JCSM
Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Validation of a Consumer Sleep Wearable Device With Actigraphy and Polysomnography in Adolescents Across Sleep Opportunity Manipulations

Xuan Kai Lee, BSc*; Nicholas I.Y.N. Chee, BSc*; Ju Lynn Ong, PhD; Teck Boon Teo, BSc; Elaine van Rijn, PhD; June C. Lo, PhD; Michael W.L. Chee, MBBS

*Centre for Cognitive Neuroscience, Neuroscience and Behavioral Disorders Program, Duke-NUS Medical School, Singapore; *Co-first authors and contributed equally*

Study Objectives: To compare the quality and consistency in sleep measurement of a consumer wearable device and a research-grade actigraph with polysomnography (PSG) in adolescents.

Methods: Fifty-eight healthy adolescents (aged 15–19 years; 30 males) underwent overnight PSG while wearing both a Fitbit Alta HR and a Philips Respironics Actiwatch 2 (AW2) for 5 nights, with either 5 hours or 6.5 hours time in bed (TIB) and for 4 nights with 9 hours TIB. AW2 data were evaluated using two different wake and immobility thresholds. Discrepancies in estimated total sleep time (TST) and wake after sleep onset (WASO) between devices and PSG, as well as epoch-by-epoch agreements in sleep/wake classification, were assessed. Fitbit-generated sleep staging was compared to PSG.

Results: Fitbit and AW2 under default settings similarly underestimated TST and overestimated WASO (TST: medium setting (M10) ≤ 38 minutes, Fitbit ≤ 47 minutes; WASO: M10 ≤ 38 minutes; Fitbit ≤ 42 minutes). AW2 at the high motion threshold setting provided readings closest to PSG (TST: ≤ 12 minutes; WASO: ≤ 18 minutes). Sensitivity for detecting sleep was ≥ 90% for both wearable devices and further improved to 95% by using the high threshold (H5) setting for the AW2 (0.95). Wake detection specificity was highest in Fitbit (≥ 0.88), followed by the AW2 at M10 (≥ 0.80) and H5 thresholds (≤ 0.73). In addition, Fitbit inconsistently estimated stage N1 + N2 sleep depending on TIB, underestimated stage N3 sleep (21–46 min), but was comparable to PSG for rapid eye movement sleep. Fitbit sensitivity values for the detection of N1 + N2, N3 and rapid eye movement sleep were ≥ 0.68, ≥ 0.50, and ≥ 0.72, respectively.

Conclusions: A consumer-grade wearable device can measure sleep duration as well as a research actigraph. However, sleep staging would benefit from further refinement before these methods can be reliably used for adolescents.

Clinical Trial Registration: Registry: ClinicalTrials.gov; Title: The Cognitive and Metabolic Effects of Sleep Restriction in Adolescents; Identifier: NCT03333512; URL: https://clinicaltrials.gov/ct2/show/NCT03333512

Keywords: actigraphy, adolescent sleep, Fitbit, polysomnography

Citation: Lee XK, Chee NIYN, Ong JL, Teo TB, van Rijn E, Lo JC, Chee MWL. Validation of a consumer sleep wearable device with actigraphy and polysomnography in adolescents across sleep opportunity manipulations. *J Clin Sleep Med*. 2019;15(9):1337–1346.

BRIEF SUMMARY

Current Knowledge/Study Rationale: Consumer sleep trackers are an attractive alternative to expensive research actigraphs for measuring sleep. However, validation studies in adolescent populations are limited and typically conducted on only 1 night of sleep. We compared a consumer sleep/activity tracker and a research-grade actigraph with polysomnography (PSG) over different sleep opportunities and across multiple nights.

Study Impact: Sleep estimation was comparable between the consumer wearable device and research-grade actigraphy on default settings. Both underestimated sleep duration compared to PSG. Sleep estimation improved in the research actigraph by adjusting sensitivity to motion. With data-driven customization, consumer wearable devices could replace research actigraphs for large-scale total sleep measurement. Sleep staging still lags behind PSG and needs further work, particularly for assessment of stage N3 sleep.

INTRODUCTION

Sleep is increasingly recognized as important for health and well-being. In addition to this growing awareness, wearable devices that incorporate accelerometers have proliferated on a massive scale. Annual global sales estimated to be under 25 million in 2014 are expected to reach 125 million by the end of 2018.[1] Growth in sales of smartwatches in particular have been even more dramatic, leaping from 5 million to 80 million in the same time interval. Originally intended to track physical activity, many wearable devices now incorporate algorithms that can provide outputs on sleep.[2–4]

Although polysomnography (PSG) remains the gold standard for quantifying sleep, wrist actigraphy based on movement (where the absence of motion implies sleep) is inexpensive and a widely available proxy for estimating sleep in nonlaboratory settings. Actigraphy is well suited for large-scale longitudinal surveys of personal and/or community sleep habits, and how these influence health and well-being. It has been well validated against PSG in adults and is widely adopted in research and clinical settings.[5–7] To date, expensive "research-grade" actigraphs remain the mainstay in scientific studies, influenced by mixed reports about the reliability,[8–12] particularly the accuracy of sleep detection of earlier consumer devices. However, constant advancements including the use of heart rate variability

(HRV) measurement to estimate sleep stages[13,14] and recent reports of good agreement with research devices[15,16] motivate a detailed re-evaluation of consumer-grade devices.

The current report has several features that might serve to better inform about the feasibility of using consumer grade activity trackers to estimate sleep in research studies. First, we collected multinight sleep data per individual across three levels of sleep opportunity (5 hours, 6.5 hours, and 9 hours), concurrently comparing a relatively new consumer wearable device that incorporates heart rate measurements to augment sleep/wake classification (Fitbit Alta HR, Fitbit Inc., San Francisco, California) with a research actigraph (Actiwatch 2, Philips Respironics Inc., Pittsburgh, Pennsylvania). Both devices were referenced to PSG sleep measurement. Second, we focused on an adolescent sample. Although actigraphy tends to overestimate sleep in adults, some studies have found underestimation of sleep in adolescents.[17–19] To examine how sensitivity to motion might affect appropriate sleep detection, we used two different motion sensitivity settings to evaluate sleep. Finally, we assessed how well the consumer wearable device could stage adolescent sleep.

METHODS

Participants

Participants consisted of 58 adolescents aged 15 to 19 years (mean ± standard deviation [SD]: 16.6 ± 0.94 years; 30 males) who took part in a study examining the cognitive and metabolic effects of sleep restriction in adolescents. They were recruited through social media, online advertisements, talks, and word of mouth. Consent was obtained from both participants and their legal guardians. Participants had no known health conditions or sleep disorders, were not habitual short sleepers (self-reported total sleep time [TST] < 5 hours on weekdays concurrent with ≤ 1 hour of weekend sleep extension), and did not travel across more than two time zones in the month prior to the study.

