HHS Public Access

Author manuscript

Chronobiol Int. Author manuscript; available in PMC 2022 July 01.

Published in final edited form as:

Chronobiol Int. 2021 July ; 38(7): 1010–1022. doi:10.1080/07420528.2021.1903481.

Performance of Fitbit Charge 3 against polysomnography in measuring sleep in adolescent boys and girls

Luca Menghini, MS[1,2], **Dilara Yuksel, PhD**[1], **Aimee Goldstone, PhD**[1], **Fiona C Baker, PhD**[1,3], **Massimiliano de Zambotti, PhD**[1,*]

[1]Center for Health Sciences, SRI International, Menlo Park, CA, USA

[2]Department of General Psychology, University of Padova, Padova, Italy

[3]Brain Function Research Group, School of Physiology, University of the Witwatersrand, Johannesburg, South Africa

Abstract

We evaluated the performance of Fitbit Charge 3™ (FC3), a multi-sensor commercial sleep-tracker, for measuring sleep in adolescents against gold-standard laboratory polysomnography (PSG). Single-night PSG and FC3 sleep outcomes were compared in thirty-nine adolescents (22 girls; 16-19 years), 12 of whom presented with clinical/subclinical DSM-5 insomnia symptoms (7 girls). Discrepancy analysis, Bland-Altman plots, and epoch-by-epoch analyses were used to evaluate FC3 performance. The influence of several factors potentially affecting FC3 performance (e.g., sex, age, body mass index, firmware version, and magnitude of heart rate changes between consecutive PSG epochs) was also tested. In the sample of healthy adolescents, FC3 systematically underestimated PSG total sleep time by about 11 min and sleep efficiency by 2.5%, and overestimated wake after sleep onset by 9 min. Proportional biases were detected for "light" and "deep" sleep duration, resulting in significant underestimation of these parameters for those participants having longer PSG N1+N2 and N3 durations, respectively. No significant systematic bias was detected for sleep efficiency and sleep onset latency. Epoch-by-epoch analysis showed sleep-stage sensitivity (average proportion of PSG epochs correctly classified by the device for a given sleep stage) of 68% for wake, 78% for "light" sleep, 59% for "deep" sleep, and 69% for rapid eye movement (REM) sleep in healthy sleepers. Similar results were found in the sample of adolescents with insomnia symptoms. Body mass index was positively associated with FC3-PSG discrepancies in wake after sleep onset (R^2 = .16, p = .048). The magnitude of the heart rate acceleration/deceleration between consecutive PSG epochs was an important factor affecting FC3 classifications of sleep stages. Our results are in line with a general trend in the literature, suggesting better performance for the recently introduced multi-sensor devices compared to motion-only devices, although further developments are needed to improve accuracy in sleep stage classification and wake detection. Further insight is needed to determine factors potentially

*Corresponding Author. Massimiliano de Zambotti • SRI International, 333 Ravenswood Avenue, Menlo Park, CA, 94025 • Tel: (650) 859-2714; Fax: (650) 859-2743 • massimiliano.dezambotti@sri.com; maxdeze@gmail.com.

Disclosure of interest. The authors report no conflict of interest related to the current work. MdZ and FCB have received research funding unrelated to this work from Ebb Therapeutics Inc., Fitbit Inc., International Flavors & Fragrances Inc., and Noctrix Health, Inc. The authors declared no agreements with Fitbit Inc. that could bias the results of this research in any way.

affecting device performance, such as accuracy and reliability (consistency of performance over time), in different samples and conditions.

Keywords

Wearable Sleep Trackers; Consumer sleep technology; Accuracy; Adolescence; Fitbit; Insomnia

Introduction

The use of consumer sleep-tracking technology (CST) including wristbands, rings, and smartwatches able to measure sleep and other behaviors (e.g., physical activity) is rapidly expanding for both research and clinical applications, providing an opportunity to advance understanding of sleep and its role in health and disease. CST adoption clearly outpaces scientific support for its use, calling for a greater understanding of this technology, its potential, and limitations. Ultimately, accuracy and reliability of CST are key factors to guarantee high-quality, trusted, and meaningful large-scale longitudinal CST data (de Zambotti et al. 2019a; Depner et al. 2019; Khosla et al. 2018).

The use of CST could be particularly relevant in investigating the biopsychosocial changes occurring in adolescence, such as the progressive shifting in adolescents' bioregulatory sleep processes (e.g., circadian phase delay, decline in slow wave sleep activity), and the interacting psychosocial factors (e.g., early school start times, academic pressure, bedtime autonomy, electronic media use) (Carskadon 2011; Colrain and Baker 2011). Adolescence is also a critical period during which sleep disturbances (e.g., insomnia disorder) and sleep-related mental disorders (e.g., major depressive disorder) frequently emerge (de Zambotti et al. 2018b). Despite having a prevalence of up to 18% in 16-18-year-olds, which is comparable to other major psychiatric disorders (e.g., anxiety and substance use), insomnia in adolescents is under-recognized, under-diagnosed, and under-treated. This under-recognition is partly due to the challenge of distinguishing insomnia from the normal developmental changes occurring in sleep regulation and the social and environmental constraints adolescents face that impact their opportunity for sleep (for a review, see de Zambotti et al. 2018a). In this context, CST may be helpful in identifying patterns of dysfunctional sleep over multiple nights in adolescents, discriminating between poor and good sleepers, and distinguishing poor sleep from normal developmental changes.

CST could serve as a low-cost and time-efficient alternative to polysomnography (PSG), the gold standard for sleep evaluation (de Zambotti et al. 2019a). Moreover, the newer generation of multi-sensor CST integrate accelerometry-based motion signals and photoplethysmography (PPG)-based heart rate (HR) and heart rate variability (HRV) measures to characterize sleep composition (epoch-by-epoch sleep staging, i.e., "light", "deep", and rapid-eye-movement sleep) and physiology (e.g., sleep autonomic function, respiration, SpO2), in addition to sleep quantity (sleep/wake patterns), providing more detailed information than what is available from research-grade actigraphy (de Zambotti et al. 2020).

Although CST has already been used to track sleep in adolescents (e.g., de Zambotti et al. 2016; Kuula et al. 2019), the evaluation of its accuracy in such populations is scarce, which is a critical barrier in CST adoption for scientific and clinical purposes (de Zambotti et al. 2019a; Depner et al. 2019; Khosla et al. 2018). Among the few CST devices tested in adolescents, the motion-based trackers, Fitbit Charge HR (de Zambotti et al. 2016), Jawbone UP (de Zambotti et al. 2015), and Polar Electro Oy (Pesonen and Kuula 2018) showed high sensitivity in detecting sleep and low specificity in detecting wake, a pattern typical of standard actigraphs (Sadeh 2011). To our knowledge, only two studies evaluated the sleep-stage classification accuracy of multi-sensor CST devices (Fitbit Alta HR and ŌURA ring) in adolescents, reporting systematic underestimation of N3 sleep and inconsistent estimations of N1+N2 and rapid eye movement (REM) sleep duration (de Zambotti et al. 2019b; Lee et al. 2019). Interestingly, two studies reported CST underestimation of sleep time and overestimation of nocturnal awakening time in adolescents (Lee et al. 2019; Pesonen and Kuula 2018), a pattern opposite to that reported by most CST validation studies conducted in healthy adults (de Zambotti et al. 2019a).

