Nature and Science of Sleep

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

# Evaluations of Commercial Sleep Technologies for Objective Monitoring During Routine Sleeping Conditions

This article was published in the following Dove Press journal:
*Nature and Science of Sleep*

Jason D Stone[1]
Lauren E Rentz[1]
Jillian Forsey[1]
Jad Ramadan[1]
Rachel R Markwald[2]
Victor S Finomore Jnr[1]
Scott M Galster[1]
Ali Rezai[1]
Joshua A Hagen[1]

[1]Rockefeller Neuroscience Institute, West Virginia University, Morgantown, WV, USA; [2] Sleep, Tactical Efficiency, and Endurance Laboratory, Warfighter Performance Department, Naval Health Research Center, San Diego, CA, USA

**Purpose:** The commercial market is saturated with technologies that claim to collect proficient, free-living sleep measurements despite a severe lack of independent third-party evaluations. Therefore, the present study evaluated the accuracy of various commercial sleep technologies during in-home sleeping conditions.

**Materials and Methods:** Data collection spanned 98 separate nights of *ad libitum* sleep from five healthy adults. Prior to bedtime, participants utilized nine popular sleep devices while concurrently wearing a previously validated electroencephalography (EEG)-based device. Data collected from the commercial devices were extracted for later comparison against EEG to determine degrees of accuracy. Sleep and wake summary outcomes as well as sleep staging metrics were evaluated, where available, for each device.

**Results:** Total sleep time (TST), total wake time (TWT), and sleep efficiency (SE) were measured with greater accuracy (lower percent errors) and limited bias by Fitbit Ionic [mean absolute percent error, bias (95% confidence interval); TST: 9.90%, 0.25 (−0.11, 0.61); TWT: 25.64%, −0.17 (−0.28, −0.06); SE: 3.49%, 0.65 (−0.82, 2.12)] and Oura smart ring [TST: 7.39%, 0.19 (0.04, 0.35); TWT: 36.29%, −0.18 (−0.31, −0.04); SE: 5.42%, 1.66 (0.17, 3.15)], whereas all other devices demonstrated a propensity to over or underestimate at least one if not all of the aforementioned sleep metrics. No commercial sleep technology appeared to accurately quantify sleep stages.

**Conclusion:** Generally speaking, commercial sleep technologies displayed lower error and bias values when quantifying sleep/wake states as compared to sleep staging durations. Still, these findings revealed that there is a remarkably high degree of variability in the accuracy of commercial sleep technologies, which further emphasizes that continuous evaluations of newly developed sleep technologies are vital. End-users may then be able to determine more accurately which sleep device is most suited for their desired application(s).

**Keywords:** wearables, consumer sleep technologies, sleep duration, sleep efficiency, sleep staging

## Introduction

With proper adherence to formal guidelines and recommendations, a healthy individual spends nearly one-third of their adult life sleeping.[1] Provided the undeniable importance of adequate sleep durations,[1,2] quantifying trends in sleep behavior introduces an opportunity for enhanced self-monitoring over one's health despite discernible challenges the scientific community has yet to overcome, such as device accuracy and scant third-party validations.[3–5] In the budding era of individualized

Correspondence: Joshua A Hagen
Tel +1 513-560-9879
Email joshua.hagen@hsc.wvu.edu

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/NSS.S270705

health monitoring, primarily as a result of technological developments related to smartphones and wearable technologies such as photoplethysmography (PPG),[3,6] there is a resulting need for effective and actionable personalized sleep monitoring.[7]

Traditionally, human sleep studies are conducted in laboratories or hospital clinic settings using a technique called polysomnography (PSG), which requires a tremendous amount of hardware, staff, and subject-matter expertise to ensure satisfactory data are collected.[8] PSG utilizes multi-lead electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG) to quantify brain, muscle, and eye activity, respectively.[5,6,8,9] In addition to concurrent recording of respiration and pulse oximetry that are used to aid in the diagnosis of sleep disorders, PSG is used to collect signals necessary to differentiate wakefulness and sleep, and to classify sleep stages based on combinations of various physiological states.[5,6,9] While PSG does provide in-depth sleep and sleep staging information, it also possess multiple limitations. PSG data collection is particularly burdensome, expensive, and time-consuming for everyone involved, thus decreasing its practicality for utility in various research objectives, including at-home monitoring.[3–6,9] Additionally, many PSG sleep studies are initiated for a clinical purpose, such as sleep apnea investigation,[10] and data from clinical subjects are unlikely to represent the entire population of sleep device users. To account for the aforementioned shortcomings inherent to PSG, other sleep assessments have been developed and often comprises subjective measures of sleep, such as the Pittsburgh Sleep Quality Index or sleep diaries, which provide greater degrees of practicality via ease of use, although their limitations primarily manifest in the form of poor veracity.[11,12] Subjective methods of assessing sleep that attempt to characterize one's personal perception(s) of their sleep do indeed offer valuable insights (eg, at home sleeping and not in a laboratory), granted the ability to accurately and objectively quantify sleep during routine sleeping conditions is still justified.

The middle ground between the highly definitive PSG and more convenient but less informative subjective sleep measures is actigraphy, which utilizes physical movement data that is later processed by the researcher or clinician to differentiate sleep and wake states.[13–16] Although actigraphy is widely accepted as being much more accommodating to routine sleep monitoring, particularly in the general population as compared to PSG, there are still inherent limitations to this methodology as well.[4,17] For example, actigraphy relies heavily on wrist movement to differentiate sleep episodes from daily activity. Awakenings that occur after sleep onset may not yield robust enough wrist movements such that this brief moment of wakefulness may be wrongfully characterized as sleep, which lends to consistent overestimation of sleep and underestimation of wake during actigraphic sleep assessments.[18–20] Further, insight on physical movement patterns alone are insufficient to categorize sleep into stages thus actigraphy is merely limited to reporting sleep and wake durations. Actigraphy also lacks a user-interface for real-time feedback on sleep behaviors thus engagement between the user and their personal data is restricted.

The aforementioned limitations associated with PSG, subjective sleep surveys and diaries, as well as actigraphy, exposes the necessity for a new middle ground: a technology capable of considerable accuracy and simplicity with instantaneous biofeedback for the clinical, research and consumer bases alike.[4,17] For instance, those in high performing populations (eg, athletics, military, first responders) may use automated, real-time feedback on their sleep quantity and quality to assist in determining their daily workload capacities, ultimately in an effort to optimize daily readiness and performance. Recent proliferations to the wearable technology consumer market provide possible alternatives to objective sleep monitoring that would significantly mitigate the existing limitations to other strategies for measuring sleep. Indeed, the number of technologies purporting to measure sleep quantity and quality continues to demonstrate exponential growth as they elude thorough independent third-party verifications of device claims in an effort to reach the consumer market faster.[3–5,7]

Despite the diversified wearable market, few commercial companies prioritize third-party assessments of their devices' advertised capabilities,[3–6] which range from claims to accurately measure total sleep time (TST) and total wake time (TWT) within a sleep opportunity to more rigorous derivations such as sleep staging (eg, duration of deep sleep). Sleep efficiency (SE), a percentage value that denotes the amount of TST relative to time in bed (sleep opportunity) is another common metric reported by commercial sleep devices. Additionally, many of these technologies report measures of heart rate [eg, mean heart rate during sleep], as fluctuations in nighttime cardiovascular physiology serves as a strong indicator for sleep quality and recovery, or lack thereof.[21,22] Still, the mere reporting of these various physiological metrics (sleep and/or heart rate) may be severely limited by the accuracy in the computation of the variables. Previously, PPG technologies demonstrated vulnerability to error, most notably through motion artifacts[23,24] and poor reliability across various skin complexions.[23,25,26] Undeterred by the lack of validation,

manufacturers continue developing sleep technologies despite the limited public knowledge regarding their accuracy and reliability.[3–5,27]

Therefore, the purpose of this study was to evaluate the sleep metrics reported by several commercial devices, which were compared to a previously validated home-based EEG sleep device.[28,29] Differing from most research that aims to assess device veracity, the assessment of sleep in a free-living environment provides the opportunity to attest to the accuracy of various technologies in the environment in which their use is most sought. Sleep laboratories, where current gold-standard device validations occur, present many limitations for evaluating commercial sleep technologies. Not only are devices generally donned immediately prior to getting into bed but the unfamiliar setting and scheduling as part of PSG evaluation undoubtedly has an effect on accurate assessments of normal sleep patterns. As such, it is possible that device accuracy outside of the sleep laboratory may present differently as compared to much of the published literature that exists. Indeed, there is a recognized gap in the literature that examines the efficacy of commercial sleep technologies in their intended environments (free-living) as a solution for convenient and routine objective sleep monitoring.[17] Third-party reporting on commercial device accuracy for sleep monitoring affords consumers (eg, researchers, clinicians, general population) the ability to make informed decisions on optimal strategies for quantifying trends in at-home sleep behavior and physiology.

# Materials and Methods

All procedures contained herein were approved by the Institutional Review Board of West Virginia University for human subject's research and were compliant with the Declaration of Helsinki guidelines. Written consent was obtained from each participant prior to engagement.

## Inclusion/Exclusion Criteria

Inclusion into the study required participants to be considered "low risk" based on the American College of Sports Medicine (ACSM) Risk Stratification.[30] While these guidelines are generally used for health screening prior to exercise testing, they also encompass many risk factors pertaining to overall health and disease. Pre-screened subjects were excluded from the study if the risk stratifications deemed them as anything other than "low risk." The ACSM classification requires individuals to present with no signs or symptoms of cardiovascular, metabolic, or renal disease, as well as identify obese, sedentary, hypertensive, or pre-diabetic individuals and those who smoke, have dyslipidemia, or a family history of heart disease. These criteria were used to ensure participants were healthy in an attempt to utilize a representative subset of the general (healthy) population. Further, individuals with known sleep or circadian rhythm disorders were also excluded.

## Subjects

In the present study, five healthy adults, two males (ages 41 and 26 years) and three females (ages 22, 23, and 27 years), volunteered to participate for a combined total of 98 nights. Participants underwent an initial health screening followed by informed consent and a familiarization session to ensure all were adept at operating the various sleep devices. Enrollment was organized such that no more than two participants were enrolled in parallel (equipment resources only allowed for the utilization of two Sleep Profilers simultaneously) at any given time. Participation in the study continued for as long as the participants were willing to wear the various devices while sleeping, which ranged from 12 to 25 nights. The average duration of enrollment per participant was 19.6 nights.

## Validation Standard: Free-Living Electroencephalography

Due to the limitations presented by PSG for in-home sleep monitoring, the Sleep Profiler (Advanced Brain Monitoring, California, United States), a technology deemed substantially equivalent to PSG by the FDA,[34] was utilized as the in-home standard for accurate data collection in the present study. Previous research efforts successfully validated the Sleep Profiler against the industry accepted gold standard, PSG.[28,29] This take-home EEG sleep device affixes to the forehead for collection of physiological data from three frontopolar channels, which generate output signals for EEG, EOG, and EMG.[28] Additionally, the technology is equipped with infrared PPG, a three-axis accelerometer, and an acoustic microphone.

## Commercial Sleep Devices

A total of eight, popular commercial sleep devices were assessed, which included the Apple Watch Series 3 (Apple, Cupertino, California, United States), Beddit Sleep Monitor 3.0 (previously Beddit, Espoo, Finland, manufacturing now owned by Apple), Fatigue Science Readiband (Fatigue Science, Vancouver, British Columbia), Fitbit Ionic (Fitbit, San Francisco, California, United States), Garmin Vivosmart 4 (Garmin, Olathe, Kansas, United States), 2nd generation Oura smart ring (ŌURA, Oulu, Finland), Polar A370 (Polar,

Kempele, Finland), and the WHOOP Strap 2.0 (WHOOP, Boston, Massachusetts, United States). Each device syncs via Bluetooth to a commercially developed smartphone application respective to each device. All applications were a product of the device manufacturer, except for the Apple Watch in which two third-party applications, SleepWatch (Bodymatter Incorporated, Newport Beach, California, United States) and Sleep++ (Cross Forward Consulting, LLC, Herndon, Virginia, United States), were concurrently used for interpretation of data. Therefore, nine separate commercial sleep entities were examined comprising eight tangible devices; due to the remaining potential for significant variation in the two Apple Watch applications' ability to interpret data, separate analyses were conducted thus the two applications were treated as separate devices. Relevant technical specifications that pertain to the mechanisms of each tangible device are listed in Table 1.

