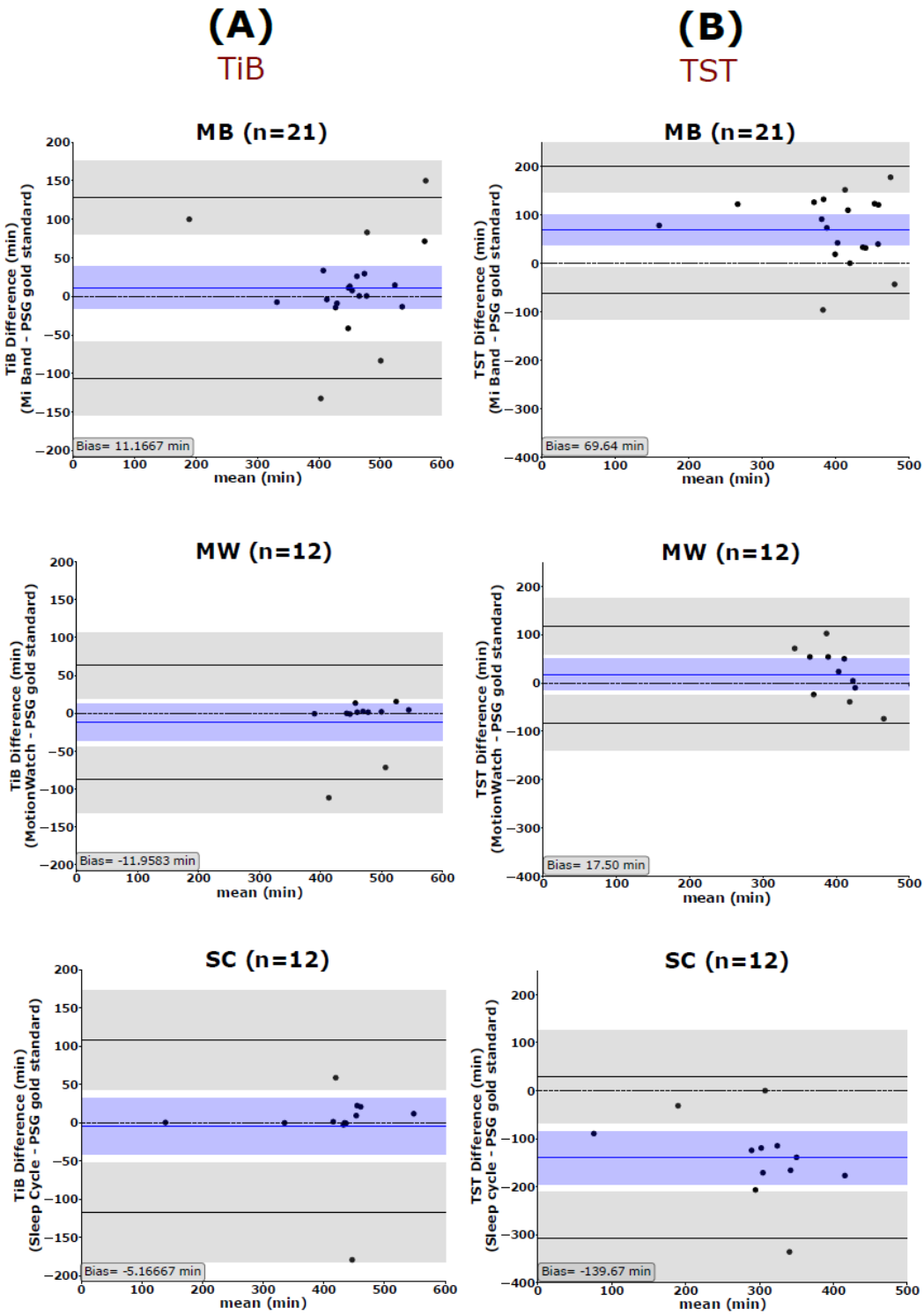

Figure 1. Bland Altman plots show the agreement of the MB, the MW and the SC with the PSG in measuring (A) TiB but not (B) TST. The blue horizontal line represents the mean difference between the two measurements and the shaded blue area represents the 95% CI of the mean difference. Black horizontal lines mark the 1.96SD from the mean and the grey shadings represent their 95%CI. The black dashed line is the line of equality (difference=0). TiB: time in bed, TST: total sleep time, MB: Mi Band, MW: MotionWatch and SC: sleep cycle application.