The current study aimed at evaluating the performance of Fitbit Charge 3™ (FC3, Fitbit, Inc.), a recent model of the popular Fitbit Charge family, in measuring sleep against gold-standard PSG in a sample of adolescents. Fitbit devices are among the most widely used and investigated CSTs (https://www.fitabase.com/research-library/). Studies evaluating previous Fitbit devices in various populations (mainly adults) showed, in general, an increase in the accuracy of multi-sensor sleep-staging compared to motion-based-only models (Haghayegh et al. 2019b). Based on recent recommendations from our group and others (de Zambotti et al. 2019a; Depner et al. 2019), standardized guidelines for assessing CST performance were applied via open-source R-based functions (available at https://github.com/SRI-human-sleep/sleep-trackers-performance), covering the main analytic steps for CST validation (Menghini et al. 2020). We also evaluated the role of several factors potentially affecting CST performance (see de Zambotti et al. 2020), including demographics (sex, age, body weight), average HR (*is FC3 accuracy different in those individuals with higher/lower HR?*), and clinical factors (presence of insomnia), in addition to device features (firmware version). Finally, we explored the CST rationale of using HR data in addition to motion to discriminate sleep/wake states and sleep stages (see de Zambotti et al. 2020, 2018b, 2019a). HR is expected to vary across sleep stages, being higher in wake and REM sleep compared to NREM sleep (de Zambotti et al. 2018d), and PPG-derived HR data are assumed to be used by FC3, along with HRV data (according to https://help.fitbit.com/) to classify sleep stage transitions (see also Beattie et al. 2017; Matar et al. 2018). However, the rationale behind this has been rarely (if ever) empirically investigated. Here, we evaluated the relationship between the magnitude of epoch-by-epoch (EBE) HR fluctuations detected by FC3, with the corresponding EBE agreement between FC3 and PSG (*is FC3 more accurate when greater HR reactivity is detected across sleep stage transitions?*). We expect that “true” shifts between different stages (based on PSG EBE transitions) would be more easily detected by FC3 when accompanied by larger HR changes.

Materials and methods

Participants

The sample included thirty-nine post-pubertal Junior and Senior (10-12th grade) high-school students (age: 16-19 years; 22 girls). Participants were recruited from the San Francisco Bay Area local high schools and community, as part of a larger study evaluating insomnia, sleep, and cardiovascular health in adolescence. The experimental protocol was conducted in accordance with accepted international ethical standards (Portaluppi et al. 2010), and the study was approved by the SRI International Institutional Review Board. All adult participants consented to participate, with minors providing written assent along with consent from a parent/legal guardian.

All participant had an in-lab initial visit to evaluate eligibility, including a structured clinical interview for DSM-5 (American Psychiatric Association 2013) disorders. All participants were free from severe mental disorders (e.g., Major Depressive Disorder, Post-Traumatic Stress Disorder) and medical conditions (e.g., Heart Diseases, Diabetes, Seizures) and were not currently using medications known to affect sleep (e.g., hypnotics) and/or the cardiovascular system (e.g., antihypertensives). Based on clinical interview, twelve participants (seven girls) reported insomnia symptoms (“difficulty initiating or maintain sleep despite an adequate opportunity to sleep, accompanied by significant daytime dysfunction, not due to the presence of another sleep disorder, mental or medical condition, or substance use”), with four of them meeting criteria for insomnia disorder and the remaining having sub-clinical insomnia symptoms. None of the participants had traveled across time zones within the prior month. Of the female participants, six healthy sleepers and 4 insomnia sufferers were taking hormonal contraceptives. None of the participants had breathing-related or leg-movement-related sleep disorders, as confirmed by in-lab clinical PSG. Demographic and other information are reported in Table 1.

Laboratory procedures

The study was conducted at the Human Sleep Research Laboratory at SRI International. After an adaptation night (i.e., first night used to get familiar with the laboratory setting and to exclude the presence of other sleep disorders, such as sleep-disordered breathing), participants underwent an additional non-consecutive standard PSG night during which they wore a FC3 on their wrist. On that night, upon arrival at the lab, the recording sensors were attached, and participants engaged in quiet activities (e.g., watching TV, reading a book) until bedtime. Lights-off and lights-on times were self-selected by participants, based on their regular sleep schedules. Electronic media use was not allowed after lights-off. When ready for bed, the PSG recording (time 0) was started at a time corresponding to a rounded FC3 time (hh:mm:00) to allow better PSG-device synchronization (de Zambotti et al. 2019a; see also supplemental material S1). Devices were removed upon awakening and data were synchronized via the Fitbit mobile App. All participants slept in sound-attenuated and temperature-controlled bedrooms. Participants were instructed to refrain from consuming caffeinated beverages and alcohol after 03:00h of the recording day, and recent alcohol/drug use was evaluated via breath alcohol (S75 Pro, BACtrack Breathalyzers) and urine drug test

(10 Panel iCup drug test, Instant Technologies, Inc.) at lab entry. Girls were scheduled irrespective of menstrual cycle phase.

Standard PSG sleep assessment

Standard laboratory PSG sleep assessment, i.e., electroencephalography (EEG; $F_{3/4}$, $C_{3/4}$, $O_{1/2}$ referred to the contralateral mastoid; 512 Hz sampled), submental electromyography, and bilateral electrooculography, was performed according to the American Academy of Sleep Medicine (AASM) guidelines (Berry et al. 2020). Recordings were performed using the Compumedics Grael® HD-PSG system (Compumedics, Abbotsford, Victoria, Australia). Sleep records were double-scored in 30 s epochs (wake, N1, N2, N3 and REM sleep) by experienced scorers (inter-rater reliability ≥ 91%; Cohen's kappa = .89 ± .20).

Fitbit Charge 3 sleep data collection

FC3 (Fitbit, Inc.) is a commercially available multi-sensor wristband. It tracks several health-related indices including fitness measures (e.g., number of steps), exercise (e.g., running, walking), cardiac measures, sleep, daily calorie consumption. According to Fitbit (https://help.fitbit.com/), the sleep-tracking feature uses a combination of indices extracted from optical PPG (HR and HRV) and motion sensors to process time spent awake and asleep (by discriminating “light”, the equivalent of PSG N1 + N2 sleep, “deep”, the equivalent of PSG N3 sleep, and REM sleep) (for details about rationale and use of a multi-sensor approach to sleep staging, see de Zambotti et al. 2018c, 2019a; Matar et al. 2018).

The device connects to mobile platforms via Bluetooth technology and data are managed via a dedicated App. All devices were connected to Fitabase (Small Steps Labs LLC.), a research-oriented data management platform enabling continuous data collection, data monitoring and easy data aggregation from multi-devices. From Fitabase, individuals’ FC3 sleep data (including timestamp and sleep stage classification) were exported in 30 s epoch resolution, while HR data were exported at 1 s resolution and then averaged in 30 s epochs to match the sleep epoch-by-epoch data resolution.

The device was placed on the participants’ dominant hand by the research staff (about an inch above the wrist bone, with the back of the device in contact with the skin), following the manufacturer’s instructions. The dominant hand was used to avoid interfering with other research-grade photoplethysmography sensors (not discussed in the present work) that were placed on the non-dominant arm. Based on previous studies of wrist actigraphy (e.g., Driller et al. 2017), no substantial differences were expected between wrists in terms of sleep measures. FC3 data were collected between Feb 2019 and March 2020. Different firmware versions were used between participants across the progression of the study, i.e., 1.49.45 (14 participants), 1.60.39 (9 participants) and 1.63.5 (19 participants). Since it is not clear whether the FC3 sleep-tracking algorithm reflects updates in firmware versions, the firmware version as well as the time of data collection (i.e., day of the year, included to cover any other potential change in the system that is not notified to the user) were recorded and explored as factors in the analyses (see Depner et al. 2019).

Data processing and statistical analyses

FC3 performance was evaluated by using a standardized framework for testing the performance of CST introduced by Menghini et al. (2020). The analyses were performed separately for the group of healthy sleepers (n = 27) and the group of adolescents with insomnia symptoms (n = 12), using R 3.5.1. (R Development Core Team 2018). The full analysis report and the script used for data analysis are publicly available from https://github.com/SRI-human-sleep/CST-performance (Supplemental material S1).

Discrepancy analysis and Bland-Altman plot—The following PSG and FC3 sleep metrics were computed separately for each participant, based on EBE data recorded between lights-off and lights-on: total sleep time (TST, min), sleep efficiency (SE, %), sleep onset latency (SOL, min), wake after sleep onset (WASO, min), “light”, “deep”, and REM sleep duration (min). For each sleep measure, group-level discrepancies were expressed in terms of systematic bias (mean FC3-PSG difference) ± 95% limits of agreement (LOAs, i.e., the limits within which most differences are expected to lie). Based on recently published recommendations on the analysis of CST performance (Depner et al. 2019; Menghini et al. 2020), discrepancies were also visualized via a modified version of the Bland-Altman plots (Bland and Altman 1999), in which FC3-PSG differences are plotted against PSG values. The significance of FC3 over- and underestimations was determined based on the 95% confidence intervals (CI) of the bias.