**Table 1** Technical Specifications for Sleep Monitoring Devices

| Device Name | Device Type | Measurement Strategy | PPG Technicalities | Battery Life |
|---|---|---|---|---|
| Sleep Profiler | Head mounted wearable | EEG (5 channel)<br>EOG<br>EMG<br>PPG<br>3D Accelerometer<br>Acoustic Microphone | Contact PPG<br>Infrared<br>Reflective | 30 hours |
| Apple Watch Series 3 | Wrist-based wearable | PPG<br>Accelerometer<br>Gyroscope | Contact PPG<br>Green LED &<br>Infrared<br>Reflective | 18 hours |
| Beddit Sleep Monitor 3.0 | Mattress affixed monitor (placed under upper body) | Piezoelectric force sensor<br>Capacitive touch sensor | N/A | N/A |
| Fatigue Science Readiband | Wrist-based wearable | 3D Accelerometer | N/A | 30 days |
| FitBit Ionic | Wrist-based wearable | PPG<br>Accelerometer | Contact PPG<br>Red LED & Infrared<br>Reflective | 5 days |
| Garmin Vivosmart 4 | Wrist-based wearable | PPG<br>Accelerometer | Contact PPG<br>Green LED<br>Reflective | 7 days |
| Oura Smart Ring (2nd Gen.) | Finger-based wearable | PPG<br>3D Accelerometer<br>Gyroscope<br>NTC temperature sensors | Contact PPG<br>Infrared<br>Transmission | 7 days |
| Polar A370 | Wrist-based wearable | PPG<br>Accelerometer | Contact PPG<br>Green LED<br>Reflective | 4 days |
| WHOOP Strap 2.0 | Wrist-based wearable | PPG<br>3D Accelerometer<br>3-Axis Gyroscope | Contact PPG<br>Green LED<br>Reflective | 2 days |

**Notes:** Table 1 includes a summary of the technical specifications for the validated Sleep Profiler as well as the various commercial sleep monitoring technologies. Device specifications, which to our knowledge are accurate, were derived from the respective company website and/or support centers.

**Abbreviations:** EEG, electroencephalography; EMG, electromyography; EOG, electrooculography; Gen, generation; LED, light emitting diode; N/A, not available; NTC, negative temperature coefficient; PPG, photoplethysmography; 3D, three dimensional.

## Experimental Design

To assess device accuracy for quantifying sleep quantity and quality, participants were asked to utilize several commercial devices concurrent to the Sleep Profiler. Since commercial sleep technologies are machines that attempt to quantify sleep metrics, they do so indiscriminately such that they will deliver objective assessments whenever they properly donned by any human being. Therefore, to provide these technologies with an ideal environment to perform optimally, the number of human subjects was small whereas the iterations to generate a total of 98 sleep nights for analysis was not. Similar experimental designs were deployed in previous investigations that sought to examine the accuracies of commercial technologies.[31–33] Participants were sized for their devices and then instructed on proper placement and usage of each technology. They were also provided a Wi-Fi enabled 6th generation iPod Touch that contained all of the commercial smartphone applications, which synced with the respective devices after each night of sleep. The Beddit, Oura smart ring, and Fatigue Science Readiband were used on randomized nights in addition to the Sleep Profiler. The wrist-based devices that utilized PPG (Apple Watch Series 3, Fitbit Ionic, Garmin Vivosmart 4, Polar A370, WHOOP Strap 2.0) were alternated nightly to mitigate sensor interference.

Participants were instructed to don all devices approximately 30 minutes before getting into bed. The Sleep Profiler was put on at the same time as the other devices but was not powered on until the individual went to bed. This schedule of events was vital for assessing the accuracy of each device in determining the state of its user, and thus, was meant to allow for some room in measurement variability. The finger in which the Oura ring was worn and the wrist for which the wrist-based devices were worn was not specified; however, once the decision was made by the participant they were encouraged to maintain consistent placement throughout the duration of the study. In order to minimize any potential issues with optical interference between PPG devices, only one PPG device was allowed to be worn on each wrist. The combination of sensors was randomized, and the subject was given instructions on which devices to utilize each night. Following each night, all devices that utilized a battery were fully charged to ensure consistency in battery level throughout each trial.

All devices, aside from the Sleep Profiler, were returned to the laboratory at the end of the enrollment period for data export. The Sleep Profiler was returned to the research laboratory every three days for data export and memory clearance to ensure data quality and enable additional nights of data collection. Data from the Sleep Profiler were uploaded for processing by the Advanced Brain Monitoring (Advanced Brain Monitoring, Carlsbad, California, United States) interface.[28] The Advanced Brain Monitoring interface uses an algorithm to analyze the various data streams collected by the Sleep Profiler; while reports provide hypnograms depicting the staging patterns throughout the night, only the nightly totals were used to summarize each sleep stage. Each commercial device assessed that attempted to stage sleep did so via automatic algorithms that classified sleep as either "light" or "deep." While most companies fail to disclose details regarding their algorithms, in most cases, "light sleep" refers to the aggregation of N1 and N2, whereas "deep sleep" generally refers to N3.[17] Considering this, the Sleep Profiler reports of N1 and N2 were summated as "light sleep" and N3 was used as comparison for "deep sleep." Data from each commercial device were manually extracted from its respective app. All data were then organized into a centralized database in version 16 of Microsoft Excel (Microsoft, Redmond, Washington, United States) for later analysis.

Lastly, it is worth noting that the collection of data from all commercial technologies occurred in a manner consistent with what is directly available to the consumer following the initial device purchase through use of the corresponding smartphone application (app). Regardless of device, additional settings or modes were not utilized throughout the study. Due to the lack of available epoch data manufacturers often provide to the consumer, aside from the Sleep Profiler, no raw or epoch-specific data were requested, acquired, or used for data analysis. Data available via the respective application for each device were merely recorded and compiled for statistical analysis. All data were manually extracted from each device's smartphone application and compiled into Microsoft Excel. Although true device validations are only accomplished via direct epoch by epoch comparisons to PSG, a position statement from the Sleep Research Society on wearable sleep technologies acknowledged that real-world assessments (eg, longitudinal at-home monitoring) for accuracy of commercial sleep technologies is a gap in the extant sleep wearables literature.[17]

## Statistical Analysis

Comparisons between the commercial devices and the Sleep Profiler were made for any "Yes" listings in Table 2 as these were sleep variables concomitantly tracked by the respective devices. Due to the varying enrollment periods, the number of nights assessed for each device also varies and are, where relevant, denoted in the tables that are later described. Variables of interest included TST, TWT, SE, as well as durations for light, deep, and REM sleep such that each of these metrics were extracted when and where the opportunity (commercial sleep technology) presented itself. In order to evaluate device performance relative to the Sleep Profiler, absolute percent error (APE) calculations were executed as well as Bland–Altman analyses, which exposed instances in which a commercial sleep technology had a propensity for over or underestimating a sleep metric.

Commercial sleep technologies attempting to quantify sleep stage durations tend to broadly group the stages as light, deep, and rapid eye movement (REM) sleep. This differs from the more specific differentiation between the three stages of non-REM sleep and REM sleep as the latter classification strategy is characterized by subtle changes in brain wave activity,[35] which presumably extends beyond the scope of most commercial sleep technologies. Therefore, device evaluations were executed for light, deep and REM sleep. Furthermore, for each of the aforementioned comparisons, the absolute percent error (APE) was calculated as

$$\frac{|\text{Device Measurement} - \text{Sleep Profiler}|}{\text{Sleep Profiler}} * 100$$

**Table 2** Sleep Assessment Variables by Device

| Device | TST | TWT | SE | Light Time | DeepTime | REM Time |
|---|---|---|---|---|---|---|
| Beddit | Yes | Yes | Yes | No | No | No |
| Fatigue Science | Yes | Yes | Yes | No | No | No |
| Fitbit | Yes | Yes | Yes | Yes | Yes | Yes |
| Garmin | Yes | Yes | Yes | Yes | Yes | Yes |
| Oura | Yes | Yes | Yes | Yes | Yes | Yes |
| Polar | Yes | Yes | Yes | No | No | No |
| Sleep++ | Yes | No | No | No | No | No |
| SleepWatch | Yes | No | No | No | No | No |
| WHOOP | Yes | Yes | Yes | Yes | Yes | Yes |

**Note:** Table 2 differentiates the sleep variables reported by each of the various technologies utilized herein.

**Abbreviations:** REM, rapid eye movement; SE, sleep efficiency; TST, total sleep time; TWT, total wake time.

For the calculation listed above, "Device Measurement" refers to each of the commercial technologies assessed. The summary statistics for APE for each sleep variable derived from each device were also calculated.[36,37] It should be noted that the magnitude of percent errors using this method of APE calculation will be influenced by the overall value of the metric being assessed. For instance, sleep efficiency values have low standard deviations on a percentage basis of the mean, so APE values will be much smaller for all devices in SE compared to metrics, like TWT and sleep staging durations, where there are much higher standard deviations. Since this study aims to not only assess validity of sleep data but to compare devices to one another, APE values are utilized but care should be taken in interpreting these values, which is detailed in the results below. The three sleep summary metrics of TST, TWT, and SE are all related as discussed above, and are available for the majority of the devices. Thus, a "composite" ranking method using these three variables was desired to help assess the consistency of device performance across those three metrics. In other words, an assessment of whether the same device(s) ranked near the top with respect to each metric. Kendall's coefficient of concordance (Kendall's W) may be used to answer this question.[38] Traditionally, Kendall's W is used to quantify "human" inter-observer agreement, and has been used extensively in literature, with applications in digital imagery[39] and ecology.[40] In this case, TST, TWT, and SE act as the "observers" who "rate" how well (or not) the device measures the sleep metrics compared to the Sleep Profiler. Once Kendall's W was calculated, a hypothesis test was constructed to determine if significant concordance (consistency in device performance) was achieved. In the case of a significant result, the three rankings per device were summed together to create an overall "composite" score, with lower-scoring devices (justifiably) being the top performers.

Another procedure of interest included Bland–Altman analyses. Previous research implemented this strategy in other wearable device validation publications.[32,37,41–45,] A Bland–Altman analysis provides worthy visualizations of the bias (whether the device over/underestimates the Sleep Profiler) and Limits of Agreement (LOA; range) between the Sleep Profiler and device measurement.[46] An individual Bland–Altman analysis was executed for each sleep metric and device combination, with proportional biases (respective $r^2$ and $p$-values) applied to On the individual level, further analysis on the bias included

a significance test to determine if the mean bias significantly differed from 0 (the value which represents "unbiasedness"). This was achieved by performing a two-sided, one-sample *t*-test on the differences in measurements between the device and Sleep Profiler with respect to each metric. Since there were a total of 55 sleep metric and device combinations where Sleep Profiler comparisons were possible, an adjustment to account for multiple comparisons was deemed necessary. Thus, a Bonferroni correction was applied such that each resulting *p*-value was multiplied by 55 to create an "adjusted" *p*-value to compare to the selected significance level of 0.05. Results where $p < 0.05$ indicate that the mean bias is significantly different from 0 for measurements from the sleep metric and device combination in question. The sign of the mean bias can then be examined to determine whether the measurements over/underestimate the Sleep Profiler.

Another potential type of bias that can occur within an individual Bland–Altman analysis is called proportional bias, which is tested by constructing a simple linear regression model using the difference in device and Sleep Profiler measurement as the outcome variable, and the average of the device and Sleep Profiler measurements as the independent variable. Significance was assessed by performing a *t*-test using the resulting $R^2$ value. As before, a Bonferroni correction (for 35 comparisons) was applied to each *p*-value. A significant result ($p < 0.05$) indicates that the $R^2$ value is greater than 0, which means that the difference between the two devices is dependent on the average.

The present investigation sought to examine objective data derived from the commercial devices. However, to account for user error (eg, forgot to charge battery, device placement shifted during the night), we elected to examine for extreme outliers. The Tukey boxplot-based outlier detection rule was used (with respect to APE) to check for extreme outliers.[47] The rule incorporated herein labels a datapoint as an extreme outlier if it is outside of the outer fence of the boxplot. The outer fence of a boxplot is defined as 3*Interquartile Range (IQR) above the third quartile, or 3*IQR below the first quartile (eg, Q1 – 3*IQR and Q3 + 3*IQR).[48] This is in contrast to the traditional rule using inner fences, where one would utilize 1.5*IQR to identify outliers. Using the more traditional rule, too many observations would have to be removed, which significantly changes the analysis of how well the device performed. A total of 55 extreme outliers (across all metrics) were identified according to the aforementioned 3*IQR rule for detection, which explains why there are discrepancies in trial numbers by a respective device observed in the tables and figures to follow. For a descriptive table that provides additional details on the context of extreme outliers by individual subjects and devices that were identified, please refer to the supplemental materials (Table S1 and S2, respectively). Moreover, in an attempt to better clarify data outliers and incongruencies in trial numbers, Kruska Wallis tests were used to assess for differences in APE across subjects per variable, with Steel-Dwass multiple comparisons where appropriate.[49] Results are also included in the Supplemental Materials (Tables S3–5 and Figures S1–3).