All 3 tested devices/applications showed inaccuracies in estimating SE, with the scientific MW giving the best results and no systematic over- or underestimation of SE (Figure 2). The mean differences indicate that the MB systematically overestimated SE (13.25%) whereas the SC application systematically underestimated (22.63%) SE. Interestingly, spearman correlations indicated that the MB linearly shows greater errors the worse the real PSG gold standard SE was, that is the MB has a strong bias towards quantifying SE better than it is.

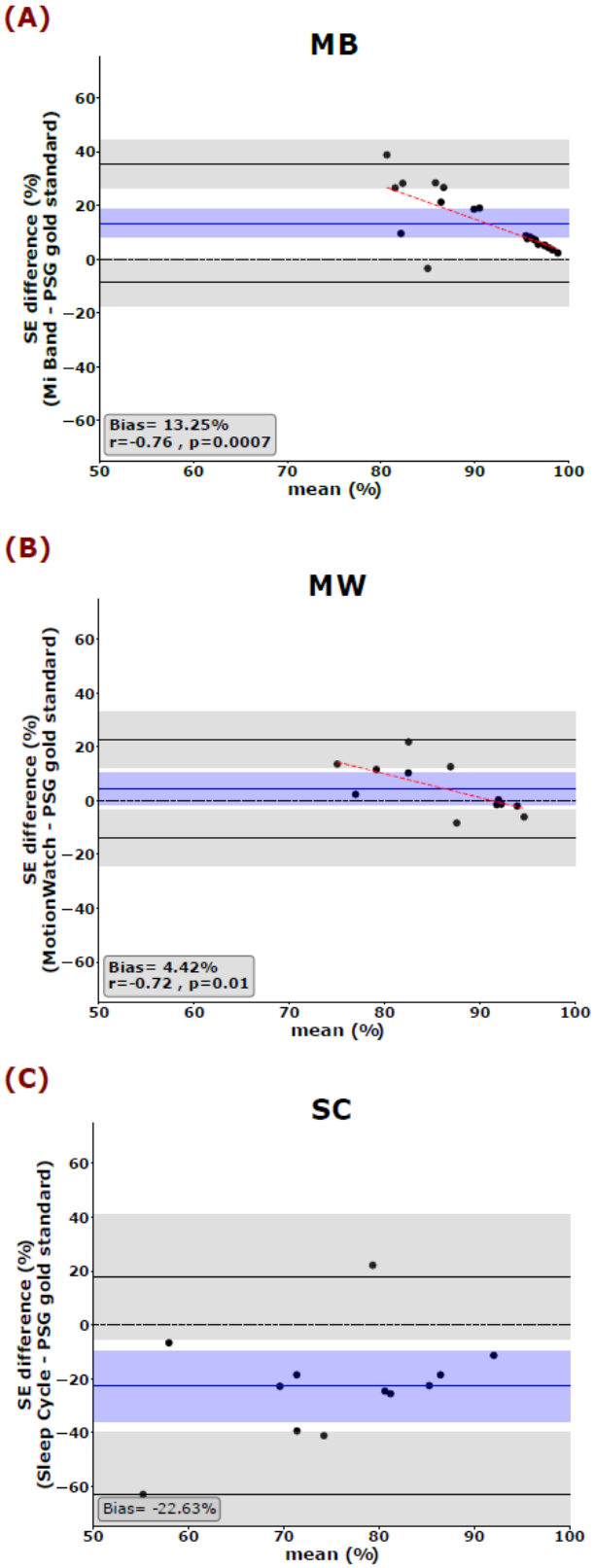

Figure 2. Bland Altman plots show that the PSG gold standard with the (A) MB, the (B) MW and (C) the SC in measuring sleep efficiency (SE). The blue horizontal line represents the mean difference between the two measurements and the shaded blue area represents the 95%-CI of the mean difference. Black horizontal lines mark the 1.96SD from the mean and the grey shadings represent their 95%CI. The black dashed line is the line of equality (difference=0) and the red dashed line represents the spearman-correlation between the difference and the average of the two measurements. SE: Sleep Efficiency, MB: mi band, MW: MotionWatch and SC: sleep cycle