Linear regression was used to test proportional biases (i.e., increasing/decreasing bias over the size of measurement) and heteroscedasticity (i.e., relationship between differences dispersion and size of measurement) for each sleep measure. When a proportional bias was detected (i.e., $p < .05$), discrepancies were modeled as a function of the size of measurement (expressed by PSG-derived measures). Similarly, when heteroscedasticity was detected LOAs were modeled as a function of the size of measurement (for details, see Menghini et al. 2020).

Epoch-by-epoch analysis—EBE agreement between FC3 and PSG sleep stage classification was evaluated by averaging the error matrices computed for each participant. That is, we computed the number of epochs classified in each stage (i.e., “light”, “deep” and REM sleep, and wake) by the two methods, and we reported the average proportion of correct (sensitivity) and incorrect (specificity) FC3 classifications over PSG classifications (proportional error matrix), with the corresponding SD and 95% CI. Following de Zambotti et al. (2019a) and Depner et al. (2019), the classical definitions of sensitivity (i.e., ability to correctly classify sleep epochs) and specificity (i.e., ability to correctly classify wake epochs), reflecting a binary classification of sleep/wake patterns (0 = wake, 1 = sleep), have been updated to meet the need of describing non-binary sleep stages classification (e.g., 0 = wake, 1 = “light”, 2 = “deep”, 3 = REM). Thus, sensitivity and specificity have been calculated separately for wake, “light”, “deep”, and REM sleep. Sensitivity represents the ability of FC3 to correctly classify a given PSG stage (wake, “light”, “deep”, or REM), while specificity is the ability of FC3 to correctly classify all the other PSG stages (see Menghini et al. 2020, for the operationalization of these definitions). The Cohen’s kappa (Cohen 1960), quantifying from 0 to 1 the proportion of classification agreement that is not

due to chance, and the prevalence-adjusted bias-adjusted kappa (PABAK) coefficient (see Byrt et al. 1993) were computed for both wake/sleep classification and each sleep stage.

Evaluation of factors potentially affecting CST performance—In the main group of healthy sleepers, multiple linear regression was used to explore several factors potentially affecting device performance (see de Zambotti et al. 2020). Absolute FC3-PSG discrepancies (min) and the agreement between FC3 and PSG epoch-by-epoch classification (0 = misclassification, 1 = agreement) were considered as the main performance metrics and used as outcome variables. For each of the absolute FC3-PSG discrepancies in TST, SE, SOL, WASO, "light", "deep", and REM sleep duration, linear regression models were specified with the following predictors: sex (male, female), age (years), BMI (kg m^{-2}), firmware version (1.49.45, 1.60.39, 1.63.5), time of data collection (number of days since the first PSG recording of the study, i.e., beginning of the data collection period). EBE agreement was modeled using logistic mixed-effects regression (LMER) with the same predictors, and a random intercept was included to account for individual variability. The average HR (bpm) measured by the FC3 between lights-off and lights-on was included as a further predictor of EBE agreement.

For each model, we assessed the underlying assumptions and the presence of influential cases (Nieuwenhuis et al. 2012), and we visually inspected the bivariate distributions of each predictor and outcome variable. To evaluate the contribution of each potential factor, each was hierarchically included to a null model (with outcomes being regressed only on intercept and residual variance). Only predictors showing stronger evidence than the null model (as indexed by the Aikake information criterion weight, AICw; see Wagenmakers & Farrell 2004) were included in the following models and statistically tested, with a level of significance set at $p < .05$.

Finally, we explored the relationship between the FC3 accuracy in classifying EBE transitions (e.g., light → light, light → wake) and the corresponding changes in FC3 HR. For each couple of consecutive epochs, the EBE transition was considered as accurately classified only if both epochs were equally scored by FC3 and PSG (see Table 2). Since HR sensitivity to 'true' shifts between different stages is a necessary condition to justify its use in sleep stage classification (see de Zambotti et al. 2018d), a first set of LMER models was used to initially evaluate FC3 HR differences across PSG sleep stages. Then, a second set of LMER models was specified to predict EBE transition agreement by the corresponding absolute percent change in FC3 HR. The type of PSG EBE transition (i.e., between PSG epochs classified with the same stage vs. different stages) and its interaction with FC3 HR changes were included as additional predictors in the models. Based on our hypothesis (i.e., higher likelihood of classifying shifts between different stages in EBE transitions associated with greater HR changes), we expected lower EBE transition agreement for consecutive epochs classified with the same stage by the PSG (e.g., light → light) but associated with larger HR changes compared to cases associated with smaller HR changes, and vice versa.

Results

Discrepancy analysis and Bland-Altman plots

Group-level discrepancies are summarized in Table 3, and the corresponding Bland-Altman plots are shown in Figure 1. Two healthy sleepers (both females) were identified as influential cases (see Nieuwenhuis et al. 2012), and, thus, they were excluded from the analyses of TST, SOL, SE, and WASO due to extreme PSG values and FC3-PSG differences (more than two SD distant from the sample mean, with both participants showing SE < 70% not attributable to detected technical failure in the recordings), whose inclusion led to false positives in bias significance. One girl with insomnia symptoms was excluded from the analysis of SOL for the same reason (see supplemental material S1 for details, in which outliers are graphically highlighted and additional analyses are performed by also including these cases).

In the sample of healthy sleepers, FC3 systematically underestimated TST and SE, whereas it overestimated WASO (see Table 3). A significant proportional bias was detected for “light” and “deep” sleep duration, with higher discrepancies (underestimation) for cases with longer duration of PSG-derived measures (i.e., N1 + N2 duration longer than about 225 min, and N3 duration longer than about 80 min) compared to cases with lower measures. No systematic biases were detected for REM sleep duration and SOL, with the former showing relatively wide LOAs, and the latter showing tighter LOAs for cases with shorter PSG SOL compared to cases with longer SOL (i.e., heteroscedasticity). Similar trends were shown in the group of adolescents with insomnia symptoms (see Figure 1), with FC3 underestimating TST and SE (but only for cases with higher PSG SE), and overestimating WASO with about twice the bias found in healthy sleepers. Contrary to what found for the main group, no proportional biases were found for “light” and “deep” sleep durations, with the latter being systematically underestimated by FC3.

Epoch-by-epoch analysis

Table 4 reports the mean proportion of correctly (diagonal) and incorrectly classified epochs (upper and lower triangle) over the total number of epochs classified in each stage by the PSG (i.e., proportional error matrix), whereas the matrix reporting the absolute number of epochs in each condition (i.e., absolute error matrix) is included in the supplemental material S2. In the sample of healthy sleepers, FC3 showed a better sensitivity to “light” and REM sleep detection (about 75-80%) compared to “deep” sleep and wake detection (about 60-70%), with about 40% of N3 epochs and 20% of wake epochs being erroneously classified as “light” sleep. On average, 20% of PSG REM epochs were also misclassified as “light” sleep. “Deep” sleep is the stage for which FC3 showed the lowest sensitivity. By combining sleep stages, FC3 shows an overall sensitivity of 94.75 ± 2.70%. Adolescents with insomnia symptoms showed a similar pattern of results, although sensitivity to wake was slightly higher (about 80%) compared with the group of healthy sleepers (about 70%). Again, the lowest FC3 sensitivity was found for “deep” sleep. Overall, sensitivity to sleep was 94.19 ± 1.78%.

The estimated kappa and PABAK coefficients for sleep/wake and sleep stages classifications are also shown in Table 4. Both coefficients indicate higher agreement between FC3 and PSG in classifying wake and REM sleep compared to both “light” and “deep” sleep. Wake and REM sleep showed the largest discrepancies between kappa and PABAK, with the latter reaching .83-.86 and .73-.74, respectively. Kappa and PABAK coefficients were similar in adolescents with and without insomnia symptoms.