All analyses were conducted using R version 4.0.0.[50] Data pre-processing was handled using the *tidyverse* package,[51] whereas plots were constructed using the *gridExtra* package[52] along with the *tidyverse* packages. To assist with the calculation of Bland–Altman statistics, the *blandr* and *rstatix* packages were utilized.[53,54]

## Results

First, an assessment as to whether or not error distributions were consistent across participants was executed. With respect to TST, significant differences were observed (Kruskal Wallis Test Statistic = 54.48, $df = 4$, $p<0.001$). However, the Steel-Dwass method identified subject 4 as the only participant with significant differences from the rest (See Table S6 in supplemental materials). None of the remaining subjects had significantly different error distributions from each other. Further details on individual variations for TST, TWT, and SE can be found in Supplemental Tables S3–S6.

APE analysis for TST is depicted as box plots in Figure 1. The precise APE summary values for TST for each device are listed in Table 3, which includes notation of the total number of nights the various devices were worn. The largest MAPE values observed were 21.94% (Garmin) and 22.84% (SleepWatch) whereas the smallest MAPE values were 7.39% (Oura) and 8.78% (WHOOP). Per the Bland–Altman plots (Figure 2A–I) and summary statistics for TST (Table 4), Beddit ($p < 0.001$), Fatigue Science ($p < 0.001$), and Sleep++ ($p = 0.01$) were the only devices whose measures resulted in a significant bias. Further examination of proportional biases (Table 5) revealed that significance was not reached for any of the devices.

Consistent trends across all devices were observed in their failure to determine the amount of time its user was awake rather than sleeping. TWT was measured by seven of the devices: Beddit, Fatigue Science, Fitbit, Garmin, Oura, Polar, and WHOOP (Figure 3). Measurements for

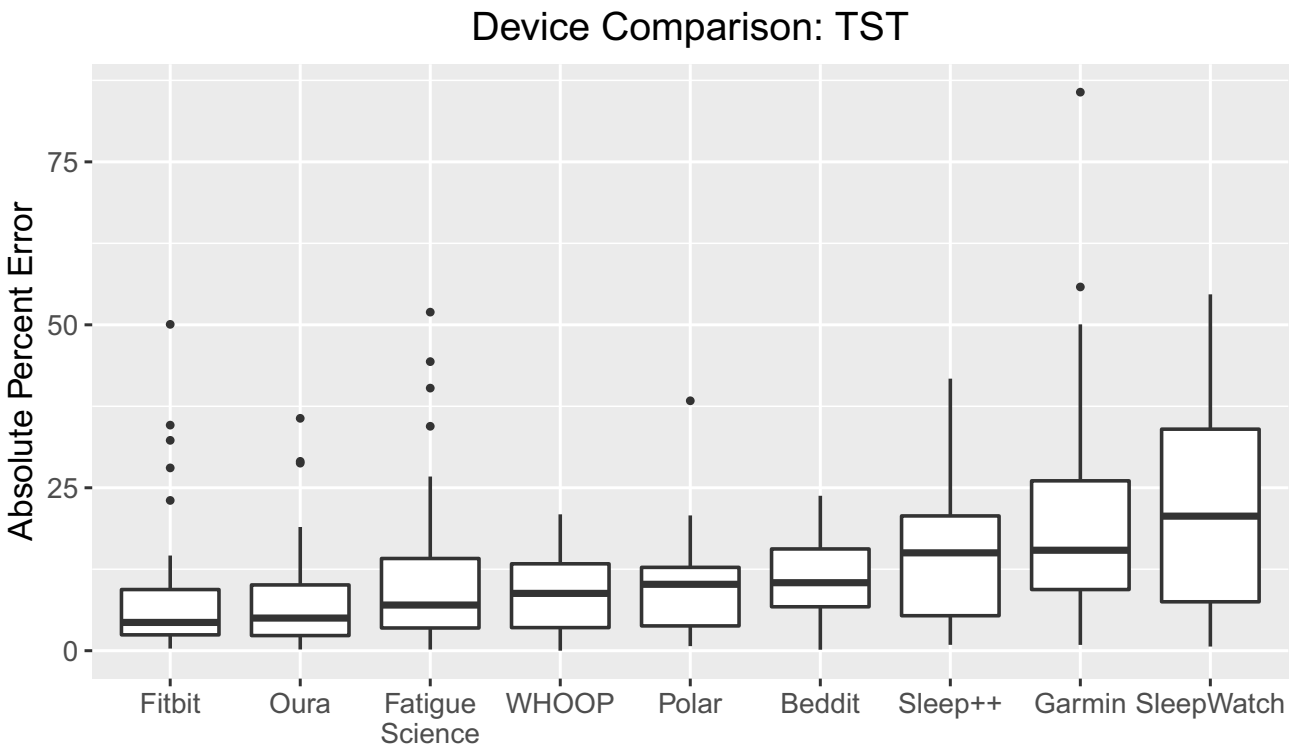


**Figure 1** TST boxplots: absolute percent error by device.
**Notes:** Figure 1 depicts a box plot representation of absolute percent error (APE) for each commercial device that reported on total sleep time.
**Abbreviation:** TST, total sleep time.

TWT, across all devices, demonstrated a much higher variability in error margins as compared to TST, with MAPE values that range from 25.64% (Fitbit) to 89.04% (Garmin), as observed in Table 3. Per Figure 4A–G and Table 4, all devices underestimated TWT, with all devices other than Fitbit ($p = 0.11$) and Oura ($p = 0.40$) exhibiting a significant bias. Additionally, proportional bias analysis demonstrated that Garmin significantly underestimated TWT as TWT durations increased ($p < 0.001$), whereas all other devices did not reach statistical significance (Table 5). Granted, after the aforementioned outlier removal Garmin still possessed a high leverage point that presumably influenced reaching significance (Figure 4D).

Similarly, as sleep efficiency is directly proportional to TST relative to TWT, consistent trends observed in both TST and TWT were generally well reflected in the findings for SE. Indeed, SE calculations illustrated a lower variability in the margins of error as compared to TST and TWT (see Figure 5), and the unit of percent rather than time is a factor in this difference. For example, MAPE for SE ranges from 3.49% (Fitbit) to 12.15% (Garmin), which is expanded in Table 3. Albeit consistent in trends that were synonymous to TWT, Fitbit and Oura were the only two devices that did not significantly differ from the measures compared to the Sleep Profiler for SE (according to Bland–Altman; see Figure 6A–G and Table 4). In fact, the same five devices that significantly underestimated TWT also significantly overestimated SE. With respect to the examination of proportional bias, there were no remarkable differences between any of the devices compared to the Sleep Profiler (Table 5).

Provided that it is possible to have two devices that report the same exact SE value despite different TST and TWT values, Kendall's W was calculated for MdAPE (0.94) as well as MAPE (0.87). The high degree of concordance for MAPE suggests that the summated rankings reveal an accurate depiction for how well the devices performed across TST, TWT, and SE (see Table 6). With respect to MAPE, Fitbit and Oura tied for the lowest (most accurate) summated ranking as these two devices outperformed all others with Garmin eliciting the highest (least accurate) ranking.

With regard to sleep staging, degrees of variability and inaccuracy vary by stage. The devices that purported to divide sleep into subcategories (eg, sleep stages) were Fitbit, Garmin, Oura, and WHOOP, which all categorized

**Table 3** Absolute Percent Error Executive Summary Statistics: All Devices

| Metric | Device | n | MAPE | Min. (%) | MdAPE | Max. (%) | IQR |
|---|---|---|---|---|---|---|---|
| TST | Oura | 59 | 7.39 | 0.19 | 5.01 | 35.65 | 7.77 |
| | WHOOP | 32 | 8.78 | 0.00 | 8.80 | 20.92 | 9.79 |
| | Fitbit | 27 | 9.90 | 0.33 | 4.34 | 50.07 | 6.95 |
| | Fatigue Science | 65 | 10.35 | 0.18 | 7.01 | 51.94 | 10.66 |
| | Polar | 19 | 10.89 | 0.70 | 10.19 | 38.34 | 8.98 |
| | Beddit | 38 | 11.11 | 0.15 | 10.43 | 23.77 | 8.87 |
| | Sleep++ | 16 | 15.61 | 0.90 | 15.01 | 41.74 | 15.31 |
| | Garmin | 21 | 21.94 | 0.89 | 15.42 | 85.69 | 16.68 |
| | SleepWatch | 13 | 22.84 | 0.64 | 20.64 | 54.68 | 26.48 |
| TWT | Fitbit | 27 | 25.64 | 0.00 | 25.00 | 79.17 | 22.41 |
| | Oura | 60 | 36.29 | 1.37 | 31.67 | 116.13 | 32.88 |
| | Fatigue Science | 70 | 43.73 | 3.23 | 42.55 | 140.58 | 37.44 |
| | Polar | 19 | 54.69 | 9.68 | 53.62 | 88.43 | 30.50 |
| | WHOOP | 33 | 60.34 | 3.85 | 70.21 | 100.00 | 30.97 |
| | Beddit | 40 | 62.37 | 2.13 | 66.74 | 125.35 | 31.40 |
| | Garmin | 21 | 89.04 | 57.90 | 93.64 | 100.00 | 12.49 |
| SE | Fitbit | 28 | 3.49 | 0.25 | 3.44 | 12.78 | 3.18 |
| | Oura | 60 | 5.42 | 0.01 | 4.79 | 19.21 | 4.74 |
| | Fatigue Science | 69 | 6.52 | 0.13 | 5.56 | 18.46 | 7.43 |
| | Polar | 19 | 9.03 | 0.04 | 5.50 | 28.48 | 8.49 |
| | WHOOP | 32 | 10.00 | 0.04 | 9.24 | 35.17 | 7.75 |
| | Beddit | 40 | 10.18 | 0.17 | 9.40 | 30.60 | 10.24 |
| | Garmin | 20 | 12.15 | 7.24 | 11.74 | 23.05 | 3.76 |

**Notes:** Table 3 includes the APE summary statistics for all devices that reported on each respective measure of sleep. Data presented are for TST, TWT, and SE, and are sorted lowest-to-highest for each metric based upon MAPE.

**Abbreviations:** APE, absolute percent error; IQR, interquartile range; MAPE, mean absolute percent error; Max, maximum; MdAPE, median absolute percent error; Min, minimum; n, number of trials; SE, sleep efficiency; TST, total sleep time; TWT, total wake time.

sleep into either light, deep, or REM sleep. Collectively, measures of the time spent in light sleep reported the lowest margin of error of the three sleep stages with an overall MAPE of 24.45% across all trials and devices. As observed in Figure 7, there was relatively low spread and minimal differences between the four devices. For light sleep, device MAPE values ranged from 13.48% (WHOOP) to 40.85% (Garmin; see Table 7 for additional APE statistics). The aforementioned MAPE values were well depicted in conjunction to the bias demonstrated in the respective Bland–Altman plots for light sleep for each device (see supplemental materials Figure S2A–D). Measurements of light sleep obtained from WHOOP were much less biased and illustrated strong consistency between trials. In contrast, the other devices demonstrated greater inaccuracy and variability between trials. Fitbit (−0.38 hours), Oura (−0.49 hours), and WHOOP (−0.17 hours) slightly underestimated the time spent in light sleep, whereas Garmin reported an overestimation in light sleep (0.56 hours), as seen in Table 8. However, none of the devices demonstrated enough statistical evidence of biased measurements, as all confidence intervals for bias contained zero suggesting the sample of measurements are nonbiased with respect to the Sleep Profiler.