Moreover, we observed a systematic error in the estimation of the WASO time by the MB and the SC but not the MW (Figure 3.A). While the MB device underestimates WASO (-33.57±42.84 min) the SC App overestimates WASO systematically (61.10 ± 49.90 min). In addition there is a linear trend in the data showing that the MB underestimates WASO time the more the longer actual WASO time gets. The mean difference of the 3 devices/applications to the gold standard is closer to zero for SOL, however it is to be noted that here also the range of possible values is much more limited (36.18 ± 38.37 min for the gold standard). Only the MB shows a linear trend with stronger underestimation of SOL the longer SOL actually was (in the gold standard) (Figure 3.B).

3.1. Epoch-wise agreement per sleep stage

Table 1 shows the overall and stage-wise agreement between the 30s-epochs scored by our PSG gold standard and both the MB device and the SC application. Note that the MW is disregarded in this respect as standard MW outputs do not provide (or claim to allow) sleep staging classifications. The overall agreement over all epochs from all subjects (16350 epochs for the MB, 11243 epochs for the MW and 9504 epochs for the SC) between the gold standard PSG scoring and the MB was relatively low (53.31% , k=0.14) and even lower for the SC device (46.34%, k=0.18). Table 1 also illustrates that the highest level of agreement for the MB was in determining light sleep (sensitivity = 70.6% and PPV = 57.8%) and the lowest sensitivity for the MB was for detecting wakefulness (sensitivity = 5.5%; PPV= 62.8%). Conversely, however, SC had moderate sensitivity in identifying awake epochs (sensitivity= 55.6%) and an unacceptable PPV value of 24.3%, meaning that only 24.3% of wake classified epochs are indeed wake according to the PSG gold-standard. On the other hand, SC had low sensitivity in detecting light sleep (40.9%) yet when it classified light sleep this was the true state in 61.2% of the cases (i.e., PPV= 61.2%). Moreover, for "deep sleep" classification we found very poor performance for the MB (sensitivity= 47.2%, PPV= 43.6%) and poor performance for the SC App (sensitivity= 52.0%, PPV= 53.0%).

Given this poor performance in correctly classifying sleep stages we then investigated the ability of these devices and the SC application to simply differentiate between sleep (light sleep, deep sleep or REM) and wakefulness and included the scientific MW device (whose software anyways only provides wake and sleep classes). Here we then found good overall agreement for the MB and the MW (>80%, cf. Table 2), and rather poor overall agreement for the SC App (65.9%). Kappa pairwise agreement indicates a "fair" agreement for the MW but poor agreements for the other two devices/apps. Specifically, the output shows that the MB and MW devices on the arm wrist are very good when only "sleep" detection is needed (MB: sensitivity= 99.5%, PPV= 86.8%; MW: sensitivity= 92.9%, PPV= 88.2%). Severe difficulties remain in assigning stage wake in all three devices/apps and therefore a proper estimation of overall sleep efficiency or sleep quality remains non feasible.