Factors potentially affecting CST performance

The inclusion of BMI in the model was associated with a higher AICw compared to the null model only when predicting absolute FC3-PSG discrepancies in TST (AICw = .67) and WASO (AICw = .76), although the estimated parameter was significant only in the latter, indicating larger WASO discrepancies in adolescents with higher BMI compared to those with lower BMI (*Coeff.* = 1.21 (.58), R^2 = .16, $t(23) = 2.09$, $p = .048$). In contrast, BMI was not significantly associated with TST discrepancies (*Coeff.* = 1.01 (.55), R^2 = .13, $t(23) = 1.82$, $p = .08$). None of the remaining factors (i.e., sex, age, firmware version, and time of data collection) was associated with stronger evidence (higher AICw) than the null model for predicting absolute discrepancies in any of the considered outcomes. Similarly, averaged nocturnal FC3 HR did not show substantial relationships with FC3 performance (AICw = .40, R^2 = .002). Results were similar when statistically controlling for the PSG measures of each sleep parameter (see supplemental materials S1 and S3 for details).

Magnitude of the heart rate acceleration/deceleration between consecutive PSG epochs—Higher FC3 HR was found in PSG epochs classified as wake compared to “light” ($\chi^2(3) = 3{,}843.93$, $p < .001$; *Coeff.* = 8.14 (.13), $t = 63.37$) “deep” (*Coeff.* = 6.90 (.14), $t = 48.61$), and REM sleep epochs (*Coeff.* = 5.78 (.14), $t = 40.01$), with no significant differences between sleep stages (see Figure 2a). Similar results were found by matching FC3 HR with FC3-based epoch classifications (see supplemental material S1 for details). Higher FC3 HR was found in girls compared to boys ($\chi^2(1) = 7.16$, p = .007, AICw = .99; *Coeff.* = 7.37 (2.61), $t = -2.67$). 196 epochs (0.82%) and 203 transitions between consecutive epochs (0.84%) were excluded from the analysis due to missing FC3 HR data.

Most EBE transitions were accurately classified by FC3 (67.03 ± 8.21%) and consisted of consecutive epochs scored with the same stage by the PSG (91.75 ± 2.26%), with “light → light” being the most frequent transition (47.04 ± 7.19%). Most transitions between equally scored PSG epochs (e.g., light ↔ light) were accurately classified by FC3 (71.41 ± 8.16%), whereas the opposite pattern was found for transitions between differently scored PSG epochs (e.g., wake → light), with only 17.89 ± 7.59% of them being correctly classified by FC3 (the number of epochs in each transition and the corresponding absolute FC3 HR change are reported in supplemental material S4).

EBE transition agreement was significantly predicted by absolute FC3 HR changes to PSG EBE transitions ($\chi^2(1) = 249.09$, $p < .001$), by the type of PSG EBE transition ($\chi^2(1) = 1{,}904.05$, $p < .001$), and by their interaction ($\chi^2(1) = 71.27$, $p < .001$, AICw = 1.00). As shown in Figure 2b, higher HR changes predicted lower agreement in classifying transitions between PSG epochs classified with the same stage (*Coeff.* = −.03 (.004), OR = .97, $t =$

−8.49), whereas the opposite pattern (greater HR changes predicter better agreement) was found for PSG transitions between different stages (*Coeff.* = .06 (.007), OR = 1.07, $t = 8.74$). Overall, EBE transition agreement was lower in PSG transitions between different stages compared to transitions between the same stage (Coeff. = −2.85 (.08), OR = .06, $t = -34.82$) (see also Supplemental material S4).

Discussion

CST is increasingly seen as an opportunity to advance sleep and circadian research. CST can be particularly useful in populations like adolescents, when continuous sleep monitoring is needed to capture change and variability in an individuals' sleep patterns and behaviors. A need for further data validation to promote better use and understanding of CST performance and outcomes in clinical and research settings has been advocated (de Zambotti et al. 2019a; Depner et al. 2019; Khosla et al. 2018). Here, we evaluated the performance of FC3 in measuring sleep in adolescents with and without insomnia symptoms. Our results indicate that FC3 systematically underestimated TST and SE, and overestimated WASO, whereas it underestimated PSG N1+N2 ("light") and N3 ("deep") sleep duration in participants/nights with longer time in that PSG-defined sleep stage compared to participants/nights showing shorter durations (proportional bias). Sensitivity to sleep stages varied from 60-70% for "deep" sleep and wake detection to 75-80% for "light" and REM sleep detection.

Our results are in line with the general trend highlighted for CST performance, suggesting an increasing accuracy for the newer generation of CST (de Zambotti et al. 2019a, Haghayegh et al. 2019b), possibly due to the implementation of a multi-sensor approach (integration of motion and PPG-derived data to quantify sleep and wake duration) (de Zambotti et al. 2020), and advancement in algorithm refinement. Indeed, over recent years several studies suggested better agreement in detecting sleep/wake patterns between PSG and multi-sensor CST devices than between PSG and motion-based devices (e.g., Chinoy et al. 2020; Haghayegh et al. 2020b; Pesonen and Kuula 2018). In the present study, the absolute biases we observed for TST (about 11 min), SE (about 2.5%), and WASO (about 9 min) were in the lower tail of bias distributions shown by motion-based Fitbit devices (i.e., 7-67 min for TST, 2-15% for SE, 6-44 min for WASO) (Haghayegh et al. 2019b), and in line with previous attempts to use Fitbit accelerometry-based and PPG-derived features for staging sleep (e.g., Beattie et al. 2017). Also, we observed relatively high accuracy in classifying sleep/wake patterns, with an overall sensitivity to sleep approaching 95% (vs. 87-99% for motion-based devices) and a sensitivity to wake >65% (vs. 10-52%) (Haghayegh et al. 2019b). Compared to recent studies reporting sleep-stage sensitivity of CST devices (Chinoy et al. 2020; Cook et al. 2019; Haghayegh et al. 2020b), our results suggested a slightly better performance of FC3 than other devices in detecting "light" and REM sleep, but not "deep" sleep.

To our knowledge, only two studies assessed the performance of multi-sensor CSTs in adolescents (de Zambotti et al. 2019b; Lee et al. 2019). In those studies, both devices underestimated "deep" sleep duration, similarly to the FC3 performance shown in our study, particularly for adolescents with longer PSG N3 duration. Moreover, in both those previous

studies and in our study, “deep” sleep was the least accurately classified stage, whereas sleep-stage sensitivity was higher for “light” and REM sleep, similar to findings reported in adult populations (e.g., Beattie et al. 2017; de Zambotti et al. 2018a). It is unclear whether CST is approaching a plateau when combining motion and PPG data for sleep staging.

Our results are also in line with previous studies showing different performance behavior between adolescents and adults, with CST overestimating WASO in the former (Lee et al. 2019; Pesonen and Kuula 2018) and generally overestimating TST in the latter (de Zambotti et al. 2019a). Sleep overestimation is a typical limitation of actigraphs, at least partially associated with a misclassification of motionless wake as sleep (low specificity) (de Zambotti et al. 2020). In contrast, overestimation of WASO might be associated with other factors, including “light” epochs being misclassified as wake (accounting for about 10% of “light” sleep classification here and in Lee et al. 2019), or contextual factors, such as sleep opportunity. For instance, Lee and colleagues (2019) manipulated the length of time in bed and found different patterns of discrepancies depending on sleep opportunity, with “light” sleep being underestimated in adolescents with longer sleep opportunity and overestimated in adolescents with shorter sleep opportunity.

With regards to LOAs (i.e., expressing the limits within which most differences between FC3 and PSG measures are expected to lie), the absence of standard criteria to establish their acceptability in sleep assessment implies some difficulties in their interpretation. For instance, Werner et al. (2008) arbitrarily defined satisfactory agreement between sleep measures as LOAs < |30 min|, a cut-off criterion that would classify as unsatisfactorily the wide LOAs we found for TST, REM sleep duration, and for cases with shorter “light” and longer “deep” sleep durations. Alternatively, Haghayegh et al. (2020) suggested providing estimates of the “minimal detectable change” (i.e., the smallest change in a given measure detected by a method that exceeds measurement error), computed as one-half the difference between the upper and lower LOAs. In our study, this would be 30.67 min, 6.87%, 30.32 min, and 42.19 min for TST, SE, WASO, and REM sleep duration, respectively, whereas it would depend on the size of measurement for SOL, ‘light’ and ‘deep’ sleep duration. However, the interpretation of these values depends again on the definition of a priori criteria of “minimal clinical important change”, the specific application, and the target population of the device under assessment. Future research efforts should be made to standardize the acceptability of LOAs. What is certain is that excessively wide LOAs would imply a poor performance even when the bias is nonsignificant.

