Performance of wearable sleep trackers during nocturnal sleep and periods of simulated real-world smartphone use

Adrian R. Willoughby, PhD, Hosein Aghayan Golkashani, MD, PhD, Shohreh Ghorbani, MSc, Kian F. Wong, BA, Nicholas I.Y.N. Chee, BSc, Ju Lynn Ong, PhD, Michael W.L. Chee, MBBS

Sleep and Cognition Laboratory, Centre for Sleep and Cognition, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Abstract

Goal and aims: To test sleep/wake transition detection of consumer sleep trackers and research-grade actigraphy during nocturnal sleep and simulated peri-sleep behavior involving minimal movement.

Focus technology: Oura Ring Gen 3, Fitbit Sense, AXTRON Fit 3, Xiaomi Mi Band 7, and ActiGraph GT9X.

Reference technology: Polysomnography.

Sample: Sixty-three participants (36 female) aged 20-68.

Design: Participants engaged in common peri-sleep behavior (reading news articles, watching videos, and exchanging texts) on a smartphone before and after the sleep period. They were woken up during the night to complete a short questionnaire to simulate responding to an incoming message.

Core analytics: Detection and timing accuracy for the sleep onset times and wake times.

Additional analytics and exploratory analyses: Discrepancy analysis both including and excluding the peri-sleep activity periods. Epoch-by-epoch analysis of rate and extent of wake misclassification during peri-sleep activity periods.

Core outcomes: Oura and Fitbit were more accurate at detecting sleep/wake transitions than the actigraph and the lower-priced consumer sleep tracker devices. Detection accuracy was less reliable in participants with lower sleep efficiency.

Important additional outcomes: With inclusion of peri-sleep periods, specificity and Kappa improved significantly for Oura and Fitbit, but not ActiGraph. All devices misclassified motionless wake as sleep to some extent, but this was less prevalent for Oura and Fitbit.

Core conclusions: Performance of Oura and Fitbit is robust on nights with suboptimal bedtime routines or minor sleep disturbances. Reduced performance on nights with low sleep efficiency bolsters concerns that these devices are less accurate for fragmented or disturbed sleep.

© 2024 The Author(s). Published by Elsevier Inc. on behalf of National Sleep Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Growing interest in personal sleep measurement to improve health and well-being has driven significant growth in the development and sales of consumer sleep trackers (CSTs). With such a proliferation of devices, it is vital to ensure that their performance is evaluated adequately and that they are used appropriately for scientific research. As part of this effort, recent studies have found that the performance of higher-quality CSTs equals or exceeds that of “research-grade” actigraphy.

However, a major concern for both traditional actigraphs and newer CSTs is their ability to correctly classify motionless wake. Typically, these trackers have sensitivity for sleep detection of over 90%, meaning they correctly detect sleep epochs most of the time. However, their specificity, or correct detection of wakefulness when participants are in bed, is often under 60%.

Most performance evaluation studies have been conducted on healthy participants in a sleep laboratory. However, good performance in laboratory studies does not necessarily translate into good performance outside the lab or when patients with insomnia use CSTs without supervision, as has been highlighted in the media. Peri-sleep behavior inside and outside the lab differs significantly. Research has shown that many people use electronic devices during sleep, which can affect sleep quality. This highlights the need for CSTs to perform well during peri-sleep periods as well as during sleep itself.
devices while in bed before sleeping. Device use in bed is associated with periods of motionless wake that wearable devices may be poor at classifying.

The correct classification of motionless wake is important not only in healthy people using electronic or other devices, but also in those who have difficulty initiating or maintaining sleep. Previous research has shown that performance of both “research-grade” actigraphy and CSTs declines as sleep quality decreases. This has been demonstrated both in a comparison between insomnia patients and healthy sleepers and in the same participants under different conditions of sleep fragmentation.

To address these issues, we designed a protocol to simulate common peri-sleep behaviors in the sleep lab, allowing both control and measurement of participant behavior, and accurate evaluation of device performance. Participants engaged in everyday activities (reading, watching videos, and texting) on a smartphone while in bed, before going to sleep and after waking up. They were also woken during the night to answer a brief questionnaire, to simulate being woken by, and replying to, an incoming message. We tested two well-known and well-developed CSTs (Oura and Fitbit), two lower-priced CSTs with unknown development histories (Axstro and Xiaomi) and one “research-grade” actigraph (ActiGraph GT9X).