Study Protocol

Participants underwent a 14-night evaluation (**Figure S1** in the supplemental material) in a boarding school under quasi-laboratory conditions. The study protocol was approved by the Institutional Review Board of the National University of Singapore and in accordance with the principles of the Declaration of Helsinki. The sleep schedule was designed to simulate one-and-a-half cycles of shortened sleep on weekdays and extended sleep on weekends in adolescents. Students who were randomized into the continuous (n = 29) and the split (n = 29) sleep groups did not significantly differ in demographic or habitual sleep characteristics (**Table S1** in the supplemental material). During the 2 baseline sleep and 4 recovery nights, all participants had a 9-hour (11:00 PM to 8:00 AM) sleep opportunity. On 8 manipulation nights, participants in the continuous sleep group had a 6.5-hour (12:15 AM to 6:45 AM) nocturnal sleep opportunity, whereas those in the split sleep group had a 5-hour (1:00 AM to 6:00 AM) nocturnal sleep opportunity plus a 1.5-hour (2:00 PM to 3:30 PM) afternoon nap following the night of restricted sleep. Actigraphy and Fitbit data were recorded throughout the protocol. PSG data were available for 9 nights: 2 baseline nights, 5 sleep restriction nights, and 2 recovery nights (**Figure S1**, asterisks). All devices were synchronized to a common Internet time server to ensure proper alignment of time-stamped data. Because the first baseline night was an adaptation night, data were not analyzed.

Polysomnography

Electroencephalography (EEG) was performed using a SOMNOtouch recorder (SOMNOmedics GmbH, Randersacker, Germany) on two channels (C3 and C4 in the international 10-20 system). Contralateral mastoids were used as references. Electrodes placed at Cz and Fpz were used as common reference and ground electrodes respectively. Electrooculography (EOG) and submental electromyography (EMG) were also used. Impedance was kept below 5 kΩ for EEG and 10 kΩ for EOG and EMG electrodes. Signals were sampled at 256 Hz and filtered between 0.2 and 35 Hz for EEG, and between 0.2 and 10 Hz for EOG.

Sleep periods were set according to the times of lights on and off. Sleep stages were automatically scored in 30-second epochs using an in-house algorithm[20] (https://z3score.com) in conjunction with the FASST toolbox (http://www.montefiore.ulg.ac.be/~phillips/FASST.html), and visually reviewed by trained technicians following criteria set by The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications to ensure accuracy of the staging.[21]

TST was computed by totaling epochs of stage N1, N2, N3, and rapid eye movement (REM) sleep, whereas wake after sleep onset (WASO) was defined by the sum of epochs scored as wake after the first stage N1 or N2 sleep epoch (because Fitbit does not distinguish between these two sleep stages). For sleep staging comparisons with Fitbit, PSG epochs classified as stage N1 and N2 sleep were categorized as "light sleep" and stage N3 sleep PSG epochs as "deep sleep."

Actigraphy

Participants wore an Actiwatch 2 (AW2) on their nondominant hand. Data were collected in 30-second epochs, with sleep periods manually defined by lights on and off times, matching those of PSG. Data were processed using Actiware (version 6.0.7, Philips Respironics Inc., Pittsburgh, Pennsylvania), using two wake threshold and immobility settings. The default setting utilizes a medium wake threshold (40 counts per epoch) with 10 immobile minutes (M10) for sleep onset and end. In addition, an optimized setting employing a high wake threshold (80 counts per epoch) and 5 immobile minutes (H5) was also included as a comparison based on prior findings suggesting increased movement during sleep in adolescents.[18,22] TST was computed by summing all sleep epochs within the sleep periods, whereas WASO was defined by summing all wake epochs after the first sleep epoch.

Consumer Wearable Device

During the protocol, each participant wore a Fitbit Alta HR, hereafter simply referred to as Fitbit, on their nondominant hand. This device tracks motion and HRV via accelerometers and optical plethysmography respectively in 30-second epochs.

A proprietary classification algorithm utilizing accelerometry and HRV signals[23] classifies epochs into wake, or one of three sleep stages: light, deep, or REM sleep. Data were then wirelessly transferred to a smartphone application, and batch extracted using a third-party data management platform (Fitabase, San Diego, California).

Given that Fitbit sleep periods are automatically defined, measures were taken to ensure that time in bed (TIB) was identical across all the three instruments. Specifically, if Fitbit sleep onset or offset timings occurred between scheduled lights off and lights on timings, wake epochs were inserted to the beginning and/or the end of the record. Conversely, if sleep onset or offset timings occurred outside of scheduled timings, wake periods at the beginning and/or the end of the record were truncated. TST was computed by summing the duration of all Light, Deep and REM epochs within each sleep period, whereas WASO was defined by summing all wake epochs after the first epoch of sleep.

Data Analysis

SPSS 24.0 (IBM Corp., Armonk, New York) was used for all statistical analyses. First, to investigate whether the discrepancy from PSG in TST and WASO estimates differed across nights within each TIB condition (5 hours, 6.5 hours, and 9 hours), a general linear mixed model was performed with device setting (M10, H5, and Fitbit) and measurement night as factors. No significant interaction effects between device setting and night were found ($P \geq .10$), indicating that biases were similar in magnitude across all measurement nights. Subsequent analyses were thus performed on intrasubject averaged TST and WASO.

Next, one-sample *t* tests against zero were used to determine if estimations of TST and WASO by devices were significantly biased, that is, overestimated or underestimated, from PSG. In addition, for each TIB condition, a repeated-measures analysis of variance (ANOVA) was conducted for TST and WASO separately to compare differences in biases between different device settings and were followed by *post hoc t* tests to discern significant pairwise differences. Furthermore, for each TIB condition, Bland-Altman[24] plots for TST were generated by plotting the bias of each device setting from PSG against the TST averaged across the device setting and PSG. Similar Bland-Altman plots were created for WASO per TIB condition. To determine if bias magnitudes were proportional to the TST or WASO measure averaged across the device setting and PSG, simple linear regression was performed. In a secondary analysis, we also investigated the effects of sex on TST estimates for each TIB condition (supplemental material). As differential effects only occurred in the 5-hour TIB condition and in actigraphy, we did not perform further sex-related analyses.