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

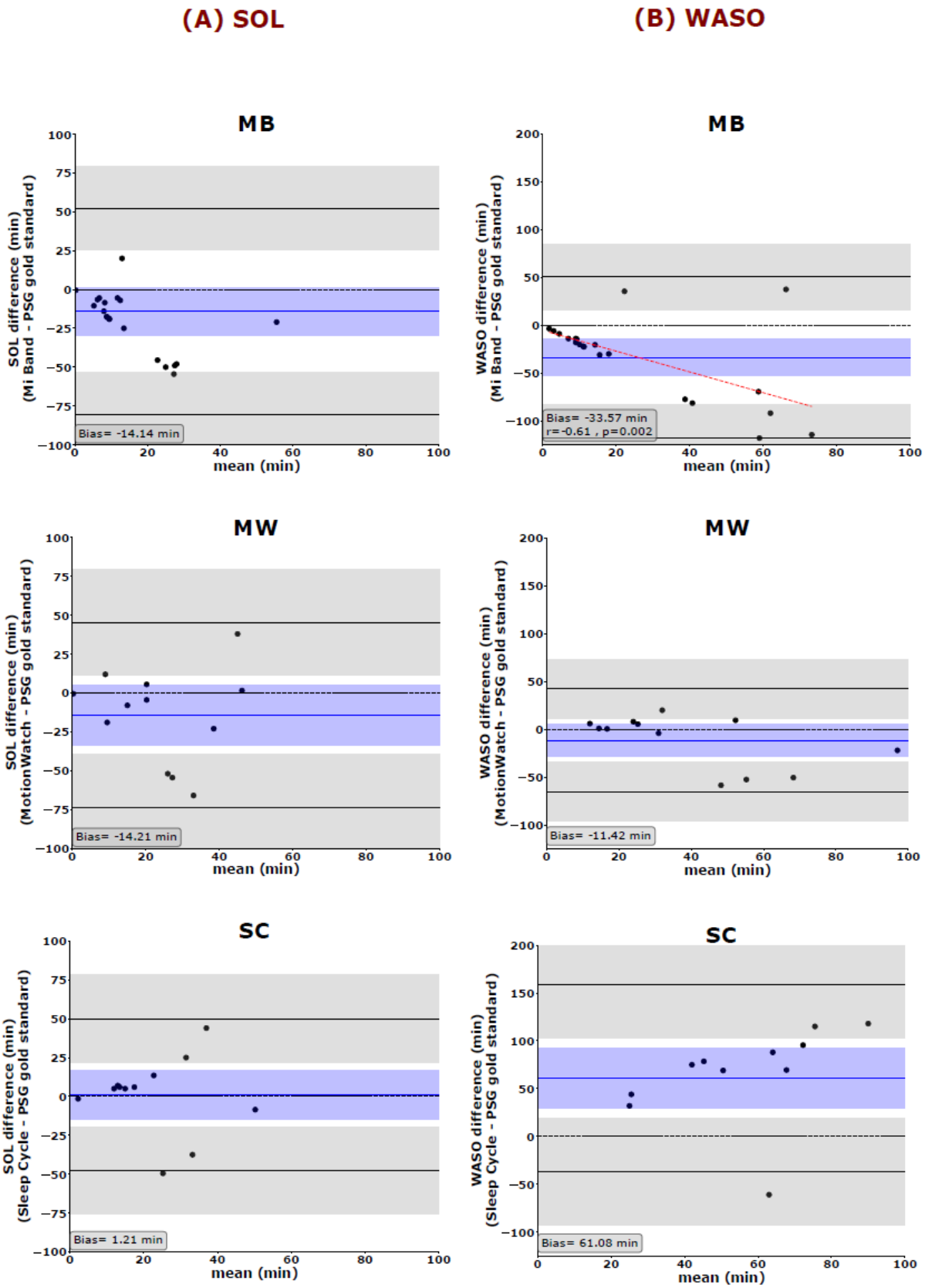

Figure 3. Bland-Altman plots of the SOL and WASO measurements showing differences between the PSG gold standard and the MB, the MW and the SC. The blue horizontal line represents the mean difference between the two measurements and the shaded blue area represents the 95%-CI of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 July 2019 doi:10.20944/preprints201907.0303.v1

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

mean difference. Black horizontal lines mark the 1.96SD from the mean and the grey shadings represent their 95% CI. The black dashed line is the line of equality (difference=0) and the red dashed line represents the spearman-correlation between the difference and the average of the two measurements. SOL: Sleep onset latency, WASO: wake after sleep onset. MB: Mi band, MW: MotionWatch and SC: sleep cycle application.

Although the SC application is as good as the wrist band devices in assigning "sleep" to an epoch, that is then the app is correct in 91.3% of these cases, it still misses a third of all sleep epochs (sensitivity= 67.4%). Importantly, however, in the case of the MB and the SC, OA increases while Kappa scores dropped when we pooled all sleep stages in one stage which likely indicates serious bias in the scoring algorithms of the MB device and the SC app.

Table 1. Percentages of agreement table for 3 stages sleep scoring (Awake/ Light sleep/ Deep sleep) between the gold standard (PSG) and the scoring of the MB and the SC.