Ability to estimate deep sleep was remarkably poor for all devices (see Table 7 and Figure 8). Collectively, all devices reported a MAPE of 67.96% across trials, whereas MAPE for each device ranged from 46.19% (WHOOP), to 133.54% (Garmin), which may be observed in Table 7. Overall, APE was substantially higher for deep sleep estimations as compared to light sleep. As depicted in the Bland–Altman plots for each device (see supplemental materials Figure S3A–D), WHOOP, Fitbit, and Oura illustrated fairly similar degrees of variability between trials, whereas Garmin demonstrated a much larger spread with respect to its LOA. All devices exhibited a bias that indicated a tendency to overestimate deep sleep to some degree; however, Oura was the only device to reach significance ($p = 0.01$) for its propensity to overestimate (0.33 hours) the deep sleep stage compared to the Sleep Profiler (see Table 8). Although Garmin reported

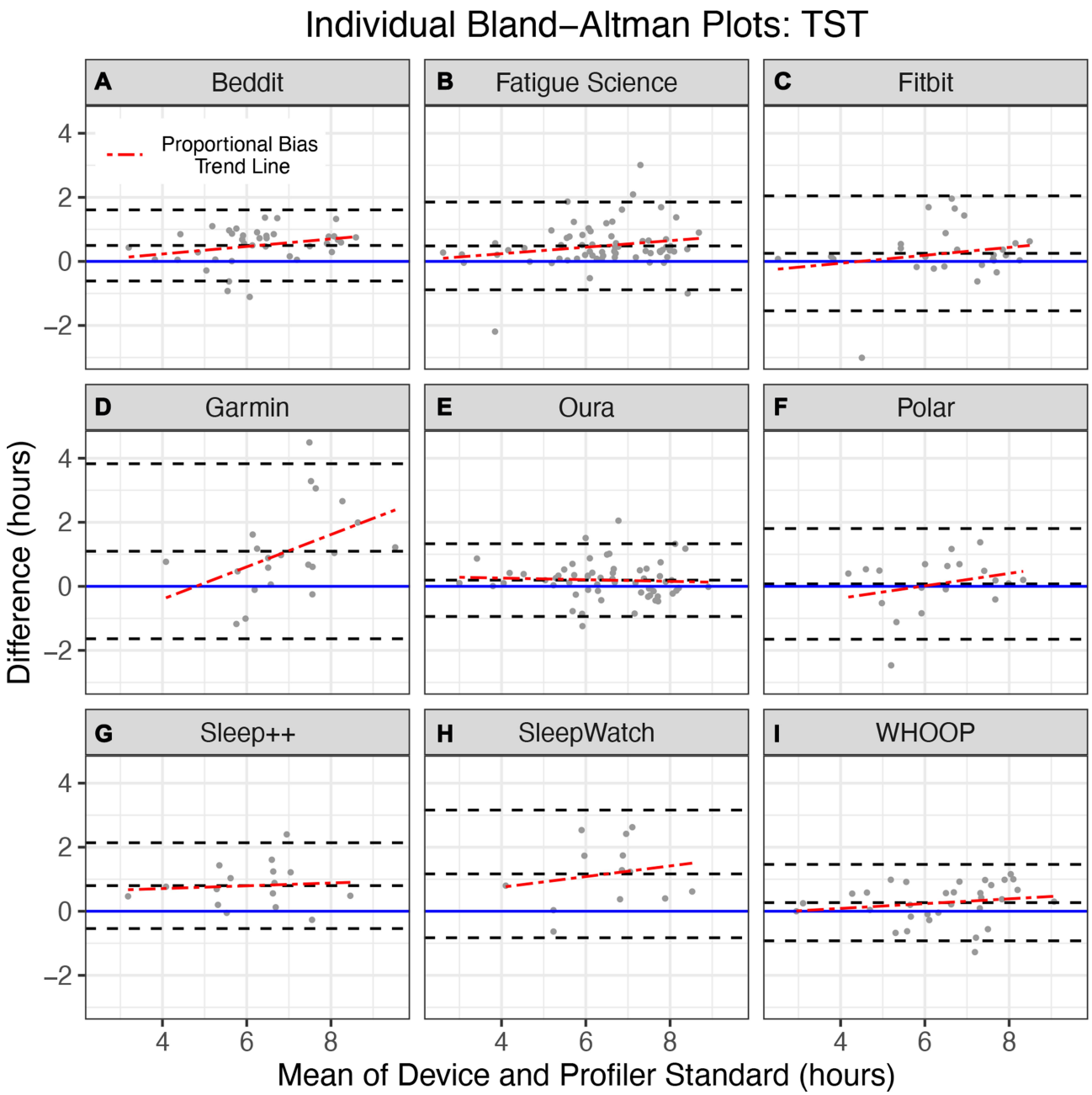


**Figure 2** (**A**–**I**) TST Bland–Altman plots for all devices.
**Notes:** Figure 2 depicts individual Bland–Altman plots for all commercial sleep technologies that measured TST. (**A**) Beddit; (**B**) Fatigue Science; (**C**) Fitbit; (**D**) Garmin; (**E**) Oura; (**F**) Polar; (**G**) Sleep++; (**H**) SleepWatch; (**I**) WHOOP.
**Abbreviation:** TST, total sleep time.

the largest degree of bias, the bias was insignificant, which is attributable to the lower number of trials assessed as well as greater variability as evidenced by the remarkably large LOA range (6.1 hours), even after extreme outlier removal. Garmin's LOA range was nearly triple that of the other three devices reporting deep sleep. The large LOA range exhibited by Garmin for deep sleep contributed to a larger standard error around the bias estimate thus eliciting a wider confidence interval.

The third sleep stage that was estimated by the commercial devices, that is the amount of time spent in REM sleep, demonstrated greater degrees of inaccuracy and variability between trials relative to the other two stages reported by the commercial devices (deep and light sleep). All devices reported multiple measures that skewed the spread, even after removal of the extreme outliers (Figure 9). Commercial devices ranged from a MAPE of 57.83% (WHOOP) to 79.73% (Oura; see Table 7) and the four devices collectively

**Table 4** Bland–Altman Executive Sleep Summary Statistics: All Devices

| Metric | Device | n | Bias (95% CI) | Adjusted p-value | Lower LOA (95% CI) | Upper LOA (95% CI) | LOA Range |
|---|---|---|---|---|---|---|---|
| TST | Beddit | 38 | 0.50 (0.31–0.68) | <0.001* | −0.61 (−0.92, −0.30) | 1.61 (1.30, 1.92) | 2.22 |
| | Oura | 59 | 0.19 (0.04, 0.35) | 0.444 | −0.94 (−1.19, −0.69) | 1.33 (1.07, 1.58) | 2.27 |
| | WHOOP | 32 | 0.27 (0.05, 0.49) | 0.609 | −0.92 (−1.29, −0.56) | 1.46 (1.10, 1.83) | 2.39 |
| | Sleep++ | 16 | 0.80 (0.44, 1.16) | 0.01* | −0.54 (−1.12, 0.04) | 2.14 (1.55, 2.72) | 2.68 |
| | Fatigue Science | 65 | 0.48 (0.31, 0.66) | <0.001* | −0.89 (−1.18, −0.60) | 1.85 (1.56, 2.14) | 2.74 |
| | Polar | 19 | 0.08 (−0.35, 0.50) | 1 | −1.65 (−2.34, −0.96) | 1.80 (1.11, 2.49) | 3.46 |
| | Fitbit | 27 | 0.25 (−0.11, 0.61) | 1 | −1.54 (−2.14, −0.94) | 2.05 (1.45, 2.64) | 3.59 |
| | SleepWatch | 13 | 1.17 (0.55, 1.78) | 0.048* | −0.83 (−1.80, 0.14) | 3.16 (2.19, 4.13) | 3.99 |
| | Garmin | 21 | 1.10 (0.46, 1.73) | 0.062 | −1.64 (−2.67, −0.60) | 3.83 (2.79, 4.86) | 5.46 |
| TWT | Fitbit | 27 | −0.17 (−0.28, −0.06) | 0.109 | −0.71 (−0.89, −0.53) | 0.37 (0.19, 0.55) | 1.08 |
| | Polar | 19 | −0.55 (−0.74, −0.37) | <0.001* | −1.31 (−1.62, −1.01) | 0.21 (−0.10, 0.51) | 1.52 |
| | Garmin | 21 | −0.86 (−1.07, −0.65) | <0.001* | −1.78 (−2.12, −1.43) | 0.06 (−0.29, 0.40) | 1.83 |
| | WHOOP | 33 | −0.59 (−0.76, −0.42) | <0.001* | −1.53 (−1.82, −1.25) | 0.35 (0.07, 0.64) | 1.89 |
| | Beddit | 40 | −0.58 (−0.73, −0.42) | <0.001* | −1.53 (−1.78, −1.27) | 0.37 (0.11, 0.63) | 1.89 |
| | Oura | 60 | −0.18 (−0.31, −0.04) | 0.399 | −1.21 (−1.44, −0.98) | 0.85 (−0.63, 1.08) | 2.07 |
| | Fatigue Science | 70 | −0.29 (−0.41, −0.16) | <0.001* | −1.32 (−1.53, −1.11) | 0.75 (0.54, 0.96) | 2.07 |
| SE | Garmin | 20 | 10.6 (9.26, 11.94) | <0.001* | 4.99 (2.81, 7.17) | 16.21 (14.03, 18.39) | 11.21 |
| | Fitbit | 28 | 0.65 (−0.82, 2.12) | 1 | −6.79 (−9.22, −4.36) | 8.09 (5.66, 10.52) | 14.88 |
| | Fatigue Science | 69 | 4.24 (2.94, 5.54) | <0.001* | −6.38 (−8.57, −4.18) | 14.86 (12.66, 17.05) | 21.23 |
| | Polar | 19 | 7.53 (4.88, 10.18) | <0.001* | −3.24 (−7.54, 1.06) | 18.30 (14.00, 22.59) | 21.54 |
| | Oura | 60 | 1.66 (0.17, 3.15) | 1 | −9.61 (−12.11, −7.11) | 12.93 (10.43, 15.43) | 22.54 |
| | WHOOP | 32 | 7.38 (4.99, 9.76) | <0.001* | −5.59 (−9.55, −1.63) | 20.35 (16.39, 24.31) | 25.94 |
| | Beddit | 40 | 7.20 (4.91, 9.48) | <0.001* | −6.80 (−10.61, −2.99) | 21.19 (17.38, 25.00) | 27.99 |

**Notes:** Table 4 includes Bland–Altman summary statistics, which assesses the degree of bias between reports of the Sleep Profiler and a given device, for all devices across the three sleep summary metrics. Significant results are indicated with a "*", and in these cases, there is enough evidence to conclude that the device is not "unbiased" thus demonstrating a tendency to overestimate or underestimate a given metric. Data presented for each individual metric are sorted lowest-to-highest based upon the range between upper and lower LOA.

**Abbreviations:** CI, confidence interval; LOA, limits of agreement; n, number of trials; SE, sleep efficiency; TST, total sleep time; TWT, total wake time.

reported a MAPE of 72.09% across all trials, suggesting remarkably poor accuracy at measuring REM sleep. The Bland–Altman analysis for each device (see supplemental materials Figure S4A–D) further exposed device inaccuracies with respect to REM sleep (Table 8). The bias indicated for each device that they all overestimated REM sleep time, despite none of them reaching significance (see Table 8), a likely result of the aforementioned degree of high variability observed in each device. To further examine proportional biases discovered for each of the devices with respect to light, deep, and REM sleep, please also refer the supplemental materials (Table S7).

## Discussion

Growing public intrigue related to self-monitoring trends in sleep parameters stimulated a large number of companies to develop in-home sleep technologies with limited oversight for device accuracy.[3–5] Provided the scarcity of third-party validations for the majority of these devices, the present study examined the validity of numerous commercial devices for quantifying sleep metrics by executing direct comparisons to a home-based FDA 510(k) approved EEG-based device (Sleep Profiler).[28] The main finding was that among all of the metrics that are reported by commercial sleep technologies, there is a large variation in accuracy across not only devices but across the metrics themselves as compared directly to the Sleep Profiler. For non-staging sleep metrics (TST, TWT, and SE), Fitbit and Oura reported lower error values overall (eg, MAPE ≤ 10%) and reported equal summated rankings (5) via Kendall's W that considers all three metrics, placing both devices atop the rankings. However, when assessing accuracy of sleep staging, all the commercial devices tested struggled with accuracy. With the exception of one device on light sleep duration (WHOOP), all devices showed >20% MAPE for all sleep staging durations, with over half of the measures exceeding 40% MAPE. High inaccuracy in sleep

**Table 5** Bland–Altman Proportional Bias Summary Statistics: All Devices

| Metric | Device | $R^2$ | Adjusted p-value |
|---|---|---|---|
| TST | Beddit | 0.073 | 1 |
| | Fatigue Science | 0.042 | 1 |
| | Fitbit | 0.041 | 1 |
| | Garmin | 0.189 | 1 |
| | Oura | 0.003 | 1 |
| | Polar | 0.073 | 1 |
| | Sleep++ | 0.007 | 1 |
| | SleepWatch | 0.038 | 1 |
| | WHOOP | 0.032 | 1 |
| TWT | Beddit | 0.083 | 1 |
| | Fatigue Science | 0.001 | 1 |
| | Fitbit | 0.188 | 0.828 |
| | Garmin | 0.753 | < 0.001* |
| | Oura | 0.003 | 1 |
| | Polar | 0.256 | 0.951 |
| | WHOOP | 0.188 | 0.410 |
| SE | Beddit | 0.006 | 1 |
| | Fatigue Science | 0.046 | 1 |
| | Fitbit | 0.012 | 1 |
| | Garmin | 0.266 | 0.42 |
| | Oura | 0.115 | 0.281 |
| | Polar | 0.303 | 0.513 |
| | WHOOP | 0.036 | 1 |

**Notes:** Table 5 includes Bland–Altman summary statistics for proportional biases, which assesses the degree of proportionality with respect to the bias between reports of the Sleep Profiler and a given device, for all devices across the three sleep summary metrics. Significant results are indicated with a "*", and in these cases, there is enough evidence to conclude that the device is not "proportionally unbiased" thus demonstrating a tendency to overestimate or underestimate at higher or lower magnitudes for a given metric. Data presented for each individual metric are sorted alphabetically by device for each of the three sleep metrics displayed.