Finally, to quantify the agreement in sleep-wake categorization between actigraphy/Fitbit and PSG, epoch-by-epoch (EBE) analyses were conducted for deriving three agreement measures: accuracy (ability to correctly classify epochs), sensitivity (ability to detect sleep), and specificity (ability to detect wake) for each device setting. Repeated-measures ANOVAs for each EBE agreement measure were similarly conducted within each TIB condition to compare differences in agreement performance between different device settings and were also followed up with *post hoc t* tests. Also, to quantify the agreement in sleep staging between Fitbit and PSG, for each sleep stage (light sleep, deep sleep, and REM), an EBE analysis was conducted, and a Bland-Altman plot was generated to show the duration discrepancies between instruments against the average duration assessed with Fitbit and PSG.

RESULTS

Fifty-seven patients contributed to the final sample, as one participant in the continuous sleep group dropped out. In addition, data loss from technical issues from Fitbit (58 records), Actiwatch 2 (2 records,) and PSG (12 records), and the exclusion of 11 outlier recordings (> 3 SDs) resulted in the final sample consisting 386 nights of data common to all devices, with each participant contributing between 3 to 7 nights of data.

PSG determined sleep architecture for the final sample is provided in **Table 1**. Results of one-sample *t* tests used to determine the significance of device setting-PSG biases are summarized in **Table 2**. Bland-Altman plots representing device setting-PSG biases for sleep-wake and sleep-stage analyses are presented in **Figure 1** and **Figure 2**, respectively. Results of simple linear regression used to investigate proportional biases are summarized in **Table 3**. Accuracy, sensitivity and specificity values for sleep-wake and sleep-stage classification are presented in **Table 4** and **Table 5** whereas EBE classification metrics are provided in **Table 6**.

Actiwatch 2 M10 Versus PSG

M10 significantly underestimated TST and overestimated WASO in all TIB conditions. M10 underestimated TST by an average of 24 to 38 minutes ($t \geq 8.19$, $P < .001$; **Table 2**). WASO duration was overestimated by an average of 22 to 38 minutes ($t \geq 10.14$, $P < .001$). EBE comparisons indicated comparable agreement across all TIBs. Sleep-wake discrimination accuracy was excellent (0.89 to 0.90; Table 6), with sensitivities ranging from 0.90 to 0.91, and good specificities from 0.80 to 0.86.

Actiwatch 2 H5 Versus PSG

H5 showed better agreement with PSG but still underestimated TST and overestimated WASO in all TIB conditions. TST was underestimated by an average of 7 to 12 minutes ($t \geq 3.49$, $P \leq .002$; **Table 2**) while WASO was overestimated by an average of 11 to 18 minutes ($t \geq 8.19$, $P < .001$). Sleep-wake accuracies ranged from 0.93 to 0.94 (Table 6). Sensitivity was 0.95 across all TIBs, whereas specificities were acceptable, ranging from 0.64 to 0.73.

Fitbit Versus PSG

Fitbit significantly underestimated TST and overestimated WASO across all TIB conditions. TST was underestimated by an average of 24 to 47 minutes ($t \geq 15.62$, $P < .001$: **Table 2**) whereas WASO was overestimated by an average of 21 to 41 minutes ($t \geq 15.14$, $P < .001$). EBE comparisons indicated excellent accuracy and sensitivity of around 0.90 across all TIB conditions (Table 6). Specificity was between 0.88 and 0.90.

Table 1—Polysomnography-determined sleep architecture.

	5-hour TIB (n = 29)	6.5-hour TIB (n = 28)	9-hour TIB (n = 57)
TIB	300.80 (0.52)	390.52 (0.08)	540.52 (0.13)
TST	276.09 (10.18)	368.61 (7.41)	497.77 (21.09)
Stage N1 sleep	6.57 (2.55)	7.02 (3.29)	11.44 (5.38)
Stage N2 sleep	128.88 (14.67)	173.52 (25.17)	260.93 (29.34)
Stage N1 + N2 sleep	135.45 (13.53)	180.54 (24.9)	272.37 (28.79)
Stage N3 sleep	89.18 (12.82)	114.52 (24.84)	109.89 (27.27)
REM sleep	51.47 (13.66)	73.56 (14.02)	115.52 (20.88)
WASO	5.90 (4.29)	7.07 (4.05)	16.23 (12.55)
Sleep efficiency (%)	91.79 (3.44)	94.39 (1.89)	92.09 (3.90)

Data presented as mean (standard deviation) in minutes unless otherwise indicated. Participants in both 5-hour or 6.5-hour TIB groups all had 9-hour TIB nocturnal sleep opportunities on some days of the protocol. REM = rapid eye movement, TIB = time in bed, TST = total sleep time, WASO = wake after sleep onset.

Table 2—Biases of each device setting from polysomnography, grouped by TIB condition.