These results highlight the importance of evaluating CST performance in specific populations. The adolescents’ increasing preference for later bedtimes as they get older (Colrain and Baker 2011), combined with factors such as early school start time and high homework load, might lead to changes in TST, possibly playing a role in CST accuracy in relation to age. A relationship between adolescents’ age and the performance of a CST (Jawbone UP) was found by de Zambotti and colleagues (2015). In their sample, TST was overestimated in adolescents 12-14 y of age and underestimated (with higher absolute discrepancies) in adolescents older than 16 y of age. Similarly, other studies highlighted greater discrepancies, especially in TST, SE, and WASO, in adolescents compared to school-age (6-12 y) and preschool children (3-5 y) (Meltzer et al. 2015; Toon et al. 2016). In

contrast, such a relationship was not found in our study, possibly due to the narrow age range (16-19 y) or differences in performance between motion-based (e.g., Jawbone UP) and multi-sensor (FC3) sleep-tracking algorithms.

In our study, the only potential confounder shown to play a role in device performance was BMI, with higher bias in WASO for adolescents with higher BMI compared to adolescents with lower BMI. This result might be associated with inaccuracies in HR-driven sleep classification due to sensor placement in individuals with large wrist circumferences. Indeed, BMI and wrist circumference were highly correlated and both were predictive of reduced HR accuracy in both commercial (Shcherbina et al. 2017) and research-grade PPG wristbands (Menghini et al. 2019). Even the presence of insomnia symptoms did not play a substantial role in device performance, with a few differences mainly concerning the size of bias in WASO (about twice that found in healthy sleepers) and the classification accuracy for wake epochs (80% vs.70% of healthy sleepers). This result is partially in contrast with previous studies highlighting larger biases for insomnia disorder patients than for healthy sleepers. For instance, Kang et al. (2017) reported a bias in TST and SE that was more than double in the insomnia than in the control group. In addition to considering differences in terms of target population (adolescents vs. adults) and device used (FC3 vs. Fitbit Flex), this inconsistency should be interpreted in light of a limitation of our study, in that we used a single-night protocol, making it more difficult to discriminate poor sleep between groups, especially in adolescents (e.g., PSG SE was on average 90% in both groups). Moreover, the number of recruited adolescents with insomnia symptoms was too low for testing differences between subject groups with acceptable statistical power. Future studies with larger and balanced sample sizes and using multiple-night protocols are warranted for better characterization of sleep patterns in adolescents with insomnia, as well as for better understanding of CST performance in sub-populations.

Finally, we explored the FC3 features we believed to be more relevant in its performance. First, we tested the role of firmware version and time of data collection since proprietary algorithms and their continuous update have been identified as critical factors to be considered when using CST (de Zambotti et al. 2019a; Depner et al. 2019). Indeed, since the algorithms are undisclosed, the scientific community is not aware, for instance, of potential updates affecting device performance during an ongoing study. While in the current study we found no significant relationships with any of the considered performance outcomes, these factors may become relevant in longitudinal assessments lasting a year or more, and need further consideration. Second, we provided a first attempt to explore the rationale of CST (and specifically FC3) in using HR data to classify sleep/wake patterns and sleep stages (see de Zambotti et al. 2020). Previous studies reported satisfactory accuracy of Fitbit devices in measuring HR during sleep in both healthy adolescents (de Zambotti et al. 2016) and healthy adults (Haghayegh et al. 2019a), a necessary condition for using HR data in sleep staging. A further condition (only partially met in the present study) is the FC3 HR sensitivity to sleep stages. At the individual level, we did not find a substantial relationship between EBE agreement and mean HR measured by the FC3 between lights-off and lights-on. At the EBE level, we explored the relationship between the FC3 accuracy in classifying EBE transition and the corresponding changes in FC3 HR.

Our results showed that the FC3 accuracy in classifying EBE transitions was positively associated with the magnitude of the corresponding HR changes, but only for transitions between different stages (e.g., light → deep). In contrast, transitions between epochs classified with the same stage by the PSG (e.g., light → light) were less accurately classified when associated with higher compared to lower changes in FC3 HR. This finding is possibly due to the algorithm "expectancy" of high HR stability between equally scored epochs, and when this is not the case (e.g., due to short awakenings, arousals, or changes in breathing patterns) the classification system fails. However, FC3's algorithm is proprietary, and the type of features (e.g., motion, tonic HR, short-term HRV), as well as their timing and integration, used to classify sleep/wake and sleep stages is unknown. In particular, HRV changes between epochs might be more informative of shifts between different sleep stages, compared to HR changes between epochs (for a recent discussion, see Radha et al. 2019). While motion is reasonably indicative of wake (Sadeh 2011), understanding the interplay of PPG-derived and accelerometry-based features used in sleep staging is more challenging (for promising attempts, see Beattie et al. 2017; Matar et al. 2018). Future studies should better characterize the CST's sleep-staging rationale, accounting for variables at both the EBE and individual level (see de Zambotti et al. 2018c).

Conclusion

In conclusion, while our findings support a positive trend for CST performance, they should raise awareness for researchers using CST to track sleep in adolescents, highlighting the potential sources of bias (BMI for WASO estimations, the range of measurement for "light" and "deep" sleep duration), and providing information (bias and LOAs) that can be used to calibrate the measurements collected in future studies using the FC3 device. CSTs may ultimately help advance understanding of adolescents' sleep patterns, including integration and evaluation of the presence of objective sleep deficits, quantification of circadian phase shifting across age, and night-to-night and week-weekend variability in sleep.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

All authors contributed to developing the study concept. MdZ, DY, AG and FCB contributed to the study design. Testing and data collection were performed by DY, in addition to research assistants Quimby Lee, Rena Wang, Teji Dulai, Nicole Arra, and Laila Volpe. Data analysis was performed by LM. All authors interpreted the data, drafted the manuscript, provided critical revisions, and approved the final version for submission.

Funding details.

This work was supported by the National Heart, Lung and Blood Institute (NHLBI) under Grant R01 HL139652 (MdZ). The content is solely the responsibility of the authors and does not necessarily represent the official views the National Institutes of Health.

References

American Psychiatric Association. 2013. Diagnostic and statistical manual of mental disorders: DSM 5. 5th. Arlington (VA): American Psychiatric Publishing.

Beattie Z, Oyang Y, Statan A, Ghoreyshi A, Pantelopoulos A, Russell A, Heneghan C. 2017. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. Physiol Meas. 38:1968–1979. doi:10.1088/1361-6579/aa9047. [PubMed: 29087960]

Berry RB, Brooks R, Gamaldo CE, Harding SM, Lloyd RM, Marcus CL, Vaughn B V. 2020. The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications version 2.6. Darien (IL): American Academy of Sleep Medicine.

Bland JM, Altman DG. 1999. Measuring agreement in method comparison studies. Stat Methods Med Res. 8:135–160. doi:10.1177/096228029900800204. [PubMed: 10501650]

Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. 1989. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. Psychiatry Res. 28:193–213. doi:10.1016/0165-1781(89)90047-4. [PubMed: 2748771]

Byrt T, Bishop J, Carlin JB. 1993. Bias, prevalence and kappa. J Clin Epidemiol. 46:423–429. doi:10.1016/0895-4356(93)90018-V. [PubMed: 8501467]

Carskadon MA. 2011. Sleep in adolescents: The perfect storm. Pediatr Clin North Am. 58:637–647. doi:10.1016/j.pcl.2011.03.003. [PubMed: 21600346]

Chinoy ED, Cuellar JA, Huwa KE, Jameson JT, Watson CH, Bessman SC, Hirsch DA, Cooper AD, Drummond SPA, Markwald RR. 2020. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep. doi:10.1093/sleep/zsaa291.