**Abbreviations:** SE, sleep efficiency; TST, total sleep time; TWT, total wake time.

staging in these technologies is not surprising, as measuring brain electrical signals (via EEG) is the most accurate way to assess sleep stages. None of the commercial devices tested incorporated EEG-based measures, and all rely on a sensor or combination of sensors including accelerometers and PPG optical sensors to estimate sleep stages. Based on the data presented, non-sleep staging data (sleep/wake summary metrics only) should be considered while using commercial non-EEG-based sleep devices, and even then, great care and consideration should be taken when selecting a commercial device for these metrics.

Sleep monitoring that requires high-resolution sleep data (eg, sleep staging) is traditionally measured with the PSG technique.[8,9] PSG, however, requires specialized equipment and skilled technicians to conduct the overnight procedure usually in a research laboratory or hospital clinic setting. Further, the novel environment combined with the need for multiple leads placed on the head, face and body can often result in disrupted sleep. Consequently, for many instances, PSG is not recommended due to the potential to disrupt sleep (eg, insomnia profiles) and/or impracticality in conducting longitudinal studies. While we recognize it is not the conventional criterion for assessing accuracy of sleep monitoring, the use of the Sleep Profiler rather than PSG in the current study provided multiple benefits. Previous research on the collection and automated algorithmic scoring of the Sleep Profiler data has shown to have comparable degrees of agreement between autostaging and manual scoring of PSG, and further, comparable accuracy margins as compared to agreement variation between PSG technicians when scoring the same assessment.[28] Assessments of technologies using the Sleep Profiler allowed for data collection in an innate environment without significantly sacrificing veracity, ultimately aiding in the assessment of the commercial devices in their intended environment.

## Device Performance

Results derived from the aforementioned analyses agree with previous research evaluating device performance of commercial sleep technologies that report on similar metrics. This notion holds especially true for objectively measuring sleep stages as it is apparent (supported by current and previous literature) that sleep staging is not accurately measured by the commercial sleep technologies assessed. Of greater concern, there is a consistent pattern in the extant literature in which the commercial sleep technology market reveals a large degree of volatility with respect to accurately quantifying sleep. For example, similarly significant overestimations of TST and underestimations of TWT by the Garmin Vivosmart were observed herein and previously, even though prior studies utilized older models of the Vivosmart device.[37] Additionally, negligible biases for TST and SE were observed for the Fitbit Ionic wrist device in the present study whereas previous models of Fitbit devices overestimated both metrics.[55] Varying findings stem from remarkably few opportunities that exist for direct comparisons, which is due to any combination of upgrades in device models, differing statistical analysis strategies, or a pure lack of previous, independent, third-party investigation. Moreover, industry-accepted standards for sufficient levels of device accuracy are not established thus interpreting

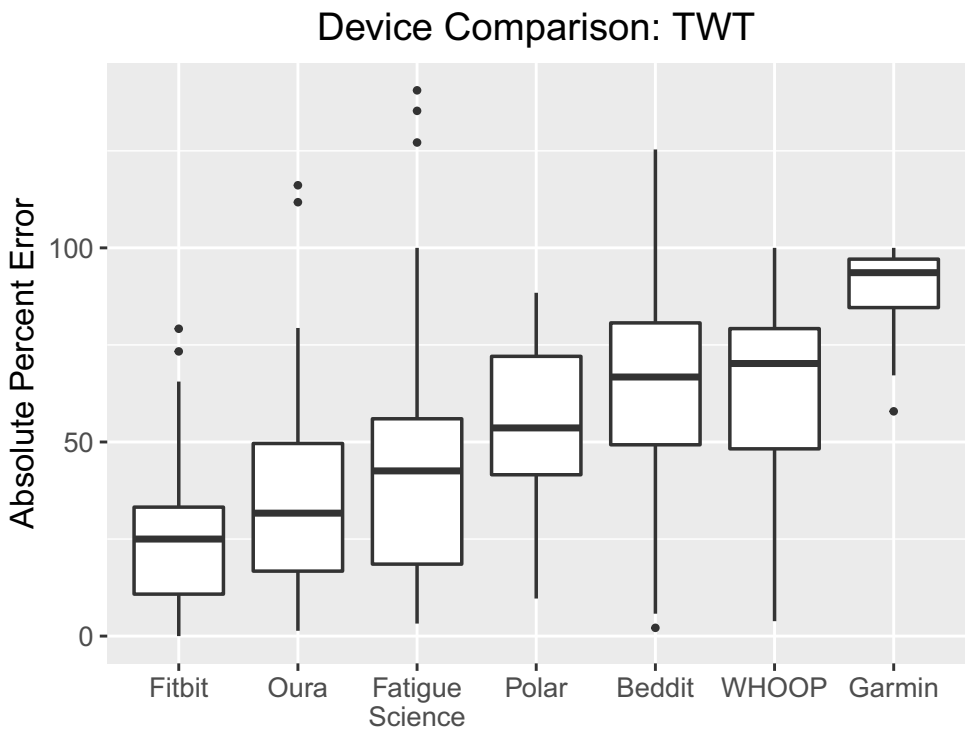


**Figure 3** TWT time boxplots: absolute percent error by device.
**Note:** Figure 3 depicts a box plot representation of absolute percent error (APE) for each commercial device that reported on total wake time.
**Abbreviation:** TWT, total wake time.

whether or not a device is "accurate enough" is impossible until there is a greater volume of research on these various devices as they are continually released to the public. Future research should also consider surveying various end-users (eg, clinicians, practitioners, general consumers) to determine thresholds by which individuals deem a sleep device as being accurate or inaccurate.

## Device Considerations

Although all of the devices assessed conclusively elicited substantial degrees of inaccuracy in staging sleep metrics, it is important to note multiple perspectives. First, the commercial sleep devices are targeted for non-clinical purposes in the general population. If a clinician is investigating a potential sleep disorder in a patient, they can use a combination of EEG-based take home devices and PSG protocols. For the general public, which may include any combination of students, sedentary workers, military personnel, first responders, or athletes, to name a few, it is important to understand the intent of utilizing the sleep device. This is obviously personalized to the individual but one simple perspective is that understanding current sleep habits, from the average time you go to bed and wake up, to the average amount of time you are awake in bed, can be enlightening and potentially utilized as an augmentation tool for bettering one's health. Ideally, this information will lead to improved sleep behaviors or lifestyle changes and, in which case, it is even more important to measure the impact of these changes on subsequent sleep patterns. Simple sleep hygiene alterations like going to bed earlier and more consistently to increase TST can accurately be assessed using some of the technologies tested and reported here. Furthermore, some of the technologies assessed herein permit the understanding of links between nutrient timing (eg, eating habits) to measures like sleep efficiency and heart rate during sleep.[56] This is where consumer education is important, since all technologies that report sleep staging do so in consumer facing smartphone applications where it is easy to fixate on a sleep staging metric despite this value having a high likelihood of being inaccurate.

The task at hand requires assessment of different states of sleep through miniscule physiological changes. This effort, specifically for sleep staging, is a challenge even for experts in the field that utilize data from the most accurate methods of data collection. In fact, repeated scoring of the same data by different PSG technicians can result in agreement differences upwards of 20%.[57]

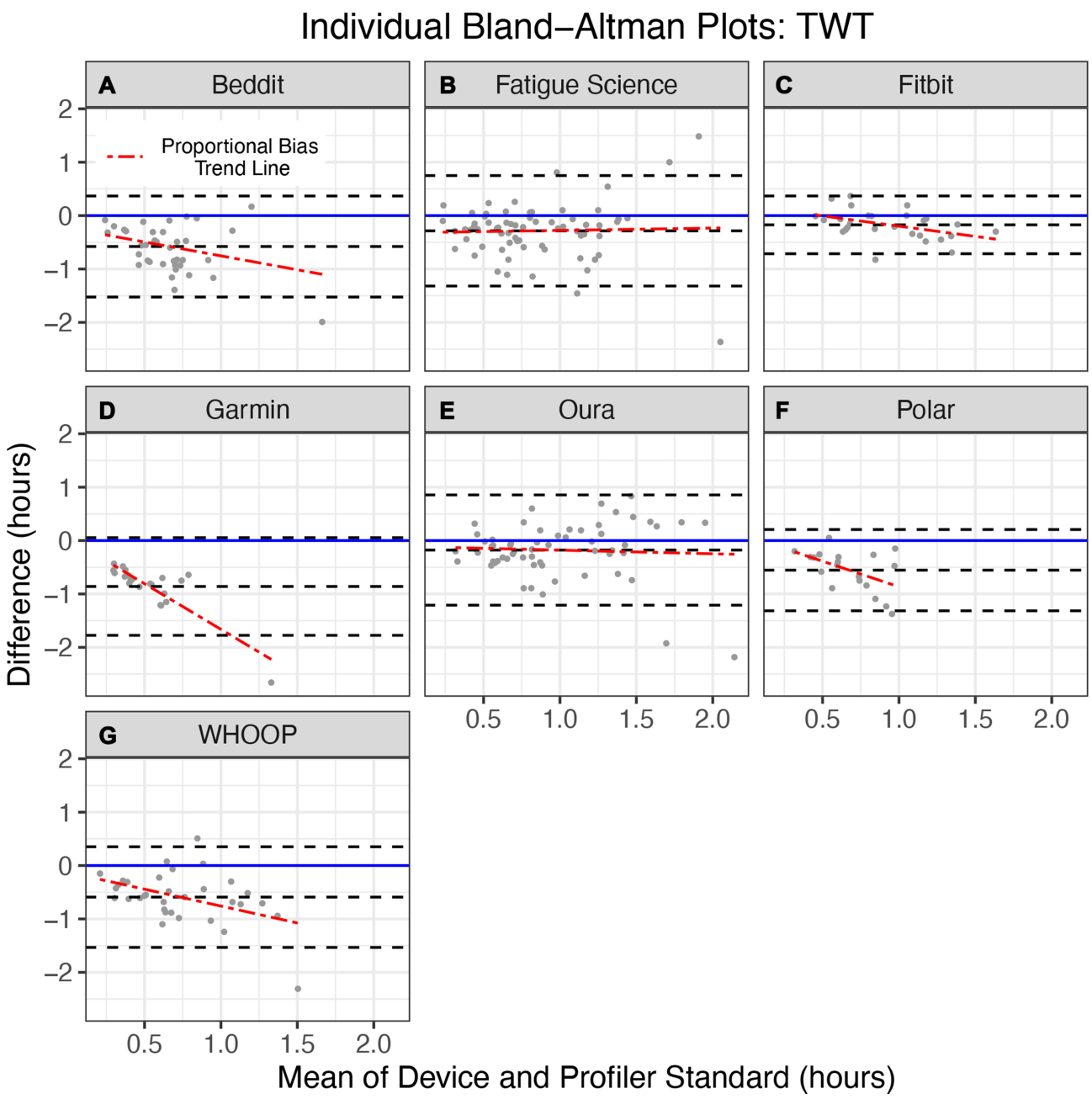


**Figure 4** (**A**–**G**) TWT Bland–Altman plots for all devices.
**Notes:** Figure 4 depicts individual Bland–Altman plots for all commercial sleep technologies that measured TWT. (**A**) Beddit; (**B**) Fatigue Science; (**C**) Fitbit; (**D**) Garmin; (**E**) Oura; (**F**) Polar; (**G**) WHOOP.
**Abbreviation:** TWT, total wake time.