This protocol allowed us to investigate the degree to which wake detection affected estimates of sleep measures and performance evaluation metrics. These findings should inform both researchers and CST users concerned about device performance on nights where there is a modest departure from an ideal bedtime routine.

Methods

Data for this study was collected on the second night of a two-night sleep protocol that evaluated performance of different categories of sleep wearable devices (please see for additional methodological details and data from the first night). The protocol was approved by the Institutional Review Board of the National University of Singapore.

Sample

Participants were recruited from the National University of Singapore and the wider university community. Sixty-six participants were recruited from three different age groups to ensure comparable representation from a wider age range than commonly published on (Young: 18-30 years of age; Middle: 31-50; Older: 51-70). Data from three participants were excluded. One participant withdrew from the study after the first night. A second had such disturbed sleep (total sleep time [TST] = 2:42:30, sleep efficiency [SE] = 54.17%) that neither Oura nor Fitbit recorded any sleep. There was a data collection failure for the third participant. This resulted in a final sample of 63 participants.

Inclusion criteria were: (1) habitual sleep duration of at least 5 hours per night between 8 PM and 10 AM, (2) body mass index (BMI) less than 35 kg/m², (3) no self-reported pre-existing sleep, neurological, or psychiatric disorders, (4) no excessive daytime sleepiness (Epworth Sleepiness Scale score less than 11), (5) low risk for Obstructive Sleep Apnea (indicated by the Berlin questionnaire), (6) no caffeine or alcohol use within 6 hours prior to bedtime, (7) no active illness, and (8) not pregnant. Mean age, BMI, and sex distribution for the three age groups are shown in Table 1.

Design, study setting, and procedures

The study was conducted in a sleep research laboratory with controlled bedroom light and temperature and sound attenuation. To simulate common bedtime behaviors that could cause a wearable device to report sleep when a person was lying still but awake, participants completed a number of peri-sleep activities on a smartphone (Samsung A30s running Android 11) before, during, and after their night’s sleep.

Directly before going to sleep, participants engaged in three Pre-Sleep activities: reading news articles, watching nature videos, and texting with a chatbot (https://www.kuki.ai). Each activity was designed to take approximately 10 minutes to complete. The order of activities was counterbalanced across participants. Full details of the peri-sleep activities can be found in Section 1.1 of the Supplementary Material.

To simulate responding to an incoming message during the night, participants were woken by an alarm and completed a short questionnaire (adapted from an existing one) on the smartphone. They were woken up 5 minutes after the first appearance of N3 sleep, within 45 minutes of initial sleep onset. If N3 was not detected, the participant was woken at 45 minutes in N2 sleep. If they were in REM sleep after 45 minutes, they were woken during the next instance of N2 sleep. This Interruption split sleep into two main periods, Sleep Period 1 (SP1) and Sleep Period 2 (SP2).

If not already awake in the morning, participants were woken by an alarm at their requested wake-up time. They then completed two Post-Sleep activities designed to take around 15 minutes, which consisted of reading four news articles and completing a standardized questionnaire to evaluate their night’s sleep and current mood. An app that recorded phone usage (App Usage) was used to record the timing and duration of phone use during the activities (app use start and end times were rounded to the nearest 30 seconds to align with polysomnography [PSG] epochs). Bedroom lights remained off for the entire time the participants were in bed (including the Pre- and Post-Sleep activity periods). Participants could sit or lie in any comfortable position in bed while they completed the peri-sleep activities. The timeline of event is shown in Fig. 1; sleep onset and wake times were determined from PSG, bedtimes were derived from phone use logs.

Focus technology

Five wearable devices were evaluated: the Oura Ring Gen3, running Oura Sleep Staging Algorithm 2.0 (OSSA 2.0; Oura Health Oy, Oulu, Finland), the Fitbit Sense (Fitbit Inc, San Francisco, CA), the ActiGraph GT9X (ActiGraph Inc, Pensacola, FL), the Xiaomi Mi Band 7 (Xiaomi Inc, Beijing, China) and the AXTRO Fit 3 Fitness Tracker (AXTRO Pte. Ltd, Singapore). For convenience, we refer to these devices by their manufacturer’s name. The wearable devices and associated apps were updated to the latest version prior to the start of the study (see Supplementary Table S1 for firmware and software details).