Cohen J 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas. 20:37–46. doi:10.1177/001316446002000104.

Colrain IM, Baker FC. 2011. Changes in sleep as a function of adolescent development. Neuropsychol Rev. 21:5–21. doi:10.1007/s11065-010-9155-5. [PubMed: 21225346]

Cook JD, Eftekari SC, Dallmann E, Sippy M, Plante DT. 2019. Ability of the Fitbit Alta HR to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: A comparison against polysomnography. J Sleep Res. 28:e12789. doi:10.1111/jsr.12789. [PubMed: 30407680]

de Zambotti M, Baker FC, Colrain IM. 2015. Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep. 38:1461–1468. doi:10.5665/sleep.4990. [PubMed: 26158896]

de Zambotti M, Baker FC, Willoughby AR, Godino JG, Wing D, Patrick K, Colrain IM. 2016. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. Physiol Behav. 158:143–149. doi:10.1016/j.physbeh.2016.03.006. [PubMed: 26969518]

de Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. 2019a. Wearable sleep technology in clinical and research settings. Med Sci Sport Exerc. 51:1538–1557. doi:10.1249/MSS.0000000000001947.

de Zambotti M, Cellini N, Menghini L, Sarlo M, Baker FC. 2020. Sensors capabilities, performance, and use of consumer sleep technology. Sleep Med Clin. 15. doi:10.1016/j.jsmc.2019.11.003.

de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. 2018a. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. Chronobiol Int. 35:465–476. doi:10.1080/07420528.2017.1413578. [PubMed: 29235907]

de Zambotti M, Goldstone A, Colrain IM, Baker FC. 2018b. Insomnia disorder in adolescence: Diagnosis, impact, and treatment. Sleep Med Rev. 39:12–24. doi:10.1016/j.smrv.2017.06.009. [PubMed: 28974427]

de Zambotti M, Javitz H, Franzen PL, Brumback T, Clark DB, Colrain IM, Baker FC. 2018c. Sex- and age-dependent differences in autonomic nervous system functioning in adolescents. J Adolesc Heal. 62:184–190. doi:10.1016/j.jadohealth.2017.09.010.

de Zambotti M, Rosas L, Colrain IM, Baker FC. 2019b. The sleep of the ring: Comparison of the ŌURA sleep tracker sgainst polysomnography. Behav Sleep Med. 17:124–136. doi:10.1080/15402002.2017.1300587. [PubMed: 28323455]

de Zambotti M, Trinder J, Silvani A, Colrain IM, Baker FC. 2018d. Dynamic coupling between the central and autonomic nervous systems during sleep: A review. Neurosci Biobehav Rev. 90:84–103. doi:10.1016/j.neubiorev.2018.03.027. [PubMed: 29608990]

Depner CM, Cheng PC, Devine JK, Khosla S, de Zambotti M, Robillard R, Vakulin A, Drummond SPA. 2019. Wearable technologies for developing sleep and circadian biomarkers: A summary of workshop discussions. Sleep. doi:10.1093/sleep/zsz254.

Driller MW, O'Donnell S, Tavares F. 2017. What wrist should you wear your actigraphy device on? Analysis of dominant vs. non-dominant wrist actigraphy for measuring sleep in healthy adults. Sleep Sci. 10:132–135. doi:10.5935/1984-0063.20170023. [PubMed: 29410743]

Haghayegh S, Kang H-A, Khoshnevis S, Smolensky MH, Diller KR. 2020a. A comprehensive guideline for bland-altman and intra class correlation calculations to properly compare two methods of measurement and interpret findings. Physiol Meas. doi:10.1088/1361-6579/ab86d6.

Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR. 2019a. Accuracy of PurePulse photoplethysmography technology of Fitbit Charge 2 for assessment of heart rate during sleep. Chronobiol Int. 36:927–933. doi:10.1080/07420528.2019.1596947. [PubMed: 30990098]

Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. 2019b. Accuracy of wristband Fitbit models in assessing sleep: Systematic review and meta-analysis. J Med Internet Res. 21:e16273. doi:10.2196/16273. [PubMed: 31778122]

Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. 2020b. Performance assessment of new-generation Fitbit technology in deriving sleep parameters and stages. Chronobiol Int. 37:47–59. doi:10.1080/07420528.2019.1682006. [PubMed: 31718308]

Kang S-G, Kang JM, Ko K-P, Park S-C, Mariani S, Weng J. 2017. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. J Psychosom Res. 97:38–44. doi:10.1016/j.jpsychores.2017.03.009. [PubMed: 28606497]

Khosla S, Deak MC, Gault D, Goldstein CA, Hwang D, Kwon Y, O'Hearn D, Schutte-Rodin S, Yurcheshen M, Rosen IM, et al. 2018. Consumer sleep technology: An American Academy of Sleep Medicine position statement. J Clin Sleep Med. 14:877–880. doi:10.5664/jcsm.7128. [PubMed: 29734997]

Kuula L, Gradisar M, Martinmäki K, Richardson C, Bonnar D, Bartel K, Lang C, Leinonen L, Pesonen AK. 2019. Using big data to explore worldwide trends in objective sleep in the transition to adulthood. Sleep Med. 62:69–76. doi:10.1016/j.sleep.2019.07.024. [PubMed: 31563008]

Lee XK, Chee NIYN, Ong JL, Teo TB, van Rijn E, Lo JC, Chee MWL. 2019. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. J Clin Sleep Med. 15:1337–1346. doi:10.5664/jcsm.7932. [PubMed: 31538605]

Matar G, Lina JM, Carrier J, Kaddoum G. 2018. Unobtrusive sleep monitoring using cardiac, breathing and movements activities: An exhaustive review. IEEE Access. 6:45129–45152. doi:10.1109/ACCESS.2018.2865487.

Meltzer LJ, Walsh CM, Peightal AA. 2015. Comparison of actigraphy immobility rules with polysomnographic sleep onset latency in children and adolescents. Sleep Breath. 19:1415–1423. doi:10.1007/s11325-015-1138-6. [PubMed: 25687438]

Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. 2020. A standardized framework for testing the performance of sleep-tracking technology: Step-by-step guidelines and open-source code. Sleep. Accepted M. doi:10.1093/sleep/zsaa170.

Menghini L, Gianfranchi E, Cellini N, Patron E, Tagliabue M, Sarlo M. 2019. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. Psychophysiology. 56. doi:10.1111/psyp.13441.

Nieuwenhuis R, Grotenhuis Manfred T, Pelzer B. 2012. influence.ME: Tools for detecting influential data in mixed effects models. R J. 4:38–47. doi:10.32614/RJ-2012-011.

Pesonen A-K, Kuula L. 2018. The Validity of a new consumer-targeted wrist device in sleep measurement: An overnight comparison against polysomnography in children and adolescents. J Clin Sleep Med. 14:585–591. doi:10.5664/jcsm.7050. [PubMed: 29609722]

Portaluppi F, Smolensky MH, Touitou Y. 2010. Ethics and methods for biological rhythm research on animals and human beings. Chronobiol Int. 27:1911–1929. doi:10.3109/07420528.2010.516381. [PubMed: 20969531]

R Development Core Team. 2018. R: A language and environment for statistical computing. Wien, Austria: R Foundation for Statistical Computing, http://www.r-project.org/.

Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, Long X, Aarts RM. 2019. Sleep stage classification from heart-rate variability using long short-term memory neural networks. Sci Rep. 9:14149. doi:10.1038/s41598-019-49703-y. [PubMed: 31578345]

Sadeh A 2011. The role and validity of actigraphy in sleep medicine: An update. Sleep Med Rev. 15:259–267. doi:10.1016/j.smrv.2010.10.001. [PubMed: 21237680]

Shcherbina A, Mattsson C, Waggott D, Salisbury H, Christle J, Hastie T, Wheeler M, Ashley E. 2017. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med. 7:3. doi:10.3390/jpm7020003.