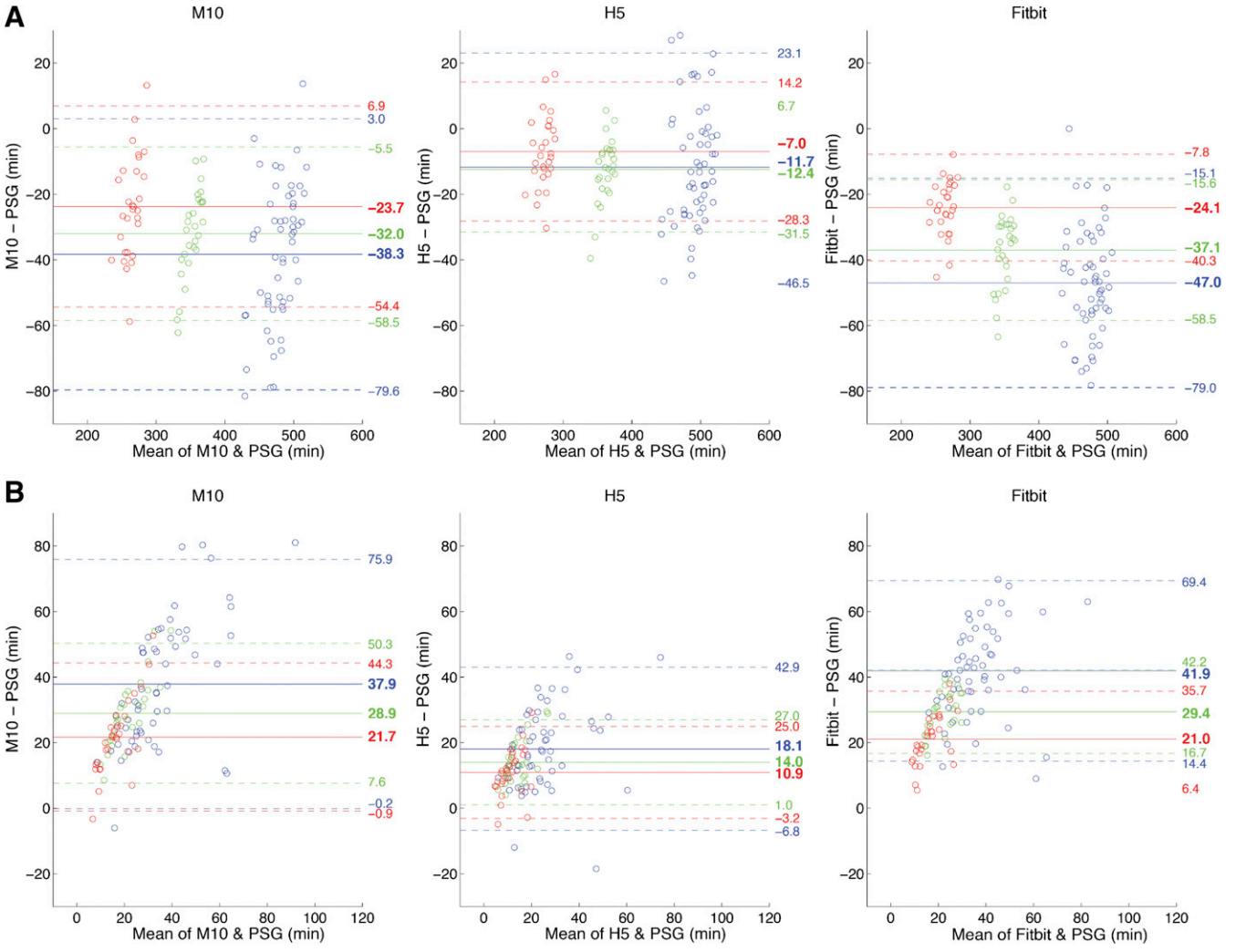
	M10	H5	Fitbit	*F*
5-hour TIB				
TST	**−23.74 (15.61)** [a]	**−7.03 (10.83)** [a,c]	**−24.05 (8.29)** [c]	86.11
Stage N1 + N2 sleep	–	–	**9.88 (19.74)**	–
Stage N3 sleep	–	–	**−37.77 (19.85)**	–
REM sleep	–	–	3.84 (16.80)	–
WASO	**21.69 (11.51)** [a]	**10.92 (7.18)** [a,c]	**21.03 (7.48)** [c]	54.61
6.5-hour TIB				
TST	**−32.02 (13.51)** [a,b]	**−12.43 (9.73)** [a,c]	**−37.05 (10.94)** [b,c]	88.98
Stage N1 + N2 sleep	–	–	3.30 (28.22)	–
Stage N3 sleep	–	–	**−46.38 (26.75)**	–
REM sleep	–	–	6.03 (20.69)	–
WASO	**28.93 (10.89)** [a]	**13.99 (6.62)** [a,c]	**29.44 (6.50)** [c]	73.24
9-hour TIB				
TST	**−38.30 (21.09)** [a,b]	**−11.73 (17.75)** [a,c]	**−47.04 (16.28)** [b,c]	153.87
Stage N1 + N2 sleep	–	–	**−20.65 (35.77)**	–
Stage N3 sleep	–	–	**−21.47 (34.03)**	–
REM sleep	–	–	−4.92 (26.42)	–
WASO	**37.87 (19.41)** [a]	**18.10 (12.67)** [a,c]	**41.87 (14.04)** [c]	109.08

Data presented as mean (standard deviation) in minutes. Significant biases are indicated in bold ($P < .05$). Analyses of variance of TST and WASO biases within each TIB were all significant ($P < .001$) even after corrections for sphericity violations. Negative values indicate underestimations. [a] M10 significantly different from H5 ($P < .05$) within each corresponding TIB. [b] M10 significantly different from Fitbit ($P < .05$) within each corresponding TIB. [c] H5 significantly different from Fitbit ($P < .05$) within each corresponding TIB. H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, REM = rapid eye movement, TIB = time in bed, TST = total sleep time, WASO = wake after sleep onset.

Concerning sleep-stage classification, biases were dependent on sleep stage and TIB condition examined. Fitbit overestimated stage N1 + N2 sleep (light sleep) by an average (SD) of 9.9 (19.7) minutes ($t = 2.70$, $P = .012$) in the 5-hour TIB condition; did not differ significantly from PSG in the 6.5-hour TIB condition, ($t = 0.62$, $P = .54$); and underestimated stage N1 + N2 sleep by an average (SD) of 20.7 (35.8) minutes ($t = 4.36$, $P < .001$) in the 9-hour TIB condition. The device consistently underestimated stage N3 sleep (deep sleep) duration in all TIB conditions, by an average of 21.5 to 46.4 minutes ($t \geq 8.19$, $P < .001$). No significant differences were observed for REM sleep ($t \leq 1.54$, $P \geq .13$) in all TIB conditions.

EBE comparisons of Fitbit's sleep staging algorithm indicated average accuracy of 0.68 to 0.71 for stage N1 + N2 sleep (light sleep), 0.50 to 0.64 for stage N3 sleep (deep sleep), and 0.72 to 0.74 for REM sleep. Confusion matrices (**Table 5**) indicate that, on average, misclassifications of PSG stage N1 + N2 sleep occurred mostly either as REM sleep (0.10 to 0.13) or

Figure 1—Bland-Altman plots for total sleep time and wake after sleep onset.



Bland-Altman plots, in minutes, of **(A)** total sleep time and **(B)** wake after sleep onset. Red, green, and blue points represent data collected from the 5-hour, 6.5-hour, and 9-hour time in bed conditions respectively. Solid lines and bolded numbers represent the mean biases of each recording, whereas dashed lines and regular numbers represent 1.96 standard deviation limits of agreement. H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, PSG = polysomnography.

wake (0.11 to 0.13); misclassifications of PSG stage N3 sleep occurred mostly as light sleep (0.31 to 0.43), and misclassifications of PSG REM sleep occurred mostly as light sleep (0.15 to 0.16).