Four of the five devices were worn concurrently (Oura, Fitbit, ActiGraph, and either Xiaomi or Axtro); this was done both for participant comfort and to keep the devices as close as possible to the

---

**Table 1**

Demographic details of the participants sorted by age group

<table>
<thead>
<tr>
<th>Age group</th>
<th>Overall</th>
<th>Young</th>
<th>Middle</th>
<th>Older</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>39.36</td>
<td>23.19</td>
<td>37.65</td>
<td>59.95</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>(15.51)</td>
<td>(2.23)</td>
<td>(4.90)</td>
<td>(6.17)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27 M/36 F</td>
<td>6 M/15 F</td>
<td>12 M/11 F</td>
<td>9 M/10 F</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>(3.33)</td>
<td>(2.44)</td>
<td>(4.12)</td>
<td>(2.70)</td>
<td>.054</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>23.84</td>
<td>22.49</td>
<td>24.86</td>
<td>24.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.33)</td>
<td>(3.33)</td>
<td>(4.12)</td>
<td>(2.70)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: BMI, body mass index.

For age and BMI, means (and standard deviation) are reported. Statistical significance column shows ANOVA main effect (for age and BMI) and chi-squared (for sex) significance. Values with different letters are significantly different from each other (pairwise comparison: p < .05).
correct wrist position). The Oura ring was worn on either the index or ring finger of the nondominant hand; the other devices were worn on the nondominant wrist. Sleep-staged epochs for each device were resampled to match the 30-second epochs of the PSG recording (see Section 1.3 of the Supplementary Material).

To demarcate the period for sleep scoring, ActiGraph requires manual entry of markers. To compare the effects of marker position, we analyzed data scored with the markers placed in two different positions. One data set was created by placing the markers when participants got into bed and out of bed (i.e., including the Pre-Sleep and Post-Sleep periods; termed ActiGraph-IB, for “In Bed”). The second data set was created by placing the markers at bedtime and at the participants’ requested wake time (i.e., excluding the Pre-Sleep and Post-Sleep periods, equivalent to the traditional “lights off” and “lights on” times; termed ActiGraph-BT, for “Bedtime”).

For three participants, Fitbit failed to record data correctly and for one participant ActiGraph data was lost due to a technical error. Final participant numbers for each device were Oura: 63, Fitbit: 60, ActiGraph: 62, Xiaomi: 29, and Axtro: 31.

Reference technology

PSG was collected using the SOMNOtouch RESP system (SOMNOmedics GmbH, Randersacker, Germany). Electroencephalogram data was recorded from electrodes C3 and C4 referenced to the con-tralateral mastoids (M2 and M1). Electro-oculogram, electromyogram, and electrocardiogram were recorded using standard procedures. The PSG was scored by consensus using three systems (Neurobit: Neurobit Inc, New York; Somnolyzer 24×7: The Siesta Group Schlafanalyse GmbH, Vienna, Austria; U-Sleep: https://sleep.ai.ku.dk25), according to a hybrid American Academy of Sleep Medicine26 (AASM) and Rechtschaffen and Kales27 system (see Section 1.4 of the Supplementary Material).

Table 2

Start time and duration of peri-sleep activity periods for participants in each age group

<table>
<thead>
<tr>
<th>Activity period</th>
<th>Measure</th>
<th>Overall</th>
<th>Age group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Young</td>
</tr>
<tr>
<td>Pre-sleep</td>
<td>Start time (HH:MM)</td>
<td>23:16 (68.3 min)</td>
<td>23:55 (55.7 min)a</td>
</tr>
<tr>
<td></td>
<td>Duration (min)</td>
<td>36.04 (7.32)</td>
<td>33.88 (6.30)a</td>
</tr>
<tr>
<td>Interruption</td>
<td>Start time (HH:MM)</td>
<td>00:36 (71.3 min)</td>
<td>01:04 (61.6 min)a</td>
</tr>
<tr>
<td></td>
<td>Duration (min)</td>
<td>2.71 (1.81)a</td>
<td>2.91 (0.98)a</td>
</tr>
<tr>
<td>Post-sleep</td>
<td>Start time (HH:MM)</td>
<td>07:06 (70.1 min)</td>
<td>07:34 (62.3 min)a</td>
</tr>
<tr>
<td></td>
<td>Duration (min)</td>
<td>15.90 (6.80)</td>
<td>12.81 (4.03)a</td>
</tr>
</tbody>
</table>

Mean (and standard deviation) start time and duration of peri-sleep activity periods. Statistical significance column shows ANOVA main effect significance. Values with different letters are significantly different from each other (pairwise comparison: p < .05).
full data recording period (including the peri-sleep periods) with data excluding the peri-sleep periods.