	PSG gold standard		
	WAKE	**LIGHT SLEEP**	**DEEP SLEEP**
MiBand (MB) staging			
Wake			
% Sensitivity	**5.5**	**0.1**	**1.5**
% PPV	**62.8**	**4.7**	**32.6**
Light sleep			
% Sensitivity	**79.2**	**70.6**	**51.3**
% PPV	**18.9**	**57.8**	**23.2**
Deep Sleep			
% Sensitivity	**15.3**	**29.3**	**47.2**
% PPV	**7.5**	**48.9**	**43.6**
Sleep Cycle (SC) staging			
Wake			
% Sensitivity	**55.6**	**37.0**	**16.9**
% PPV	**24.3**	**61.1**	**14.7**
Light sleep			
% Sensitivity	**36.4**	**40.9**	**31.1**
% PPV	**14.4**	**61.2**	**24.4**

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

Deep Sleep			
% Sensitivity	**8.0**	**22.1**	**52.0**
% PPV	**4.1**	**42.8**	**53.0**

Devices/applications	**OA (%)**	**K**
mi-band		
MB	**53.31**	**0.14**
Sleep Cycle		
SC	**46.34**	**0.18**

The agreement is demonstrated by the means of the sensitivity (%) as well as the positive predictive value (PPV). The percentage of the overall agreement (% OA) as well as the Cohen's Kappa coefficient (K) is reported for each device.

Table 2. Percentages of agreement table for 3 stages sleep scoring (Awake/ Asleep) between the gold standard (PSG) and the scoring of the MB and the SC.

	PSG gold standard	
	WAKE	**SLEEP**
Mi Band (MB) staging		
Wake		
% Sensitivity	**5.5**	**0.5**
% PPV	**62.8**	**37.2**
Sleep		
% Sensitivity	**94.5**	**99.5**
% PPV	**13.2**	**86.8**

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

Sleep Cycle (SC) staging

Wake		
% Sensitivity	**55.6**	**32.6**
% PPV	**19.9**	**80.1**
Sleep		
% Sensitivity	**44.4**	**67.4**
% PPV	**8.7**	**91.3**

MotionWatch (MW) staging

Wake		
% Sensitivity	**37.5**	**7.8**
% PPV	**47.8**	**52.2**
Sleep		
% Sensitivity	**62.5**	**92.9**
% PPV	**11.5**	**88.5**

Device/Application	OA (%)	K
Mi Band		
MB	**86.54**	**0.08**
Sleep Cycle		
SC	**65.90**	**0.13**
MotionWatch		
MW	**83.42**	**0.33**

4. Discussion

In the present study we evaluated the ability of 2 readily used consumer devices and one application for detecting true wake and sleep epochs at night. We acquired data from a commercial activity tracker, the Xiaomi Mi Band v2 (Xiaomi, Beijing, China), 2) a scientific actigraph, the CamNTech Motionwatch 8 (CamNTech, Cambridge, UK) and a readily used mobile phone application for sleep assessment, the Sleep Cycle App (v3.0.1.2511-release; Northcube, Gutenberg, Sweden). We then compared these consumer devices to our PSG gold standard which was simultaneously recorded. Overall, we revealed that these devices have an alarmingly low accuracy

Peer-reviewed version available at Sensors 2019, 19, 4160; doi:10.3390/s19194160

in scoring sleep in three categories (that is, wake, light and deep sleep) with overall agreements between 46.34% for the SC application and 53.02% for the wrist worn MB. If we tested only for the correct classification in two classes, that is wake and sleep, the devices of course performed better with an overall agreement of 65.90% for the SC, 84.69% for the MB and 81.33% for the scientific MW device. Kappa coefficients however indicate only poor agreement with the PSG gold standard for the MB and the SC (0.172 and 0.186 respectively) indicating serious bias in the scoring algorithms of the MB device and the SC app.

We showed that all devices and applications had high accuracy in estimating the most global sleep parameter, namely TiB. This makes these devices a helpful tool for objectively measuring the time spent in bed at home rather than relying solely on subjective measures such as daily sleep diaries. Especially in the case of the MW we need to note that we adjusted the start and the end of the recordings to the PSG gold standard (as usually done using additional sleep diaries) which might overestimate the fidelity of the MW device in measuring TiB. Nevertheless, our MW results are consistent with those reported in previous literature [8,9].