Furthermore, in most sleep monitoring technologies, this daunting challenge of collecting precise measurements is appointed to a device that only costs $80–$300, which is a fraction of the costs associated with the breadth of equipment and expertise needed for PSG. In order to maintain affordability of the various devices, the methods of data collection and algorithms utilized by the various technologies examined herein vary (see Table 1). While computational algorithms are generally withheld by manufacturers, methods of measurement employed by a device can have a significant impact on the accuracy of the device, and thus inferences drawn from device measurements are questioned. Therefore, due to the diversity of inaccurate reports that these devices provide, we recommend those who plan to employ a commercial sleep technology for general, clinical, or research purposes

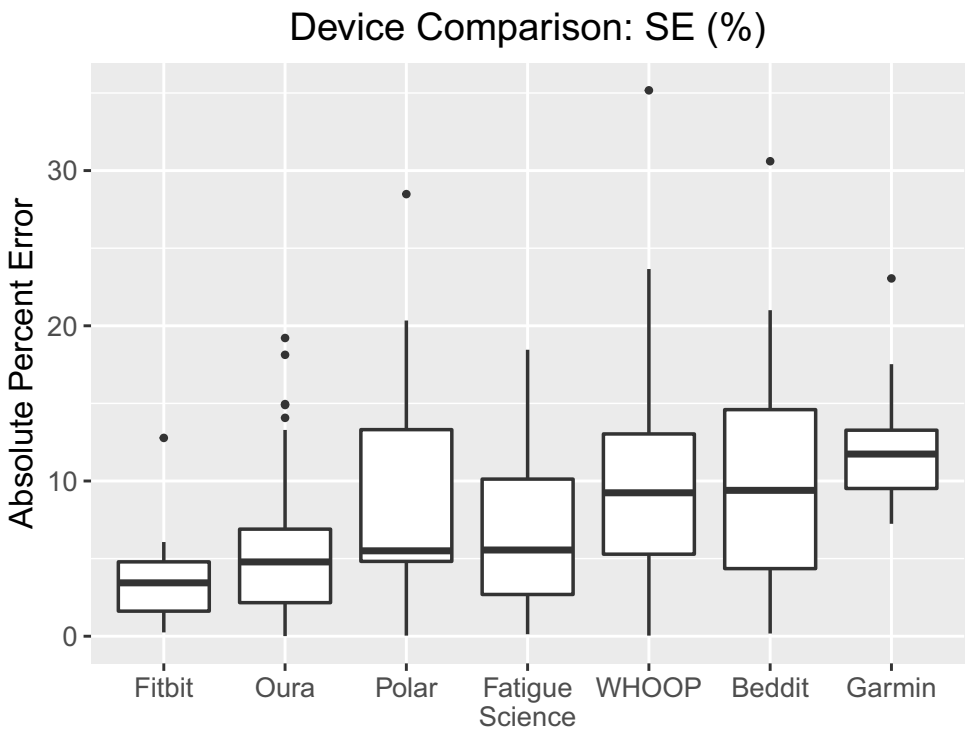


**Figure 5** SE boxplots: absolute percent error by device.
**Note:** Figure 5 depicts a box plot representation of absolute percent error (APE) for each commercial device that reported on sleep efficiency.
**Abbreviation:** SE, sleep efficiency.

thoroughly evaluate their primary objectives for monitoring prior to purchasing.

As movement patterns vary with different states of wakefulness, accelerometers are often utilized by sleep monitoring technologies. Of the devices assessed in this study, the Apple Watch Series 3 (measured via Sleep++ and SleepWatch apps), Fatigue Science Readiband, Fitbit Ionic, Garmin Vivosmart 4, Oura Smart Ring, Polar A370, and WHOOP Strap 2.0 were all equipped with accelerometers (see Table 1). Indeed, notable changes in movement occur during the transition between sleep and wake[58] and the technique of estimating sleep and wake states from accelerometer-based movement data (termed actigraphy) has been used in sleep research for several decades.[59] This is likely a major reason why commercial devices include an accelerometer and report on such metrics as TST, TWT, and SE with relatively low levels of inaccuracy and bias. When considering the devices examined in the present study that incorporated accelerometry into sleep derivations, only two devices (of the eight total), the Fatigue Science Readiband and Sleep++ smartphone app via the Apple Watch, reported a significant bias of TST.

Without insight on brain activity, accurately staging sleep requires the assessment of far more diminutive physiological signals than those required to simply determine sleep and wakefulness.[58] Many devices approach this challenge through use of PPG signals to measure changes in heart rate, heart rate variability, and respiration rates,[4,23,58,60] which additionally enhances wakefulness determinations provided by accelerometers. Further, devices' ability to measure mean heart may contribute to its ability to estimate sleep stages. Although the current study was unable to directly assess the accuracy of heart rate measurements in the technologies assessed due to a lack worthy comparison (eg, multi-lead electrocardiography), previous research suggests PPG devices that utilize a red or infrared wavelength most accurately evaluate physiological signals in the presence of little to no movement, as observed during sleep.[61] This is consistent with trends in our findings; the Fitbit Ionic and Oura Smart Ring, both of which collect data via PPG at higher wavelengths, possessed the lowest magnitudes of error when assessing TST, TWT, and SE. In fact, these two devices had lower error margins than all of the devices that did not utilize a method of PPG, and those that utilized PPG but with a green LED. Similar assertions were made in the staging variables where, for example, the Garmin

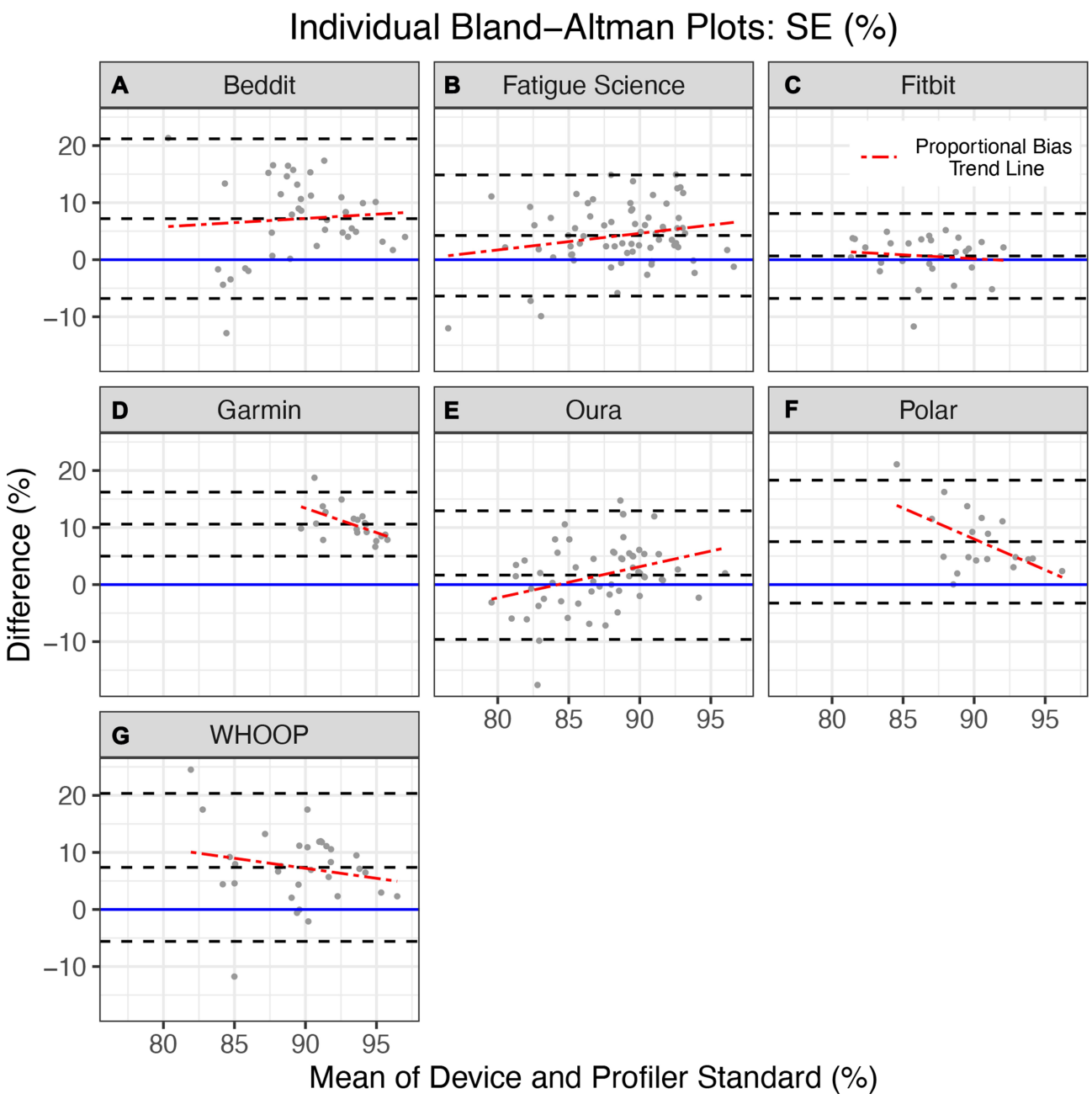


**Figure 6** (**A**–**G**) SE Bland–Altman plots for all devices.
**Notes:** Figure 6 depicts individual Bland–Altman plots for all commercial sleep technologies that measured SE. (**A**) Beddit; (**B**) Fatigue Science; (**C**) Fitbit; (**D**) Garmin; (**E**) Oura; (**F**) Polar; (**G**) WHOOP.
**Abbreviation:** SE, sleep efficiency.

Vivosmart 4, a device that wields PPG via green LED, provided remarkably higher margins of error in light and deep sleep time as compared to devices that wield PPG utilizing red or infrared wavelengths. Still, these are merely inferences drawn from the data presented and future research should consider incorporating multi-lead electrocardiography as an additional device to be worn at night so that direct comparisons of heart rate measurements during routine sleeping is possible.

Another aspect worthy of consideration in the present study surrounds the sleep data from the Apple Watch, which poses substantial limitations and the potential for a great deal of variability. Apple does not advertise any of their watches as having sleep monitoring capabilities thus it is left up to third-party app developers to address this gap.[62] With dozens of different applications now claiming to monitor sleep via Apple Watch, attempts by the developers are made to disseminate information to consumers (or researchers), but limited insight

**Table 6** Kendall's Coefficient of Concordance (W): TST, TWT, and SE

| Device | Sum Rank |
|---|---|
| Fitbit | 5 |
| Oura | 5 |
| Fatigue Science | 10 |
| WHOOP | 12 |
| Polar | 13 |
| Beddit | 18 |
| Garmin | 21 |
| Kendall's W | 0.87 |

**Notes:** Table 6 includes a summary of Kendall's Coefficient of Concordance (W) based upon ranked MAPE values from each device for TST, TWT, and SE. Device rankings from the three sleep metrics were summated (Sum Rank).
**Abbreviations:** MAPE, mean absolute percent error; SE, sleep efficiency; TST, total sleep time; TWT, total wake time; W, Kendall's coefficient of concordance.

exists regarding the methods each of these apps employ for collection and interpretation of the collected data to determine sleep metrics.[5,6] Furthermore, sufficient evaluations for any of these third-party applications are incomplete (lacking peer review) or simply do not exist. While the Apple Watch is equipped with the technologies necessary to provide competitive results with competing devices, the algorithms implemented by third parties may not utilize device capabilities to their fullest extent; this likely contributes, at large, to the consistently poor assessments of sleep by Sleep++ and SleepWatch observed in the present study. To expand on an issue similar to this, the black box effect is rampant in commercial wearable technologies.[17,63] More specifically, the black box effect refers to the notion that there is a severe lack of transparency originating from companies as it relates to computational strategies for health monitoring. As such, the variability in algorithms incorporated into objective sleep monitoring devices and the confidentiality that most companies possess for this information drastically limits understanding for how these conclusions are drawn.[62]

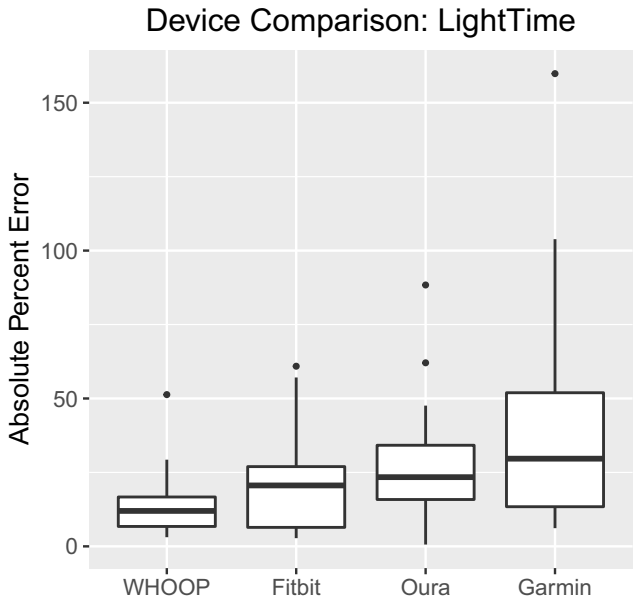


**Figure 7** Light time boxplots: absolute percent error by device.
**Note:** Figure 7 depicts a box plot representation of absolute percent error (APE) for each commercial device that reported on light sleep time.