Comparison Among Actiwatch 2 M10, Actiwatch 2 H5, and Fitbit

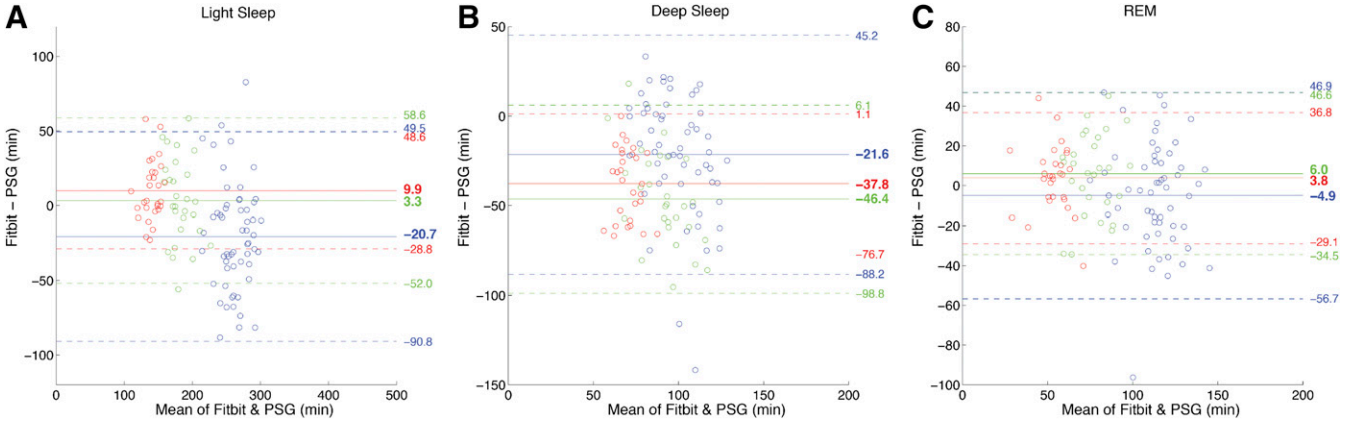
Magnitudes of device setting-PSG biases for TST and WASO (Table 2), and EBE agreement metrics (Table 6), between M10, H5, and Fitbit were compared. All ANOVAs were significant (TST: $F \geq 86.11$, $P < .001$; WASO: $F \geq 54.61$, $P < .001$; accuracy: $F \geq 33.03$, $P < .001$; sensitivity: $F \geq 79.84$, $P < .001$; specificity: $F \geq 33.80$, $P < .001$) for all TIB conditions examined.

Post hoc pairwise comparisons indicated that H5 had on average significantly less TST underestimation than M10 by 17 to 27 minutes ($t \geq 13.14$, $P < .001$), and Fitbit by 17 to 35 minutes ($t \geq 13.14$, $P < .001$) across all TIBs. Additionally, H5 had on average significantly less WASO overestimation than M10 by 11 to 20 minutes ($t \geq 9.59$, $P < .001$), and Fitbit by 10 to 24 minutes ($t \geq 10.07$, $P < .001$).

EBE analyses indicated that H5 had on average small but significantly higher sleep-wake classification accuracies than M10 by 0.03 ($t \geq 7.02$, $P < .001$), and Fitbits by 0.02 to 0.04 ($t \geq 6.35$, $P < .001$) across all TIBs. H5 also had on average small but significantly higher sensitivity values than M10 by 0.04 to 0.05 ($t \geq 10.71$, $P < .001$), and Fitbit by 0.04 to 0.05 ($t \geq 12.06$, $P < .001$) across all TIBs. However, this came at a cost of lower specificity values: H5 was lower than both M10 by 0.13 to 0.15 ($t \geq 7.30$, $P < .001$), and Fitbit by 0.16 to 0.24 ($t \geq 6.30$, $P < .001$).

M10 and Fitbit underestimated TST in the 5-hour TIB condition comparably ($t = 0.17$, $P = .87$). This underestimation of TST was larger in the 6.5-hour (mean [SD] = 5.0 [12.6] minutes, $t = 2.11$, $P = .044$) and 9-hour recordings (mean [SD] = 8.7 [19.4] minutes, $t = 3.40$, $P = .001$). However, M10 and Fitbit had similar overestimations of WASO across all TIB conditions ($t \leq 1.91$, $P \geq .06$).

Figure 2—Bland-Altman plots for sleep stages.



Bland-Altman plots, in minutes, of **(A)** stage N1 + N2 sleep (light sleep), **(B)** stage N3 sleep (deep sleep), and **(C)** REM sleep. Red, green, and blue points represent data collected from the 5-hour, 6.5-hour, and 9-hour time in bed conditions, respectively. Solid lines and bold numbers represent the mean biases of each recording, whereas dashed lines and regular numbers represent 1.96 standard deviation limits of agreement. H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, PSG = polysomnography, REM = rapid eye movement.

Table 3—Proportional biases associated with sleep duration observed in each device setting and grouped by TIB condition.

	5-hour TIB			6.5-hour TIB			9-hour TIB		
	M10	**H5**	**Fitbit**	**M10**	**H5**	**Fitbit**	**M10**	**H5**	**Fitbit**
TST	**0.70 (0.21)**	0.36 (0.18)	0.22 (0.15)	**0.94 (0.12)**	**0.65 (0.16)**	**0.69 (0.20)**	**0.40 (0.11)**	0.15 (0.11)	−0.02 (0.11)
Stage N1 + N2 sleep	–	–	0.37 (0.32)	–	–	−0.39 (0.31)	–	–	−0.08 (0.22)
Stage N3 sleep	–	–	−0.35 (0.54)	–	–	**−1.03 (0.29)**	–	–	**−0.84 (0.27)**
REM sleep	–	–	−0.26 (0.33)	–	–	0.28 (0.36)	–	–	−0.12 (0.23)
WASO	**1.31 (0.20)**	**0.74 (0.27)**	**0.86 (0.19)**	**1.23 (0.14)**	**0.80 (0.19)**	**0.78 (0.20)**	**0.75 (0.13)**	**0.32 (0.12)**	**0.31 (0.15)**

Data presented as *B* (standard error), in minutes. Bold fonts indicate significant ($P < .05$) associations. Biases were linearly regressed onto the mean of polysomnography and device setting duration. H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, REM = rapid eye movement, TIB = time in bed, TST = total sleep time, WASO = wake after sleep onset.

Sleep-wake accuracies did not significantly differ in the 6.5-hour and 9-hour TIBs ($t \leq 1.18$, $P \geq .25$) across all device settings. Fitbit was only slightly more accurate than M10 in the 5-hour TIB condition (mean [SD] = 0.01 [0.02]; $t = 2.45$, $P = .019$). Sensitivity values of M10 and Fitbit were comparable across all TIBs ($t \leq 1.73$, $P \geq .09$). Finally, although specificity was higher in Fitbit compared to M10 in both 5-hour, mean (SD) = 0.08 (0.17); $t = 2.65$, $P = .013$, and 9-hour TIB conditions, mean (SD) = 0.07 (0.12); $t = 4.56$, $P < .001$, both had comparable performance in the 6.5-hour TIB condition ($t = 1.99$, $P = .06$).