Results

Peri-sleep activity period timing and duration

Table 2 shows the timing and duration of the peri-sleep activity periods for the three age groups. In general, the activity periods started earlier, and were longer, for the Older group than the Young and Middle age groups.

PSG-derived sleep measures

PSG-derived sleep measures for each age group are shown in Table 3. Overall sleep patterns are consistent with research characterizing changes in sleep with age. Participants in the Older group had shorter and less efficient sleep than the other groups; participants in both the Older and Middle groups showed less Deep sleep and more Light sleep than those in the Young group.

Bedtime detection

Bedtime is typically defined as the time at which a person goes to bed with the intention to sleep. In laboratory studies, this usually corresponds to “lights off” time. In this study, participants got into bed, and lights were turned off, prior to the Pre-Sleep period. Therefore, bedtime was defined as the first epoch following completion of the Pre-Sleep activities, indicating the point at which participants first tried to fall asleep.

Of the devices evaluated, only Oura and Fitbit provided automated measures of bedtime. Both devices reported bedtime as later than it actually was; Oura showed a mean difference of 8.57 minutes (SD = 24.65) and Fitbit 9.75 minutes (SD = 23.97). This is consistent with findings indicating that wearables typically underestimate SOL.

Fig. 2 shows the distribution of differences in bedtimes reported by

<table>
<thead>
<tr>
<th>PSG sleep measure</th>
<th>Overall</th>
<th>Young</th>
<th>Middle</th>
<th>Older</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIB (min)</td>
<td>433.21  (43.13)</td>
<td>425.60 (48.16)</td>
<td>428.33 (41.81)</td>
<td>447.53 (37.14)</td>
<td>.22</td>
</tr>
<tr>
<td>TST (min)</td>
<td>379.49  (42.66)</td>
<td>364.33 (45.74)</td>
<td>391.54 (40.79)</td>
<td>359.55 (35.74)</td>
<td>.04</td>
</tr>
<tr>
<td>WASO (min)</td>
<td>43.45   (32.38)</td>
<td>29.69 (20.96)</td>
<td>31.28 (17.95)</td>
<td>73.39 (37.06)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>SE (%)</td>
<td>87.88   (7.98)</td>
<td>90.48 (6.29)</td>
<td>91.47 (4.53)</td>
<td>80.68 (8.45)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>SP1 SOL (min)</td>
<td>10.46   (12.90)</td>
<td>11.64 (15.13)</td>
<td>5.50 (5.20)</td>
<td>15.15 (15.16)</td>
<td>.04</td>
</tr>
<tr>
<td>SP2 SOL (min)</td>
<td>10.14   (16.65)</td>
<td>9.76 (18.74)</td>
<td>8.43 (14.92)</td>
<td>12.63 (16.79)</td>
<td>.72</td>
</tr>
<tr>
<td>Light sleep (min)</td>
<td>222.89  (42.99)</td>
<td>200.64 (46.41)</td>
<td>232.43 (31.35)</td>
<td>235.92 (40.10)</td>
<td>.01</td>
</tr>
<tr>
<td>Deep sleep (min)</td>
<td>65.95   (32.31)</td>
<td>92.48 (30.60)</td>
<td>60.21 (23.91)</td>
<td>44.08 (23.76)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>REM sleep (min)</td>
<td>90.74   (34.90)</td>
<td>91.86 (22.78)</td>
<td>98.91 (19.44)</td>
<td>79.61 (29.67)</td>
<td>.04</td>
</tr>
</tbody>
</table>

Abbreviations: Deep sleep, N3 sleep; Light sleep, N1 and N2 sleep; REM sleep, rapid eye movement sleep; SE, sleep efficiency; SOL, sleep onset latency; SP1/2, sleep period 1/2; TIB, time in bed; TST, total sleep time; WASO, wake after sleep onset.