Only for TiB correlational analysis also showed significant positive correlations between the gold standard and all 3 sleep trackers. This raises the question of whether the faulty estimation of values such as TST, SE, WASO or SOL are due to a priori knowledge of these sleep trackers of the amount of time the average person actually sleeps or needs to fall asleep. If such information is included in the algorithms and outputs of the consumer devices, this would explain why the largest errors occur primarily for “non-average” sleep profiles and nights. However, to date this argument remains speculative as all tested devices do not allow raw data access or are black boxes when it comes to their staging algorithms. Similarly, when comparing the agreement between the MB and SC with the PSG gold standard for 3 sleep-wake classes (light sleep, deep sleep, and wake) as compared to 2 classes (sleep vs. wake) we found the expected increase in the OA yet a drop in the Kappa scores. Especially for the MB, looking at the sensitivity scores we observed extremely low sensitivity in detecting wakefulness (5.5%) and a very high sensitivity in detecting sleep (99.5%). That is, by assigning “sleep” to basically every epoch the device also cannot miss sleep epochs, yet it of course strongly overestimates sleep and has a vast amount of false alarms for stage “sleep”. Although the MB was the least sensitive between all the 3 devices and applications, it had the highest precision in scoring wakefulness (PPV: 62.8% for the MB, 47.8% for the MW and 24.3% for the SC). That is, the MB does not score awakenings from sleep unless they are almost unmistakable. This indicates a strong bias of the MB algorithm (as observed in the very low kappa values: 0.08; [10]), which again raises the question if such a biased output can be of any benefit to the end-user.

Regarding the other key parameters evaluated, our results raise serious doubts whether such consumer devices and applications can to-date provide any reliable information about sleep-related health issues. Especially the revealed misjudgment in estimating key features of sleep such as SE, SOL and WASO are worrisome as they are important diagnostic criteria for quantifying clinically relevant bad sleep and sleep disorders such as insomnia [11]. On the contrary, by providing such inaccurate information these consumer devices might even run risk to contribute to worse sleep and life quality as end-users may be concerned by the sometimes negative output highlighting bad nights of sleep [4].

Comparing the devices and SC app, our results suggest that wrist worn devices (MB and MW) tend to have better a performance than mobile phone applications (SC) in measuring the key

features of sleep. This might be attributed to the fact that these devices have direct contact with the body and hence are more accurate in capturing changes in physiological activity that accompanies sleep of an individual and are therefore also more resilient to the environmental factors such as noise or movement from the bed partner, child, pet or loud neighbors.

One very important drawback of the sleep trackers not mentioned yet is their inability to provide any information about REM or "dreaming" sleep. Due to an inherent absence of needed measurements for quantifying REM sleep (that is, most importantly eye movements via EOG and brain activity via EEG) the devices on the market today cannot provide the full spectrum of sleep even if the algorithms and sensors would be considerably improved. The incorporation of additional sensors such as an eye or brain electrode might add substantially to the ability of these devices to track and score sleep more accurately and in the long-run similar to a professional polysomnography in the sleep laboratory.

An inherent limitation of our evaluation study is that most of our analysis need to build upon the simple (graphical) outputs of the devices in form of plots provided for the end-user. For the tested MB device and SC app there is no way to access the raw data. We therefore needed to come up with a way to quantify the data and extract information that can be analyzed statistically (for details see suppl. material). The MW device is a scientific device out of the price range of the usual consumer and allows raw data access. Interestingly this device is likely the most accurate device tested and yet its software only provides two outputs, sleep and wake classes, as it does not claim to be able to classify sleep more fine-grained than that (as compared to consumer devices including the MB and SC).

In summary, the currently available consumer devices for sleep tracking do not provide reliable information about one's sleep. However, devices of that kind could be very promising tools for tracking sleep outside the laboratory in the future given that they adhere more to the scientific standards of sleep staging and analysis. Moreover, by refining their algorithms or even adding sensors these devices might be able to reliably monitor and classify sleep across its full range from wakefulness to light sleep, deep sleep, and "REM" dreaming sleep.

Author Contributions: conceptualization, M.S. and M.A.; methodology, M.A. and M.S.: software, L.C. and M.A.; validation, M.A. and M.S.; formal analysis, M.A. and L.C; investigation, M.A, L.C., T.H. and M.H. resources, M.S.; data curation, M.A. and L.C.; writing—original draft preparation, M.A.; writing—review and editing, M.S. and M.A; visualization, L.C., M.A. and M.S.; supervision, M.S.; project administration, M.S.; funding acquisition, M.S."