Toon E, Davey MJ, Hollis SL, Nixon GM, Horne RSC, Biggs SN. 2016. Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. J Clin Sleep Med. 12:343–350. doi:10.5664/jcsm.5580. [PubMed: 26446248]

Wagenmakers E-J, Farrell S. 2004. AIC model selection using Akaike weights. Psychon Bull Rev. 11:192–196. doi:10.3758/BF03206482. [PubMed: 15117008]

Werner H, Molinari L, Guyer C, Jenni OG. 2008. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. Arch Pediatr Adolesc Med. 162:350–358. doi:10.1001/archpedi.162.4.350. [PubMed: 18391144]

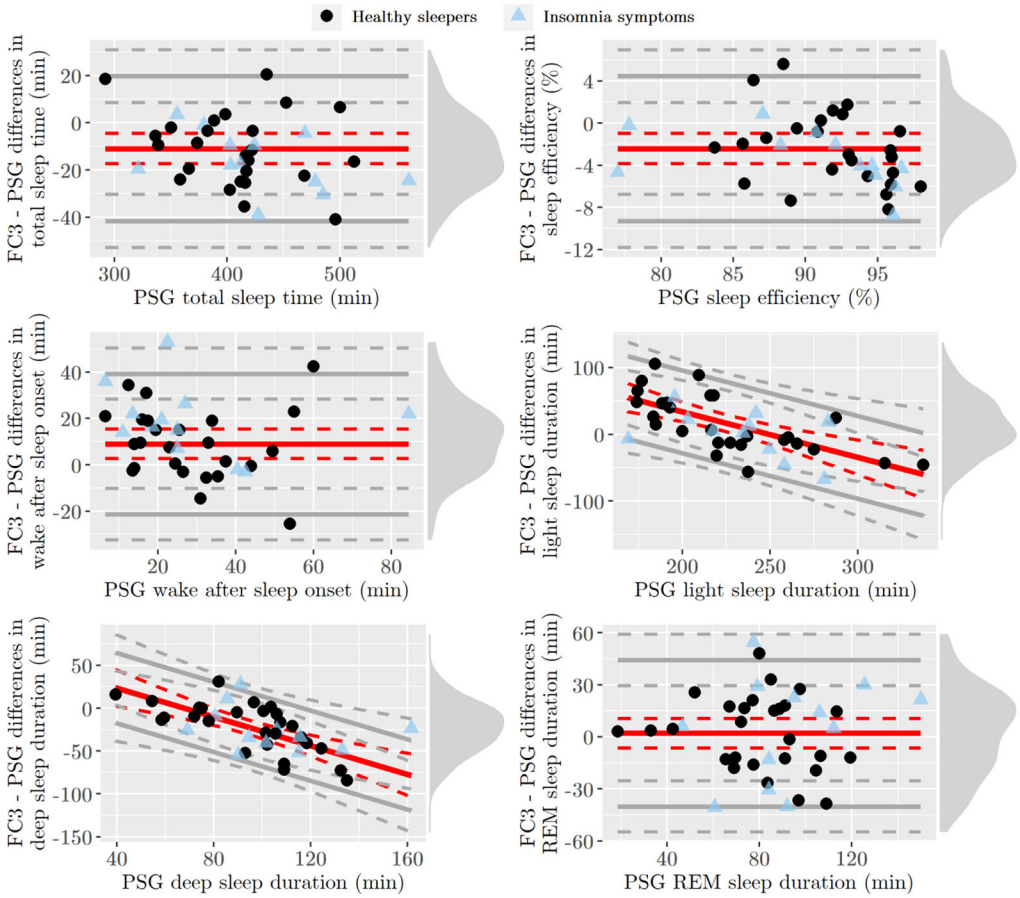

Figure 1.

Bland-Altman plots of sleep measures in the sample of healthy sleepers (black dots) and adolescents with insomnia symptoms (blue triangles). **PSG**, polysomnography; **REM**, Rapid-Eye-Movement. Red solid lines indicate bias, whereas gray solid lines indicate the 95% limits of agreement (LOAs), both with their 95% confidence intervals (dotted lines), computed from the group of healthy sleepers. Density diagram on the right side of each plot represents the distribution of FC3-PSG differences among healthy sleepers.

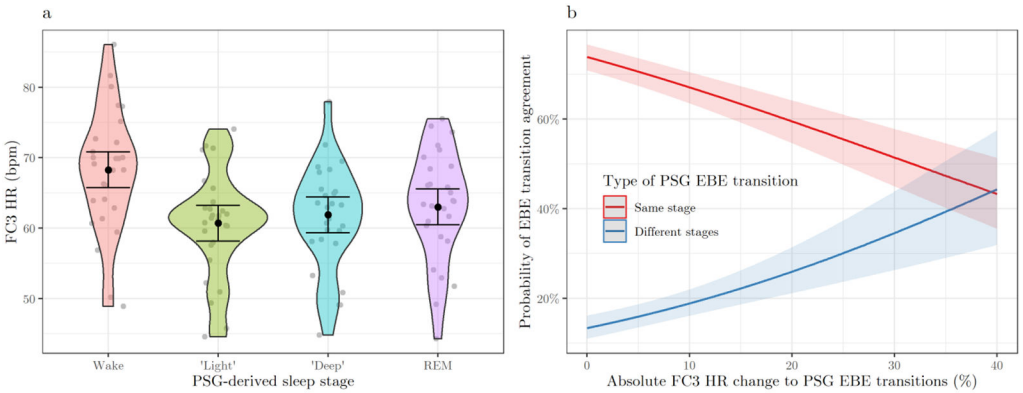

Figure 2.

Fitbit Charge 3™ heart rate by polysomnographic sleep stage (a) and predicted probability of epoch-by-epoch (EBE) transition agreement by absolute Fitbit Charge 3™ heart rate changes to polysomnographic EBE transitions (b), in the sample of healthy sleepers. **FC3**, Fitbit Charge 3™, **PSG**, polysomnography, **HR**, heart rate; **bpm**, beats per minute; **"light",** PSG-based N1 + N2 sleep; **"deep"**, PSG-based N3 sleep; **REM**, Rapid-Eye-Movement sleep. Figure (a) shows the distribution of FC3 HR values averaged by participant (gray dots) for each PSG sleep stage (error bars indicate 95% confidence intervals). Figure (b) shows the predicted probabilities (with 95% confidence intervals) of EBE transition agreement depending on absolute percent FC3 HR change to transitions between consecutive epochs classified with the same stage (solid red line) or with different stages (solid blue line) by the PSG.

Table 1.

Sample characteristics reported as mean (SD).

	Healthy sleepers	Insomnia group
Sample, No.	27	12
Sex, No. M/F	12/15	5/7
Age, y	17.7 (.6)	17.6 (.8)
Caucasian, No.	25	11
BMI, kg m^{-2}	22.5 (3.7)	23.3 (5.3)
PSQI, total score	3.4 (2.2)	6.5 (2.1)

BMI, Body Mass Index; **PSQI,** Pittsburg Sleep Quality Index. BMI was calculated based on in-lab measurements of height and weight. The PSQI(Buysse et al. 1989) is a 19-items questionnaire investigating habitual sleep over the past month, with higher scores reflecting poorer sleep (total score ranging from 0 to 21).

Table 2.

Agreement between polysomnographic and Fitbit Charge 3™ classification of epoch-by-epoch transitions by considering the corresponding cardiac reactivity (absolute percent change in heart rate).

	…	Epoch 234	Epoch 235	Epoch 236	Epoch 237	Epoch 238	…
PSG epoch classification	…	Light	Light	REM	REM	Wake	…
PSG EBE transition (current vs. previous epoch)	…	…	Light → Light	Light → REM	REM → REM	REM → Wake	
Type of PSG EBE transition	…	…	Same stage	Different stages	Same stage	Different stages	…
FC3 epoch classification	…	Light	Light	REM	Light	Wake	…
FC3 EBE transition (current vs. previous epoch)		…	Light → Light	Light → REM	REM → Light	Light → Wake	
EBE transition agreement	…	…	1	1	0	0	…
FC3 HR (bpm)	…	50	52	54	49	58	…
Absolute FC3 HR change (%) (current vs. previous epoch)	…	…	4.00	3.85	9.26	18.37	…

PSG, Polysomnography; **FC3**, Fitbit Charge 3™; **HR**, Heart rate; **REM**, Rapid-Eye-Movement. EBE transitions between consecutive epochs were used to evaluate the relationship between EBE transition accuracy and FC3 HR changes to PSG EBE transitions. EBE transition agreement was considered as accurate (1) when both the current and the previous epoch were equally classified by FC3 and PSG, inaccurate (0) otherwise. The type of PSG EBE transition (same stage vs. different stages) was used as a further predictor of EBE transition agreement.