Mean (and standard deviation) PSG-derived sleep measures are shown. Statistical significance column shows ANOVA main effect significance. Values with different letters are significantly different from each other (pairwise comparison: p < .05).
Our and Fitbit compared to the phone usage defined bedtime for all participants. While the mean and median differences were relatively small, the MAE was notably larger (Oura: 16.43 (SD = 20.19) minutes; Fitbit: 15.76 (SD = 20.45) minutes). There were also several instances when devices misjudged bedtime by over an hour.

**Sleep Period 1**

An epoch-by-epoch analysis (see Section 2.1 of the Supplementary Material) showed that ActiGraph misclassified wake as sleep more often than Oura and Fitbit. However, epoch-by-epoch analyses characterize classification performance without regard to error distribution and may not reflect how accurately devices detect sleep/wake transitions during the night. In the following sections, we describe detection and timing accuracy for these timepoints in temporal order.

Sleep Period 1 SOT could be detected in one of three time windows: incorrectly during the Pre-Sleep period (before bedtime), during the correct window after bedtime and before the Interruption (when the participant was either trying to fall asleep or was asleep) or after the Interruption. Fig. 3 shows the detection rate for each device in time each time window. Table 4 (first section) presents this information together with the distribution of instances, and the average difference from the PSG-derived SOT across participants.

ActiGraph-BT detected SP1 SOT accurately most of the time, with no sleep onsets detected late. There were also several instances when devices missed the correct window. Oura and Fitbit performed fairly accurately, reporting sleep onset early only once, but detecting it late nine times. Fitbit and Xiaomi performed similarly with a roughly even split in errors detecting the sleep onset early and late. Axtro, while never missing the sleep onset, reported it before bedtime for half of the participants.

**Sleep Efficiency and Sleep Period 1 SOT detection accuracy**

SE is often used as an indicator of sleep continuity. Fig. 5 shows, for each device, the relationship between SE and accuracy of SP1 SOT estimation. Members of the three age groups are shown in different colors. For most devices, at higher SE there is a cluster of points relatively close to zero. The slope of this cluster is low, suggesting that when SE is high, SOT is estimated accurately. However, at lower SE (below ~85%) points are more spread out, indicating that SOT estimation is less accurate. The slope of the regression lines suggest that ActiGraph reported SOT increasingly earlier for participants with lower SE.

**Mid-sleep interruption**

All participants were woken up from N2 or N3 sleep to complete the activities during the Interruption period. The Interruption Wake Time (IWT) could only be estimated if the device identified sleep beforehand; therefore, this analysis includes only those participants where the device recorded sleep prior to the Interruption.

The IWT was considered correctly detected if it was identified in the time window from the epoch prior to the PSG-defined awakening caused by the alarm (to allow for discrepancies caused by splitting 1-minute epochs for some devices) to Sleep Period 2 SOT (i.e., while the participant was awake following the Interruption). There are two potential types of errors; one where the participant was incorrectly classified as awake prior to the Interruption alarm and one where wakefulness during the Interruption period was missed. In both cases, the correct transition from sleep to wake caused by the alarm was not identified correctly. Fig. 6 illustrates the detection windows and IWT detection rates, with additional results shown in the second section of Table 4.