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Sleep scoring procedure of the sleep cycle application output hypnogram. Figure S2: Scatter plot showing the positive correlations between the time in bed measurement of the PSG gold standard and those of the Mi Band, MotionWatch and Sleep Cycle. Table S1: Spearman correlation results between the PSG gold standard and the Mi Band, the MotionWatch and the Sleep Cycle application for the key sleep parameters. Table S2: Epoch-wise agreement (Sleep/wake) while discarding epochs scored as REM by the PSG gold standard. Table S3: Epoch-wise agreement comparison between 3 stage scoring (Wake/light Sleep/Deep sleep) and 2-stage scoring (Sleep/wake).

Funding: This research received no external funding.

Acknowledgments: the authors would like to thank Ass.-Prof. Dr. rer. nat. Kerstin Hödlmoser, Monika Angerer and Frank van Schalkwijk for their support throughout the whole process.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Portier, F., Portmann, A., Czernichow, P., Vascaut, L., Devin, E., Benhamou, D., … Muir, J. F. (2000). Evaluation of Home versus Laboratory Polysomnography in the Diagnosis of Sleep Apnea Syndrome. American Journal of Respiratory and Critical Care Medicine, 162(3), 814–818. https://doi.org/10.1164/ajrccm.162.3.9908002
2. Bruyneel, M., Sanida, C., Art, G., Libert, W., Cuvelier, L., Paesmans, M., … Ninane, V. (2011). Sleep efficiency during sleep studies: results of a prospective study comparing home-based and in-hospital polysomnography. Journal of Sleep Research, 20(1pt2), 201–206. https://doi.org/10.1111/j.1365-2869.2010.00859.x
3. Edinger, J. D., Fins, A. I., Sullivan, R. J., Marsh, G. R., Dailey, D. S., Hope, T. V., … Vasilas, D. (1997). Sleep in the Laboratory and Sleep at Home: Comparisons of Older Insomniacs and Normal Sleepers. Sleep, 20(12), 1119–1126. https://doi.org/10.1093/sleep/20.12.1119
4. Baron, K. G., Abbott, S., Jao, N., Manalo, N., & Mullen, R. (2017). Orthosomnia: Are Some Patients Taking the Quantified Self Too Far? Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine, 13(2), 351–354. https://doi.org/10.5664/jcsm.6472
5. Anderer, P., Gruber, G., Parapatics, S., Woertz, M., Miazhynskaia, T., Klösch, G., … Dorffner, G. (2005). An E-Health Solution for Automatic Sleep Classification according to Rechtschaffen and Kales: Validation Study of the Somnolyzer 24 × 7 Utilizing the Siesta Database. Neuropsychobiology, 51(3), 115–133. https://doi.org/10.1159/000085205
6. Anderer, P., Moreau, A., Woertz, M., Ross, M., Gruber, G., Parapatics, S., … Dorffner, G. (2010). Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 × 7. Neuropsychobiology, 62(4), 250–264. https://doi.org/10.1159/000320864
7. American Academy of Sleep Medicine. (2007). The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. Westchester, IL: American Academy of Sleep Medicine, 23.
8. Kushida, C. A., Chang, A., Gadkary, C., Guilleminault, C., Carrillo, O., & Dement, W. C. (2001). Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. Sleep Medicine, 2(5), 389–396. https://doi.org/10.1016/S1389-9457(00)00098-8
9. Sadeh, A. (2011). The role and validity of actigraphy in sleep medicine: An update. Sleep Medicine Reviews, 15(4), 259–267. https://doi.org/10.1016/j.smrv.2010.10.001
10. Flight, L., & Julious, S. A. (2015). The disagreeable behaviour of the kappa statistic. Pharmaceutical Statistics, 14(1), 74–78. https://doi.org/10.1002/pst.1659
11. Schutte-Rodin, S., Broch, L., Buysse, D., Dorsey, C., & Sateia, M. (2008). Clinical Guideline for the Evaluation and Management of Chronic Insomnia in Adults. Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine, 4(5), 487–504.