Proportional Biases

Across all TIBs, M10 demonstrated significant increases in TST and WASO estimation biases with increasing sleep durations (Table 3). The amount of underestimation increased by 0.40 to 0.94 minute per minute of TST, whereas the amount of overestimation increased by 0.75 to 1.31 minutes per minute of WASO ($F \geq 10.72$, $P \leq .003$). H5 and Fitbit demonstrated a similar relationship for TST only in the 6.5-hour TIB condition (H5: $B = 0.65$ minutes, $F = 17.18$, $P < .001$; Fitbit: $B = 0.69$ minutes, $F = 12.01$, $P = .002$); no other TIB condition demonstrated significant relationships ($F \leq 3.98$, $P \geq .06$). However, H5 and Fitbit demonstrated increasing estimation biases for WASO across all TIBs, with H5 biases increasing by 0.32 to 0.80 minutes ($F \geq 6.76$, $P \leq .012$) and Fitbit biases increasing by 0.31 to 0.86 minutes ($F \geq 4.46$, $P \leq .039$) per minute of WASO. There was generally no significant relationship between the amount of bias by Fitbit and sleep stage duration across all TIBs ($F \leq 1.56$, $P \geq .223$). The exception to this was a decrease in the magnitude of stage N3 sleep (deep sleep) bias by 1.03 minutes in the 6.5-hour TIB condition ($F = 12.33$, $P = .002$) and by 0.84 minutes in the 9-hour TIB condition ($F = 9.81$, $P = .003$) per minute of stage N3 sleep (deep sleep).

DISCUSSION

We assessed how well a contemporary consumer-grade wearable device assessed sleep compared to a research-grade actigraph and PSG. At default settings, both Fitbit Alta HR and AW2 performed comparably. Both devices systematically underestimated sleep in adolescents by an

Table 4—Confusion matrices of each device setting by TIB group.

			M10		H5		Fitbit	
			Sleep	Wake	Sleep	Wake	Sleep	Wake
PSG	Sleep	5-hour TIB	**0.90 (0.04)**	0.10 (0.04)	**0.95 (0.02)**	0.05 (0.02)	**0.90 (0.03)**	0.10 (0.03)
		6.5-hour TIB	**0.91 (0.04)**	0.10 (0.04)	**0.95 (0.02)**	0.05 (0.02)	**0.90 (0.03)**	0.10 (0.03)
		9-hour TIB	**0.91 (0.04)**	0.09 (0.04)	**0.95 (0.02)**	0.05 (0.02)	**0.90 (0.03)**	0.10 (0.03)
	Wake	5-hour TIB	0.20 (0.21)	**0.80 (0.21)**	0.36 (0.24)	**0.64 (0.24)**	0.12 (0.10)	**0.88 (0.10)**
		6.5-hour TIB	0.14 (0.10)	**0.86 (0.10)**	0.27 (0.16)	**0.73 (0.16)**	0.10 (0.09)	**0.90 (0.09)**
		9-hour TIB	0.19 (0.16)	**0.81 (0.16)**	0.33 (0.20)	**0.67 (0.20)**	0.12 (0.09)	**0.88 (0.09)**

Mean (standard deviation) of proportions, referenced to PSG, of sleep/wake agreements are displayed. Bold values indicate specificities for sleep/wake categories; the classification accuracy of epochs into sleep or wake. H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, PSG = polysomnography, TIB = time in bed.

Table 5—Confusion matrices of Fitbit sleep staging by TIB group.

			Fitbit			
			Light Sleep	Deep Sleep	REM Sleep	Wake
PSG	Stage N1 + N2 Sleep	5-hour TIB	**0.71 (0.06)**	0.06 (0.04)	0.11 (0.07)	0.12 (0.04)
		6.5-hour TIB	**0.68 (0.08)**	0.06 (0.04)	0.13 (0.06)	0.13 (0.04)
		9-hour TIB	**0.71 (0.06)**	0.08 (0.04)	0.10 (0.05)	0.11 (0.03)
	Stage N3 Sleep	5-hour TIB	0.43 (0.13)	**0.50 (0.14)**	0.02 (0.03)	0.05 (0.03)
		6.5-hour TIB	0.40 (0.12)	**0.51 (0.12)**	0.02 (0.03)	0.06 (0.04)
		9-hour TIB	0.31 (0.14)	**0.64 (0.14)**	0.01 (0.02)	0.04 (0.02)
	REM Sleep	5-hour TIB	0.15 (0.11)	0.00 (0.01)	**0.74 (0.15)**	0.11 (0.06)
		6.5-hour TIB	0.16 (0.11)	0.01 (0.03)	**0.73 (0.15)**	0.11 (0.07)
		9-hour TIB	0.16 (0.10)	0.00 (0.01)	**0.72 (0.14)**	0.12 (0.07)
	Wake	5-hour TIB	0.09 (0.07)	0.02 (0.06)	0.02 (0.02)	**0.88 (0.10)**
		6.5-hour TIB	0.07 (0.05)	0.01 (0.04)	0.03 (0.03)	**0.90 (0.09)**
		9-hour TIB	0.08 (0.06)	0.01 (0.03)	0.03 (0.03)	**0.88 (0.09)**

Mean (standard deviation) of proportions, referenced to PSG, of each sleep stage classification are displayed. Bold values indicate classification accuracies for each sleep stage category. H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, PSG = polysomnography, REM = rapid eye movement, TIB = time in bed.

average of approximately 30 minutes. This underestimation increased when sleep opportunity was lengthened, likely as a result of reduced sleep efficiency (greater wake within a given sleep opportunity) and a tendency of these devices to overestimate wakefulness. Reducing motion sensitivity during sleep in the AW2 device yielded TST and WASO measurements closer to those obtained with PSG, but at the expense of slightly worsened detection of wakefulness. Fitbit estimation of sleep stages was good for stage N1 and N2 sleep, as well as REM sleep, but stage N3 sleep was systematically underestimated.

Consumer-Grade Actigraphy Has Caught Up With Research-Grade Devices

A key finding of the current work is that at default settings, and for the assessment of adolescent sleep, the Fitbit Alta HR performed comparably with the research grade AW2, costing about three times more. Additionally, Fitbit readouts showed less deviation in TST measurement relative to PSG than Actiwatch 2 at default (medium sensitivity) settings. Fitbit demonstrated the best wake specificities of all device settings considered across the different TIB conditions and this likely reflects the benefit of incorporating heart rate sensing to the classification of sleep and wake. Current findings document a clear advance of consumer wearable devices for the purpose of measuring sleep relative to prior comparisons between consumer and research devices. An additional advantage of Fitbit devices is that sleep data can be wirelessly synchronized by participants to a data cloud, allowing monitoring of data as it is collected. This saves time from having to physically download data using a proprietary dock one unit at a time as when using conventional research devices.