Oura, Fitbit, and ActiGraph were very sensitive to awakening during the Interruption period; it was missed only once. Axtro and Xiaomi, however, missed most or all of the Intermittents. Fig. 7 shows the IWT bias when it was not missed (including instances when the device reported the wake interruption before its detection by PSG). Oura, Fitbit, and ActiGraph frequently detected the wake Interruption within one epoch (30 seconds) of the PSG wake epoch (see Supplementary Fig. S1 for the Oura outlier). When Axtro successfully detected the wake Interruption, it was reported on average over 5 minutes after PSG. Xiao is not shown in this figure because it missed all of the IWTs (see Supplementary Fig. S2).
<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection accuracy, timing accuracy, and error magnitude for Sleep Period 1 Sleep Onset Time, Interruption Wake Time, Sleep Period 2 Sleep Onset Time and Morning Wake Time</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Device</th>
<th>Detection accuracy</th>
<th>Bias: Median (MAD, min)</th>
<th>Bias: Mean (SD, min)</th>
<th>MAE (SD, min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleep Period 1 Sleep Onset Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oura</td>
<td>0.82 (1/56 early)</td>
<td>0.04 (7/53 early)</td>
<td>16.93 (20.99)</td>
<td></td>
</tr>
<tr>
<td>Fitbit</td>
<td>0.66 (9/56 late)</td>
<td>0.66 (11/53 late)</td>
<td>15.16 (18.32)</td>
<td></td>
</tr>
<tr>
<td>ActiGraph-IB</td>
<td>0.04 (53/55 early)</td>
<td>3.33 (22.00)</td>
<td>39.07 (17.87)</td>
<td></td>
</tr>
<tr>
<td>ActiGraph-BT</td>
<td>0.96 (2/55 early)</td>
<td>-36 (10.38)</td>
<td>8.11 (12.67)</td>
<td></td>
</tr>
<tr>
<td>Axtro</td>
<td>0.5 (14/28 early)</td>
<td>-5.31 (14.10)</td>
<td>25.41 (19.15)</td>
<td></td>
</tr>
<tr>
<td>Xiaomi</td>
<td>0.6 (14/28 late)</td>
<td>-15.5 (30.39)</td>
<td>5.5 (35.30)</td>
<td></td>
</tr>
<tr>
<td>Interruption Wake Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oura</td>
<td>0.94 (3/54 early)</td>
<td>-0.5 (3.74)</td>
<td>0.96 (1.50)</td>
<td></td>
</tr>
<tr>
<td>Fitbit</td>
<td>0.83 (8/47 early)</td>
<td>0.7 (0)</td>
<td>5.43 (3.87)</td>
<td></td>
</tr>
<tr>
<td>Axtro</td>
<td>0.63 (22/62 early)</td>
<td>0 (0.74)</td>
<td>5 (2.97)</td>
<td></td>
</tr>
<tr>
<td>Xiaom</td>
<td>0.23 (24/62 missed)</td>
<td>5 (3.74)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Sleep Period 2 Sleep Onset Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oura</td>
<td>1 (63/63 correct)</td>
<td>1 (60/60 correct)</td>
<td>5.64 (8.79)</td>
<td></td>
</tr>
<tr>
<td>Fitbit</td>
<td>2 (0.45)</td>
<td>-0.75 (5.56)</td>
<td>7.33 (16.25)</td>
<td></td>
</tr>
<tr>
<td>Axtro</td>
<td>1.44 (10.37)</td>
<td>-5.96 (17.76)</td>
<td>9.10 (16.35)</td>
<td></td>
</tr>
<tr>
<td>Xiaom</td>
<td>1.68 (7.64)</td>
<td>5.11 (17.26)</td>
<td>10.44 (14.30)</td>
<td></td>
</tr>
<tr>
<td>Morning Wake Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oura</td>
<td>0.98 (1/61 early)</td>
<td>1</td>
<td>3.50 (7.41)</td>
<td></td>
</tr>
<tr>
<td>Fitbit</td>
<td>0.97 (2/58 early)</td>
<td>1 (60/60 correct)</td>
<td>3.50 (7.41)</td>
<td></td>
</tr>
<tr>
<td>Axtro</td>
<td>0.48 (1/61 missed)</td>
<td>1 (60/60 correct)</td>
<td>10.44 (14.30)</td>
<td></td>
</tr>
<tr>
<td>Xiaom</td>
<td>0.71 (18.18)</td>
<td>1</td>
<td>11.43 (13.37)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: MAD, median absolute deviation; MAE, mean absolute error.
Detection accuracy: Proportion (and number) of participants where the SOT/WT was detected in the correct window. Bias: Average difference between device and PSG-derived SOT/WT. The best-performing device for each timepoint and each measure is shown in bold (there are ties for some measures).
Sleep Period 2

All participants were able to fall asleep again following the Interruption. For each device, Sleep Period 2 SOT was defined as the first wake/sleep transition following the Interruption alarm up to 1 hour after the PSG-defined SP2 SOT (Fig. 8). SP2 SOT could only be calculated for participants where the device either recorded a wake period related to the Interruption or where Sleep Period 1 was missed (in which case, the first sleep onset recorded was for SP2). SP2 SOT could not be detected if the device identified SP1 but no Interruption wake period (see Supplementary Fig. S2). All devices detected SP2 SOT during the correct window when they identified wake during the Interruption period (see third section of Table 4).