## Limitations and Future Directions

A few limitations were apparent in the present study. The complexion of participants' skin tones were not accounted for, a variable of which has been shown to affect measures obtained via PPG.[23,24] Participants all had light skin tones, thus the sample excluded relatively darker skins tones, which are often associated with high error rates from PPG technology that utilize a green LED.[23,25,26] However, the design of the current study should work in the favor of PPG accuracy as the data presented could be representative of the best case scenario for PPG performance. It is also expected that these sleep technologies utilize different scoring rules (eg, filtering techniques, data signal trimming) that likely differ across technologies. Further research should be performed to understand the links between scoring approaches and accuracy, which would likely necessitate commercial companies being willing to share raw data signals. Additionally, because this study was an evaluation of the devices as they are available to the consumer at the time of the study, we were unable to account for any cloud-based firmware updates that may have been pushed to a device, during a single subject's data collection.[6,17] New device models are released at different times for each company as well, so to stay exactly current on the most up-to-date models is difficult. The most recent devices commercially available were procured at the time of the study. This issue lends credence to needing validation studies to be continuously executed as new technologies emerge, and modifications of existing technologies are released. Furthermore, direct observations of sleep nights by the researchers were not possible since data collection occurred as part of participants' normal sleeping routines. Consequently, only the aforementioned assurances of data quality, such as an initial familiarization session, routine data exports and reviews executed by the researchers, as well as extreme outlier detection strategies, were capable of practical implementation. However, this design lends itself to

**Table 7** Absolute Percent Error Staging Summary Statistics: All Devices

| Metric | Device | n | MAPE | Min. (%) | MdAPE | Max. (%) | IQR |
|---|---|---|---|---|---|---|---|
| Light Time | WHOOP | 30 | 13.48 | 3.09 | 11.98 | 51.30 | 9.96 |
| | Fitbit | 26 | 20.41 | 2.77 | 20.59 | 60.91 | 20.54 |
| | Oura | 62 | 25.63 | 0.60 | 23.40 | 88.37 | 18.34 |
| | Garmin | 22 | 40.85 | 6.15 | 29.65 | 159.84 | 38.54 |
| Deep Time | WHOOP | 32 | 46.19 | 1.64 | 32.68 | 142.86 | 49.08 |
| | Fitbit | 26 | 58.82 | 9.86 | 42.02 | 137.84 | 61.39 |
| | Oura | 59 | 60.46 | 0.00 | 55.81 | 184.34 | 58.41 |
| | Garmin | 21 | 133.54 | 8.73 | 62.42 | 561.29 | 135.33 |
| REM Time | Garmin | 19 | 62.06 | 8.02 | 46.58 | 184.51 | 80.54 |
| | WHOOP | 30 | 57.83 | 3.30 | 53.72 | 230.77 | 47.40 |
| | Fitbit | 27 | 79.17 | 6.91 | 30.20 | 311.54 | 98.80 |
| | Oura | 56 | 79.73 | 3.73 | 61.08 | 303.39 | 74.60 |

**Notes:** Table 7 includes APE summary statistics for all devices that reported on each respective measure of sleep staging. Data presented for each individual staging metric are sorted lowest-to-highest based upon MAPE.
**Abbreviations:** APE, absolute percent error; IQR, interquartile range; MAPE, mean absolute percent error; Max, maximum; MdAPE, median absolute percent error; Min, minimum; n, number of trials; REM, rapid eye movement.

**Table 8** Bland–Altman Sleep Staging Statistics: All Devices

| Metric | Device | n | Bias (95% CI) | Adjusted p-value | Lower LOA (95% CI) | Upper LOA (95% CI) | LOA Range |
|---|---|---|---|---|---|---|---|
| Light Time | WHOOP | 30 | −0.17 (−0.43, 0.08) | 1 | −1.52 (−1.95, −1.10) | 1.17 (0.75, 1.60) | 2.70 |
| | Fitbit | 26 | −0.38 (−0.82, 0.06) | 1 | −2.51 (−3.23, −1.79) | 1.75 (1.03, 2.47) | 4.26 |
| | Oura | 62 | −0.49 (−0.78, −0.20) | 0.05 | −2.74 (−3.23, −2.24) | 1.76 (1.27, 2.25) | 4.50 |
| | Garmin | 22 | 0.33 (−0.38, 1.04) | 1 | −2.81 (−3.97, −1.65) | 3.47 (2.31, 4.64) | 6.28 |
| Deep Time | Fitbit | 26 | 0.15 (−0.09, 0.39) | 1 | −1.01 (−1.40, −0.61) | 1.30 (0.91, 1.70) | 2.31 |
| | WHOOP | 32 | 0.03 (−0.19, 0.25) | 1 | −1.16 (−1.52, −0.79) | 1.22 (0.86, 1.58) | 2.38 |
| | Oura | 59 | 0.33 (0.16, 0.49) | 0.006* | −0.91 (−1.19, −0.63) | 1.57 (1.29, 1.84) | 2.48 |
| | Garmin | 21 | 0.56 (−0.15, 1.26) | 1 | −2.49 (−3.65, −1.34) | 3.61 (2.45, 4.76) | 6.10 |
| REM Time | WHOOP | 30 | 0.10 (−0.15, 0.35) | 1 | −1.22 (−1.63, 0.80) | 1.42 (1.00, 1.83) | 2.63 |
| | Garmin | 19 | −0.05 (−0.49, 0.39) | 1 | −1.83 (−2.54, −1.12) | 1.73 (1.02, 2.44) | 3.56 |
| | Fitbit | 27 | 0.31 (0.00, 0.62) | 1 | −1.22 (−1.73, −0.71) | 1.83 (1.32, 2.34) | 3.05 |
| | Oura | 56 | 0.19 (−0.05, 0.42) | 1 | −1.52 (−1.92, −1.13) | 1.90 (1.51, 2.29) | 3.42 |

**Notes:** Table 8 includes Bland–Altman summary statistics, which assesses the degree of bias between reports of the Sleep Profiler and a given device, for all devices across the three sleep staging metrics. Significant results are indicated with a "*", and in these cases, there is enough evidence to conclude that the device is not "unbiased" thus demonstrating a tendency to overestimate or underestimate a given metric. Data presented for each individual staging metric are sorted lowest-to-highest based upon the range between upper and lower LOA.
**Abbreviations:** CI, confidence interval; LOA, limits of agreement; n, number of trials; REM, rapid eye movement.

a true commercial evaluation of technologies, where the consumer follows device directions and implements device utilization as a part of their daily routine.

As technologies continue to advance, it will be vital for concerning parties to continue evaluating the accuracy and reliability of new devices and updates to algorithms as they are released. Future research should address the dynamic nature of the field and continue encouragement towards manufacturers to devote a greater degree of prioritization towards third-party validations of their devices. Additionally, validations on the various sensors of the devices (eg, accelerometers, PPG), of which raw data are collected and used to estimate various sleep metrics, should be conducted. This may provide valuable insight necessary for further validation of not only the device itself but also the algorithms used for interpretation of collected data. Once accuracy and reliability have been confirmed in a general population, modifications to algorithms can be explored for application to various clinical and diseased populations. Moreover, a limitation to not only the present study but

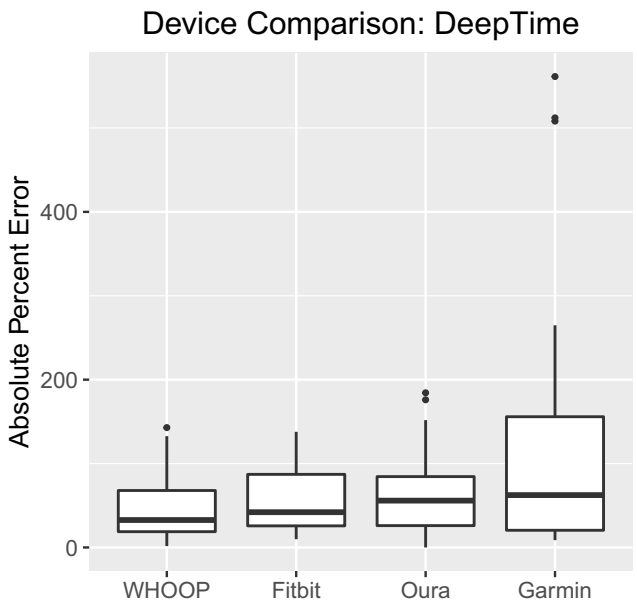


**Figure 8** Deep time boxplots: absolute percent error by device.
**Note:** Figure 8 depicts a box plot representation of absolute percent error (APE) for each commercial device that reported on deep sleep time.

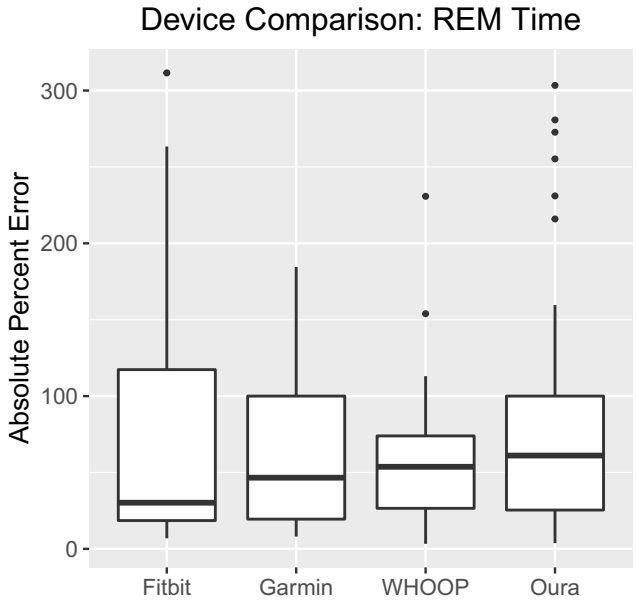


**Figure 9** REM time boxplots: absolute percent error by device.
**Note:** Figure 9 depicts a box plot representation of absolute percent error (APE) for each commercial device that reported on REM time.
**Abbreviation:** REM, rapid eye movement.

seemingly every other one that seeks to examine commercial device accuracies is that they are unable to determine what is "accurate enough." Provided that the intended purpose for these technologies will likely vary to a large degree based upon the end-user(s), future research should survey various application domains so that companies, researchers, and consumers can all have a concrete answer to what is "accurate enough."

## Conclusion

The ability to accurately monitor longitudinal trends in sleep through commercial sleep technologies presents a valuable opportunity to end-users who are interested in tracking this information. Considering the unequivocal importance of adequate sleep, a commercial device's ability to accurately and objectively measure sleep characteristics (eg, sleep, wake, and stage durations) affords invaluable insights to consumers, researchers, and clinicians alike. However, findings from the present study, which align with extant literature, reveal large variations between commercial devices; the greatest amounts of confidence in device accuracy were associated with measurements of TST, TWT, and SE whereas little confidence exists in attempts to measure sleep staging metrics. Although accurate sleep staging derived from commercial technologies would generate a whole new realm of application in sleep research, it does not appear that any of the devices tested are capable of such reporting. Thus, we do not recommend that these sleep staging metrics are incorporated into any decision-making as it relates to routine monitoring (eg, consumer-based self-monitoring, clinical sleep monitoring) at this time. Yet, there are possibilities for reasonably accurate monitoring of TST, TWT, and SE. The Fitbit Ionic wrist device and the Oura Smart Ring possessed the lowest degrees of error for these three metrics as they were the only devices that provided unbiased estimates in each case and elicited the lowest APE values. With that said, WHOOP performed similarly to Fitbit and Oura with respect to TST, although error margins for TWT and SE were nearly doubled. Additionally, WHOOP reported the most accurate results among commercial devices analyzed for light and deep sleep durations, albeit APE values being relatively high. For these reasons, it is advisable that those individuals contemplating future sleep monitoring consider the data quality of these three devices compared to the others. If the primary interest of sleep monitoring is to assess sleep durations and the general efficiency of that sleep, there first must be a foundational understanding that more detailed interpretations of sleep physiology would be possible with distinguishing sleep stages, yet no commercial device appears capable of accurately doing so. As such, the incorporation of devices that report TST, TWT, and SE, may be feasible if an end user is aware of or able to determine what their acceptable thresholds for error might be. Unfortunately, which device is "accurate enough" is a question that remains.