Table 3.

Sleep measures and group-level discrepancies in the sample of healthy sleepers and adolescents with insomnia symptoms.

		Fitbit Charge 3™ Mean (SD)	PSG Mean (SD)	Bias Mean (SD) [95% CI]	Lower LOA [95% CI]	Upper LOA [95% CI]
TST (min)	Healthy Sleepers	396.76 (51.40)	407.72 (53.03)	−10.96 (15.65) [−17.42, −4.50] *	−41.63 [−52.82, −30.44]	19.71 [8.52, 30.90]
	Insomnia Symptoms	410.21 (60.17)	426.46 (64.46)	−16.25 (12.55) [−23.17, −9.50] *	−40.85 [−47.72, −34.05]	8.35 [1.68, 15.10]
SE (%)	Healthy Sleepers	89.46 (4.23)	91.89 (4.02)	−2.43 (3.51) [−3.87, −.98] *	−9.30 [−11.80, −6.79]	4.44 [1.94, 6.95]
	Insomnia Symptoms	86.98 (5.96)	90.42 (6.83)	14.32 - .20 × PSG b0 = [3.19, 70.71], b1 = [−.80, −.01] *	Bias - 4.58 [1.70, 5.71]	Bias + 4.58 [1.70, 5.71]
SOL (min)	Healthy Sleepers	9.36 (7.39)	7.40 (7.46)	1.96 (5.16) [−.17, 4.09]	Bias - 2.46(1.10 + .33 × PSG) c0 = [−.17, 2.37], c1 = [.20, .45] *	Bias + 2.46(1.10 + .33 × PSG) c0 = [−.17, 2.37], c1 = [.20, .45] *
	Insomnia Symptoms	13.62 (8.99)	16.25 (22.79)	−2.62 (17.64) [−13.50, 5.54]	−37.21 [−47.66, −29.33]	31.96 [21.12, 39.83]
WASO (min)	Healthy sleepers	37.46 (19.99)	28.46 (14.91)	9.00 (15.47) [2.61, 15.39] *	−21.32 [−32.38, −10.26]	39.32 [28.26, 50.38]
	Insomnia Symptoms	47.00 (22.60)	28.12 (20.69)	18.88 (15.48) [10.75, 27.62] *	−11.46 [−19.59, −2.76]	49.21 [41.09, 57.76]
Light (min)	Healthy Sleepers	241.89 (34.49)	225.41 (43.26)	170.26 - .68 × PSG b0 = [101.18, 239.35], b1 = [−.98, −.38] *	Bias - 62.00 [49.74, 80.23]	Bias + 62.00 [49.74, 80.23]
	Insomnia Symptoms	234.62 (34.51)	230.21 (35.78)	4.42 (35.93) [−15.46, 23.83]	−66.00 [−86.79, −47.59]	74.84 [54.88, 93.63]
Deep (min)	Healthy Sleepers	72.2 (21.26)	94.65 (24.28)	56.93 - .84 × PSG b0 = [22.32, 91.54], b1 = [−1.19, −.48] *	Bias - 40.96 [34.16, 52.17]	Bias + 40.96 [34.16, 52.17]
	Insomnia Symptoms	78.04 (29.44)	103.58 (25.20)	−25.54 (25.01) [−38.29, −11.04] *	−74.55 [−87.05, −60.22]	23.47 [10.93, 37.68]
REM (min)	Healthy Sleepers	82.15 (28.63)	80.11 (23.94)	2.04 (21.53) [−6.48, 10.55]	−40.16 [−54.91, −25.41]	44.23 [29.48, 58.98]
	Insomnia Symptoms	97.54 (46.87)	92.67 (28.05)	4.88 (30.14) [−11.58, 20.79]	−54.19 [−71.03, −38.28]	63.94 [47.32, 79.82]

TST, Total sleep time; **SE**, Sleep efficiency; **SOL**, Sleep onset latency; **WASO**, Wake after sleep onset; **Light**, "light" sleep duration (i.e., PSG-derived N1 + N2); **Deep**, "deep" sleep duration (i.e., PSG-derived N3); **REM**, Rapid-Eye-Movement sleep; **SD**, standard deviation; **PSG**, polysomnography; **LOA**, limit of agreement; **CI**, confidence intervals (computed using percentile bootstrap for the group with insomnia symptoms due to skewed distributions and small sample size)

*, cases showing a significant bias, proportional bias or heteroscedasticity. When a proportional bias was detected, a linear model predicting the discrepancies by the corresponding PSG measures was specified, and 95% CI were reported for the model's intercept (b0) and slope (b1). When heteroscedasticity was detected, a linear model predicting the absolute residuals of the previous model by PSG-derived measures was specified, and 95% CI were reported for the model's intercept (c0) and slope (c1).

Table 4.

Group-level proportional error matrix in the sample of healthy sleepers and adolescents with insomnia symptoms.

			Fitbit Charge 3™					
			Wake	**“light”**	**“deep”**	**REM**	**kappa**	**PABAK**
PSG	Wake	Healthy sleepers	**.68 (.18) [.60, .75]**	.21 (.15) [.16, .27]	.01 (.02) [.00, .02]	.10 (.11) [.06, .14]	.52 (.14) [.47, .58]	.83 (.10) [.79, .87]
		Insomnia symptoms	**.81 (.12) [.74, .87]**	.12 (.08) [.08, .17]	.00 (.01) [.00, .01]	.07 (.07) [.03, .10]	.59 (.14) [.52, .66]	.86 (.03) [.84, .87]
	“light”	Healthy sleepers	.08 (.05) [.06, .09]	**.78 (.07) [.75, .81]**	.07 (.04) [.05, .09]	.07 (.05) [.05, .09]	.48 (.14) [.43, .54]	.48 (.14) [.42, .54]
		Insomnia symptoms	.08 (.02) [.06, .09]	**.71 (.13) [.64, .78]**	.08 (.08) [.04, .12]	.13 (.10) [.08, .19]	.42 (.18) [.32, .52]	.43 (.18) [.33, .52]
	“deep”	Healthy sleepers	.02 (.02) [.01, .02]	.39 (.24) [.29, .48]	**.59 (.25) [.50, .69]**	.00 (.01) [.00, .01]	.48 (.14) [.43, .54]	.48 (.14) [.42, .54]
		Insomnia symptoms	.02 (.02) [.01, .03]	.40 (.17) [.31, .49]	**.57 (.17) [.48, .66]**	.01 (.03) [.00, .03]	.42 (.18) [.32, .51]	.43 (.18) [.33, .52]
	REM	Healthy sleepers	.03 (.03) [.02, .04]	.20 (.17) [.14, .27]	.01 (.03) [.00, .02]	**.76 (.18) [.68, .83]**	.56 (.25) [.47, .66]	.74 (.14) [.68, .79]
		Insomnia symptoms	.05 (.06) [.02, .09]	.26 (.17) [.17, .36]	.00 (.01) [.00, .01]	**.69 (.19) [.58, .79]**	.56 (.18) [.47, .66]	.73 (.12) [.68, .80]

“light”, PSG-based N1 + N2; **“deep”**, PSG-based N3; **REM**, Rapid-Eye-Movement sleep; **PSG**, polysomnography; **PABAK**, prevalence-adjusted bias-adjusted kappa. Results are reported as mean (standard deviation) [95% confidence intervals]. Bold type indicates sleep-stage sensitivity (device’s ability to correctly detect a given stage), whereas the remaining cells indicate percentages of misclassifications. Confidence intervals were computed using percentile bootstrap for the group with insomnia symptoms due to skewed distributions and small sample size.