Actigraphy Underestimates Sleep of Healthy Adolescents

Currently, there are conflicting data about the accuracy of actigraphy for assessing adolescent sleep. In two studies actigraphy underestimated sleep in adolescents[18,19] whereas

Table 6—EBE agreement metrics, referenced to PSG, of each device setting grouped by TIB condition.

	M10	H5	Fitbit	*F*
5-hour TIB				
Sleep-wake accuracy	0.89 (0.04) [a,b]	0.93 (0.02) [a,c]	0.90 (0.03) [b,c]	33.03
Wake specificity	0.80 (0.21) [a,b]	0.64 (0.24) [a,c]	0.88 (0.10) [b,c]	33.85
Sleep sensitivity	0.90 (0.04) [a]	0.95 (0.02) [a,c]	0.90 (0.03) [c]	84.44
Sleep stage accuracies				
Light sleep	–	–	0.71 (0.06)	–
Deep sleep	–	–	0.50 (0.14)	–
REM sleep	–	–	0.74 (0.15)	–
6.5-hour TIB				
Sleep-wake accuracy	0.90 (0.03) [a]	0.94 (0.02) [a,c]	0.90 (0.03) [c]	46.72
Wake specificity	0.86 (0.10) [a]	0.73 (0.16) [a,c]	0.90 (0.09) [c]	33.80
Sleep sensitivity	0.91 (0.04) [a]	0.95 (0.02) [a,c]	0.90 (0.03) [c]	79.84
Sleep stage accuracies				
Light sleep	–	–	0.68 (0.08)	–
Deep sleep	–	–	0.51 (0.12)	–
REM sleep	–	–	0.73 (0.15)	–
9-hour TIB				
Sleep-wake accuracy	0.90 (0.03) [a]	0.93 (0.02) [a,c]	0.90 (0.02) [c]	45.49
Wake specificity	0.81 (0.16) [a,b]	0.67 (0.20) [a]	0.88 (0.08) [b,c]	80.49
Sleep sensitivity	0.91 (0.04) [a]	0.95 (0.02) [a,c]	0.90 (0.03) [c]	119.47
Sleep stage accuracies				
Light sleep	–	–	0.71 (0.06)	–
Deep sleep	–	–	0.64 (0.14)	–
REM sleep	–	–	0.72 (0.14)	–

ANOVAs of sleep sensitivities, wake specificities and sleep-wake accuracies within each TIB were all significant ($P < .001$), even after corrections for sphericity violations. [a]M10 significantly different from H5 ($P < .05$) within each corresponding TIB. [b]M10 significantly different from Fitbit ($P < .05$) within each corresponding TIB. [c]H5 significantly different from Fitbit ($P < .05$) within each corresponding TIB. EBE = epoch by epoch, H5 = Actiwatch 2 high wake threshold with 5 immobile minutes for sleep onset and end, M10 = Actiwatch 2 medium wake threshold with 10 immobile minutes for sleep onset and end, PSG = polysomnography, REM = rapid eye movement, TIB = time in bed.

another found either correct estimation or overestimation of sleep by actigraphy in older adolescents, depending on device sensitivity settings.[16] Conversely, at least two sleep diary + actigraphy studies have shown significant underestimation of adolescent sleep with actigraphy,[22,25] but neither had PSG confirmation of sleep duration.

The current work shows actigraphy to clearly underestimate sleep and to overestimate WASO in healthy older adolescents studied over multiple nights with PSG and over different sleep opportunity durations. A likely reason for sleep underestimation relates to greater movement compared to adults during healthy adolescent sleep.[18,22] Inclusion of data from a clinical population may have masked such increased movement during sleep in an earlier study, as patients tend to move less.[16]

Estimation biases increased with TIB duration. Underestimation of TST was more pronounced at 9-hour TIB compared to 5-hour TIB. Conversely, WASO was overestimated with longer TIB. In our sample, sex effects were not sufficiently significant across different sleep schedules to merit correction. This information regarding estimation biases as a function of adolescence and TIB provides for finer grained customization of sleep evaluation using actigraphy and could make for better estimates of TST and WASO in future consumer wearable devices. The value of "tuning" sleep detection to the patient is illustrated in the comparison between M10 and H5 (lower sensitivity to motion) settings in Actiwatch devices, the latter giving rise to superior accuracy of sleep detection with some tradeoff in the form of reduced sensitivity to wakefulness detection.

Fitbit Sleep Staging

REM sleep estimation by Fitbit was accurate on average across all TIB conditions considered. However, stage N3 sleep (deep sleep) was consistently underestimated. Our findings replicate those of de Zambotti and colleagues.[11] Stage N3 sleep underestimation was more pronounced in shorter TIB conditions as compared to the longer 9-hour TIB condition. The estimation of stage N1 + N2 sleep (light sleep) was also affected by sleep duration, where it was overestimated in the 5-hour TIB condition, and underestimated it in the 9-hour TIB condition.

EBE comparisons of Fitbit-PSG help shed some light onto its overall sleep staging performance. Stage N1 + N2 and REM sleep demonstrated accuracies of approximately 70% across all TIB conditions. Stage N3 sleep classification accuracies were much poorer, especially at shorter recording durations of 5-hour and 6.5-hour TIBs. Contributing to these errors, stage N3 sleep epochs, similar to REM sleep epochs, were most commonly misclassified as light sleep. As sleep restriction has been shown to cause an increase in sympathetic activity evidenced by alterations to HRV,[26] this could have affected accurate staging.

Strengths and Limitations

A key strength of the current study is the evaluation of more than 50 healthy adolescents over multiple nights in carefully controlled settings and with concurrent PSG. The consumer wearable device tested belongs to a new generation of devices where other physiological sensors other than motion (eg, heart rate, skin temperature, conductance) are used to differentiate sleep and wake. A fuller evaluation of newer consumer-grade wearable devices would need to test other devices, and include young adults as well as older persons in the study sample to confirm our suggestion regarding the utility of age and sleep duration customization of sleep measurement to improve accuracy. We are also unable to comment on how this particular wearable device would perform in persons with medical conditions. In addition, because of the limited number of EEG channels afforded by the PSG setup, only C3 and C4 derivations were recorded. This could have affected the scoring of (1) N1 sleep onset based on occipital alpha rhythm attenuation, as well as (2) the amount of N3 recorded as signals from the central electrodes are typically less prominent than those recorded from the frontal electrodes.