As previously mentioned, future investigations must consider this thought and being surveying relevant populations for what magnitudes of accuracy (or lack thereof) justify the utilization of a particular commercial device.

## Abbreviations

APE, absolute percent error; ACSM, American College of Sports Medicine; bpm, beats per minute; CI, confidence interval; EEG, electroencephalography; EMG, electromyography; EOG, electrooculography; eg, exempli gratia; Q1, first quartile; FDA, Food and Drug Administration; Gen, generation; HR, heart rate; cm, height in centimeters; IQR, interquartile range; LED, light emitting diode; LOA, limits of agreement; Max, maximum; MAPE, mean absolute percent error; MdAPE, median absolute percent error; Min, minimum; NTC, negative temperature coefficient; n, number of trials; PPG, photoplethysmography; PSG, polysomnography; REM, rapid eye movement; SE, sleep efficiency; app, smartphone application; SD, standard deviation; Q3, third quartile; TST, total sleep time; kg, weight in kilograms; y, years of age.

## Acknowledgments

The authors would like to express gratitude to the Rockefeller Neuroscience Institute at West Virginia University for support in data collection.

## Funding

This study was funded internally by the West Virginia University, Rockefeller Neuroscience Institute.

## Disclosure

Financial competing interests: none. Non-financial competing interests: none. The authors report no conflicts of interest for this work.

## References

1. Watson NF, Badr MS, Belenky G, et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American academy of sleep medicine and sleep research society. *Sleep*. 2015. doi:10.5665/sleep.4716
2. Itani O, Jike M, Watanabe N, Kaneita Y. Short sleep duration and health outcomes: a systematic review, meta-analysis, and meta-regression. *Sleep Med*. 2017;32:246–256. doi:10.1016/j.sleep.2016.08.006
3. Kelly J, Strecker R, Bianchi M. Recent developments in home sleep monitoring devices. *ISRN Neurol*. 2012;2012. doi:10.5402/2012/768794
4. de Zambotti M, Cellini N, Goldstone A, Colrain I, Baker F. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exerc*. 2019;51:1538–1557. doi:10.1249/MSS.0000000000001947
5. Ibañez V, Silva J, Navarro E, Cauli O. Sleep assessment devices: types, market analysis, and a critical review on accuracy and validation. *Expert Rev Med Devices*. 2019;16(12):1041–1052. doi:10.1080/17434440.2019.1693890
6. Roomkham S, Lovell D, Cheung J, Perrin D. Promises and challenges in the use of consumer-grade devices for sleep monitoring. *IEEE Rev Biomed Eng*. 2018;11:53–67. doi:10.1109/RBME.2018.2811735
7. Khosla S, Deak MC, Gault D, et al. Consumer sleep technology: an American academy of sleep medicine position statement. *J Clin Sleep Med*. 2018;14(05):877–880. doi:10.5664/jcsm.7128
8. Rundo J, Downey R. Polysomnography. In: Levin KH, Chauvel P, editors. *Handbook of Clinical Neurology*. Vol. 160. Elsevier; 2019:381–392.
9. Simons PJ, Overeem S. Polysomnography: recording, analysis and interpretation. *Sleep Disorders Neurol*. 2018;13–29.
10. Ahmadi N, Shapiro GK, Chung SA, Shapiro CM. Clinical diagnosis of sleep apnea based on single night of polysomnography vs. two nights of polysomnography. *Sleep Breath*. 2009;13(3):221–226. doi:10.1007/s11325-008-0234-2
11. Kaplan K, Hirshman J, Hernandez B, et al. When a gold standard isn't so golden: lack of prediction of subjective sleep quality from sleep polysomnography. *Biol Psychol*. 2017;123:37–46. doi:10.1016/j.biopsycho.2016.11.010
12. Kushida C, Chang A, Gadkary C, Guilleminault C, Carrillo O, Dement W. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Med*. 2001;2:389–396. doi:10.1016/S1389-9457(00)00098-8
13. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev*. 2011;15(4):259–267. doi:10.1016/j.smrv.2010.10.001
14. Martin JL, Hakim AD. Wrist actigraphy. *Chest*. 2011;139(6):1514–1527. doi:10.1378/chest.10-1872
15. Marino M, Li Y, Rueschman MN, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*. 2013;36(11):1747–1755. doi:10.5665/sleep.3142
16. 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. doi:10.1093/sleep/26.3.342
17. Depner CM, Cheng PC, Devine JK, et al. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;43(2):1–13. doi:10.1093/sleep/zsz254
18. Sitnick SL, Goodlin-Jones BL, Anders TF. The use of actigraphy to study sleep disorders in preschoolers: some concerns about detection of nighttime awakenings. *Sleep*. 2008;31(3):395–401. doi:10.1093/sleep/31.3.395
19. Meltzer LJ, Montgomery-Downs HE, Insana SP, Walsh CM. Use of actigraphy for assessment in pediatric sleep research. *Sleep Med Rev*. 2012;16(5):463–475. doi:10.1016/j.smrv.2011.10.002
20. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. *Sleep*. 2007;30(10):1362–1369. doi:10.1093/sleep/30.10.1362
21. Boneva RS, Decker MJ, Maloney EM, et al. Higher heart rate and reduced heart rate variability persist during sleep in chronic fatigue syndrome: a population-based study. *Auton Neurosci*. 2007;137(1–-2):94–101. doi:10.1016/j.autneu.2007.08.002
22. Cincin A, Sari I, Oğuz M, et al. Effect of acute sleep deprivation on heart rate recovery in healthy young adults. *Sleep Breath*. 2015;19(2):631–636. doi:10.1007/s11325-014-1066-x
23. Allen J. Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas*. 2007;28(3):R1–R39.
24. Butler MJ, Crowe JA, Hayes-Gill BR, Rodmell PI. Motion limitations of non-contact photoplethysmography due to the optical and topological properties of the skin. *Physiol Meas*. 2016;37:N27–N37. doi:10.1088/0967-3334/37/5/N27

25. Bent B, Goldstein B, Kibbe W, Dunn J. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med*. 2020;3. doi:10.1038/s41746-020-0226-6
26. Sañudo B, Hoyo M, Muñoz-Lopez A, Perry J, Abt G. Pilot study assessing the influence of skin type on the heart rate measurements obtained by photoplethysmography with the apple watch. *J Med Syst*. 2019;43(7):195. doi:10.1007/s10916-019-1325-2
27. Wen D, Zhang X, Liu X, Lei J. Evaluating the consistency of current mainstream wearable devices in health monitoring: a comparison under free-living conditions. *J Med Internet Res*. 2017;19(3):e68. doi:10.2196/jmir.6874
28. Levendowski D, Ferini-Strambi L, Gamaldo C, Cetel M, Rosenberg R, Westbrook P. The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers. *J Clin Sleep Med*. 2017;13(6):791–803. doi:10.5664/jcsm.6618
29. Finan P, Richards J, Gamaldo C, et al. Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers. *J Clin Sleep Med*. 2016;12(11):1443–1451.
30. Riebe D, Ehrman J, Liguori G, Magal M, Medicine A. *ACSM's Guidelines for Exercise Testing and Prescription*. 10th ed. Philadelphia Baltimore New York: Wolters Kluwer; 2018.
31. Kaewkannate K, Kim S. A comparison of wearable fitness devices. *BMC Public Health*. 2016;16:1–16. doi:10.1186/s12889-016-3059-0
32. Nelson BW, Allen NB. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: intraindividual validation study. *JMIR Mhealth Uhealth*. 2019;7(3):e10828. doi:10.2196/10828
33. Nakano N, Sakura T, Ueda K, et al. Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. *Front Sports Act Living*. 2020;2:50. doi:10.3389/fspor.2020.00050
34. USFDA. 510 (k) premarket notification: automatic event detection software for polysomnograph with electroencephalograph. Administration USFaD, ed. K120450. Vol 882.14002012.
35. Schupp M, Hanning CD. Physiology of sleep. *BJA CEPD Rev*. 2003;3(3):69–74. doi:10.1093/bjacepd/mkg069
36. Berryhill S, Morton C, Dean A, et al. Effect of wearables on sleep in healthy individuals: a randomized cross-over trial and validation study. *J Clin Sleep Med*. 2020;16(5):775–783. doi:10.5664/jcsm.8356
37. Tedesco S, Sica M, Ancillao A, Timmons S, Barton J, O'Flynn B. Validity evaluation of the fitbit charge2 and the garmin vivosmart HR + in free-living environments in an older adult cohort. *JMIR Mhealth Uhealth*. 2019;7(6):e13084. doi:10.2196/13084
38. Kendall MG, Smith BB. The problem of m rankings. *Ann Math Stat*. 1939;10(3):275–287. doi:10.1214/aoms/1177732186
39. Gearhart A, Booth DT, Sedivec K, Schauer C. Use of Kendall's coefficient of concordance to assess agreement among observers of very high resolution imagery. *Geocarto Int*. 2013;28(6):517–526. doi:10.1080/10106049.2012.725775
40. Gouhier TC, Guichard F, Gonzalez A. Synchrony and stability of food webs in metacommunities. *Am Nat*. 2010;175(2):E16–E34. doi:10.1086/649579
41. Tuominen J, Peltola K, Saaresranta T, Valli K. Sleep parameter assessment accuracy of a consumer home sleep monitoring ballistocardiograph beddit sleep tracker: a validation study. *J Clin Sleep Med*. 2019;15(3):483–487. doi:10.5664/jcsm.7682
42. Kinnunen HO, Rantanen A, Kenttä TV, Koskimäki H. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. *Physiol Meas*. 2020;41:04NT01. doi:10.1088/1361-6579/ab840a
43. Pesonen A-K, 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. doi:10.5664/jcsm.7050
44. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res*. 2019;21(11):e16273. doi:10.2196/16273
45. de Zambotti M, Rosas L, Colrain IM, Baker FC. The sleep of the ring: comparison of the OURA sleep tracker against polysomnography. *Behav Sleep Med*. 2019;17:124–136. doi:10.1080/15402002.2017.1300587
46. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986;327(8476):307–310. doi:10.1016/S0140-6736(86)90837-8
47. Tukey JW. *Exploratory Data Analysis*. Vol. 2. Reading, Mass dawli; 1977.
48. Dawson R. How significant is a boxplot outlier? *Stat Educ*. 2011;19(2).
49. Hsu J. *Multiple Comparisons: Theory and Methods*. CRC Press; 1996.
50. Team RC. R: a language and environment for statistical computing. 2019. Available from: https://www.R-project.org/.
51. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686
52. Auguie B. gridExtra: miscellaneous functions for "grid" graphics. *R Package Ver*. 2017;2:602.
53. Datta D. blandr: a Bland-Altman method comparison package for R. Zenodo. 2017.
54. Kassambara A. rstatix: pipe-friendly framework for basic statistical tests. 2019.
55. Cook J, Prairle M, Plante D. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy. *J Affect Disord*. 2017;217:299–305. doi:10.1016/j.jad.2017.04.030
56. Markwald RR, Wright KP. Circadian misalignment and sleep disruption in shift work: implications for fatigue and risk of weight gain and obesity. In: *Sleep Loss and Obesity*. Springer; 2012:101–118.
57. Collop N. Scoring variability between polysomnography technologists in different sleep laboratories. *Sleep Med*. 2002;3:43–47. doi:10.1016/S1389-9457(01)00115-0
58. Willemen T, Van Deun D, Verhaert V, et al. An evaluation of cardiorespiratory and movement features with respect to sleep stage classification. *IEEE J Biomed Health*. 2013;18(2):661–669. doi:10.1109/JBHI.2013.2276083
59. Brown AC, Smolensky MH, D'Alonzo GE, Redman DP. Actigraphy: a means of assessing circadian patterns in human activity. *Chronobiol Int*. 1990;7(2):125–133. doi:10.3109/07420529009056964
60. de Zambotti M, Cellini N, Menghini L, Sarlo M, Baker F. Sensors capabilities, performance, and use of consumer sleep technology. *Sleep Med Clin*. 2020;15(1):1–30. doi:10.1016/j.jsmc.2019.11.003
61. Lee J, Matsummura K, Yamakoshi KI, Rolfe P, Tanaka S, Yamakoshi T. Comparison between red, green and blue light reflection photoplethysmography for heart rate monitoring during motion. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology (EMBC); 2013.
62. De Arriba-pérez F, Caeiro-Rodríguez M, Santos-Gago J. Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. *Sensors*. 2016;16(9):1538. doi:10.3390/s16091538
63. Tuovinen L, Smeaton AF Unlocking the black box of wearable intelligence: ethical considerations and social impact. Paper presented at: 2019 IEEE Congress on Evolutionary Computation (CEC); January 10, 2019; Wellington, New Zealand.