Notably, although many previous reports were motivated by investigators seeking to use actigraphy in clinical settings, the rapid growth in adoption of consumer wearable devices is driven by aspirations to improve personal health and wellbeing in mostly healthy persons. The favorable price-to-performance ratio of these new devices makes them very attractive for large-scale longitudinal bio-bank type studies where sleep is being increasingly recognized as a health variable that should be tracked and analyzed when creating models of healthy life styles.

CONCLUSIONS

In healthy adolescents, a new generation of consumer-grade wearable activity/sleep trackers exemplified by the Fitbit Alta HR generates sleep/wake data that are comparable to default settings used in a well-known research actigraph costing about three times more. Age and sleep opportunity should be considered as variables for tuning the performance of such wearable devices for sleep/wake estimation in the future. Wearable device sleep staging, although somewhat adequate for detecting stage N1 + N2 and REM sleep, significantly underestimates stage N3 sleep and consumers should be made aware of this point to allay anxiety when comparing their sleep stages to PSG based norms.

ABBREVIATIONS

AW2, Actiwatch 2
EBE, epoch-by-epoch
EEG, electroencephalography
EMG, electromyography
EOG, electrooculography
H5, Actiwatch high wake threshold with 5 immobile minutes for sleep onset and end
HRV, heart rate variability
M10, Actiwatch medium wake threshold with 10 immobile minutes for sleep onset and end
PSG, polysomnography
REM, rapid eye movement
TIB, time in bed
TST, total sleep time
WASO, wake after sleep onset

REFERENCES

1. International Data Corporation. New wearables forecast from IDC shows smartwatches continuing their ascendance while wristbands face flat growth. https://www.idc.com/getdoc.jsp?containerId=prUS44000018. Accessed August 28, 2018.
2. Dickinson DL, Cazier J, Cech T. A practical validation study of a commercial accelerometer using good and poor sleepers. *Health Psychol Open*. 2016;3(2):2055102916679012.
3. Liang Z, Chapa Martell MA. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *J Healthc Inform Res*. 2018;2(1-2):152–178.
4. Wright SP, Hall Brown TS, Collier SR, Sandberg K. How consumer physical activity monitors could transform human physiology research. *Am J Physiol Regul Integr Com Physiol*. 2017;312(3):R358–R367.
5. Sadeh A, Acebo C. The role of actigraphy in sleep medicine. *Sleep Med Rev*. 2002;6(2):113–124.
6. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;26(3):342–392.
7. Van de Water AT, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography–a systematic review. *J Sleep Res*. 2011;20(1 Pt 2):183–200.
8. de Zambotti M, Baker FC, Colrain IM. Validation of sleep-rracking technology compared with polysomnography in adolescents. *Sleep*. 2015;38(9):1461–1468.
9. Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;16(3):913–917.
10. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int J Behav Nutr Phys Act*. 2015;12:159.
11. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2 compared with polysomnography in adults. *Chronobiol Int*. 2018;35(4):465–476.
12. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep*. 2015;38(8):1323–1330.
13. Takeda T, Mizuno O, Tanaka T. Time-dependent sleep stage transition model based on heart rate variability. *Conf Proc IEEE Eng Med Biol Soc*. 2015;2015:2343–2346.
14. Aktaruzzaman M, Migliorini M, Tenhunen M, Himanen SL, Bianchi AM, Sassi R. The addition of entropy-based regularity parameters improves sleep stage classification based on heart rate variability. *Med Biol Eng Comput*. 2015;53(5):415–425.

15. Werner H, Molinari L, Guyer C, Jenni OG. Agreement rates between actigraphy, diary, and questionnaire for children's sleep patterns. *Arch Pediatr Adolesc Med*. 2008;162(4):350–358.
16. Meltzer LJ, Walsh CM, Traylor J, Westin AM. Direct comparison of two new actigraphs and polysomnography in children and adolescents. *Sleep*. 2012;35(1):159–166.
17. Lo JC, Ong JL, Leong RL, Gooley JJ, Chee MW. Cognitive performance, sleepiness, and mood in partially sleep deprived adolescents: The Need for Sleep Study. *Sleep*. 2016;39(3):687–698.
18. Johnson NL, Kirchner HL, Rosen CL, et al. Sleep estimation using wrist actigraphy in adolescents with and without sleep disordered breathing: a comparison of three data modes. *Sleep*. 2007;30(7):899–905.
19. Pesonen AK, Kuula L. 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*. 2018;14(4):585–591.
20. Patanaik A, Ong JL, Gooley JJ, Ancoli-Israel S, Chee MWL. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*. 2018;41(5).
21. Iber C, Ancoli-Israel S, Chesson AL, Jr, Quan SF. for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.
22. Short MA, Gradisar M, Lack LC, Wright H, Carskadon MA. The discrepancy between actigraphic and sleep diary measures of sleep in adolescents. *Sleep Med*. 2012;13(4):378–384.
23. Fitbit Inc. Start sleeping better with Fitbit. https://www.fitbit.com/sg/sleep-better. Accessed July 24, 2018.
24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307–310.
25. Short MA, Gradisar M, Lack LC, Wright HR, Chatburn A. Estimating adolescent sleep patterns: parent reports versus adolescent self-report surveys, sleep diaries, and actigraphy. *Nat Sci Sleep*. 2013;5:23–26.
26. Dettoni JL, Consolim-Colombo FM, Drager LF, et al. Cardiovascular effects of partial sleep deprivation in healthy volunteers. *J Appl Physiol (1985)*. 2012;113(2):232–236.

ACKNOWLEDGMENTS

The authors are grateful for the invaluable assistance of Alyssa Ng, James Cousins, Amiya Patanaik, Jesisca Tandi, Ruth Leong, Shirley Koh, Ksenia Vinogradova, Andrew Dicom, Shamsul Azrin Jamaluddin, Lydia Teo, James Teng, Kian Foong Wong, and Karen Sasmita in collecting and collating data for this study. We thank all our participants for contributing their time and effort to this study.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication September 15, 2018
Submitted in final revised form January 18, 2019
Accepted for publication January 18, 2019
Address correspondence to: Dr. Michael W.L. Chee, Centre for Cognitive Neuroscience, Duke-NUS Medical School, Singapore 169857;
Email: michael.chee@duke-nus.edu.sg

DISCLOSURE STATEMENT

All authors have reviewed and approve of the manuscript. All adolescent participants and their legal guardians provided written informed consent. This work was approved by the Institutional Review Board of the National University of Singapore, and was supported by the National Medical Research Council, Singapore (NMRC/StaR/015/2013) and the Far East Organization. Alta HR units were provided under an unrestricted gift from Fitbit. The authors report no conflicts of interest.