In general, when devices detected SP2 SOT, it was fairly accurate, as illustrated by the clustering around zero in Fig. 9. There were some instances in which participants took a long time to return to sleep and devices reported SOT considerably earlier than PSG (see Supplementary Fig. S2).

Morning wake up

A morning alarm woke participants at their requested wake-up time, if they were not already awake. The Morning Wake Time (MWT) was considered missed if no transitions from sleep to wake were detected from 45 minutes before the final PSG awakening to the end of the Post-Sleep period (Fig. 10). Table 4 (bottom section) shows the detection accuracy and timing for the MWT (either in response to the alarm, or the final awakening before the alarm). Oura, Fitbit, and ActiGraph successfully detected a MWT for each participant; however, Axtro and Xiaomi frequently missed it.

Fig. 11 shows the differences between PSG and device estimates of MWT when it was not missed. Oura, Fitbit, and ActiGraph are all fairly accurate apart from a few instances where the MWT was estimated much earlier than the PSG-derived MWT. These were cases where the participant woke up early (which was detected by the device) and then fell back to sleep before waking up for the final time (which was not detected by the device; see Supplementary Fig. S3). Both Axtro and Xiaomi frequently failed to detect the morning awakening; furthermore, when it was detected, it was reported later than the PSG-derived MWT.

Sleep measures and performance evaluation metrics across nights

Table 5 shows PSG sleep measures and device discrepancy (upper section), and device performance metrics (lower section) for both nights of the protocol. The second night was analyzed both including the peri-sleep periods (labeled Night 2 IB, for “In Bed”) and excluding them (labeled Night 2 BT, for “Bedtime”).

PSG data showed that, when excluding the peri-sleep periods, TST and SE were similar on Night 1 and Night 2. However, both measures were underestimated by Oura and Fitbit on Night 2 compared with Night 1 (upper bold values in Table 5).

Including the peri-sleep periods in the analysis had a significant impact on performance evaluation metrics. For both Oura and Fitbit, specificity increased, while for ActiGraph specificity decreased. Kappa scores for Oura and Fitbit also increased, while that for ActiGraph remained unchanged (lower bold values in Table 5).

Discussion

Main results and implications

Using a novel protocol to simulate common peri-sleep behavior, we found that modest deviation from a good bedtime routine does not significantly compromise the performance of higher-quality CSTs. Consistent with data from the first night of the study, the high-quality CSTs performed considerably better than the lower-cost ones.

Oura and Fitbit displayed good detection accuracy for Sleep Period 1 SOT. While Oura’s detection accuracy was higher than Fitbit’s, it was biased towards detecting sleep onset later than it actually was. This is consistent with results from the first night of the protocol where Oura significantly overestimated sleep onset latency. While Axtro and Xiaomi performed moderately well for the initial
sleep onset, their inability to detect awakenings may severely limit their usefulness.

ActiGraph’s Sleep Period 1 SOT accuracy was significantly influenced by marker placement. Both detection and timing accuracy were markedly better for the ActiGraph-BT data, when the algorithm excluded the Pre-Sleep period. This improvement is not surprising because wake misclassification during the Pre-Sleep period, resulting in earlier estimation of sleep onset, is not considered in the ActiGraph-BT data; however, it does highlight the importance of correct marker placement for accurate sleep measurement. This pattern of results is consistent with ActiGraph-IB’s overestimation of TST and SE, resulting from wake misclassified as sleep in the peri-sleep periods. It also suggests that in a controlled lab environment, where participants do not engage in typical peri-sleep activities, accuracy of sleep measures may be artificially inflated. Oura and Fitbit’s performance during these periods suggests that this may be less of a concern for these devices.

Sleep Period 1 SOT timing accuracy was less reliable for those with more fragmented sleep. This is consistent with previous research, including Night 1 of this protocol. Sleep fragmentation is associated with normal aging as well as neurodegenerative disorders such as Alzheimer’s disease and Parkinson’s disease; therefore, data from wearable devices in these populations is likely to be less reliable than from those with well-consolidated sleep.

Detection accuracy for both the Interruption and Morning Wake Times was very high for Oura, Fitbit, and ActiGraph; however, these transitions were missed by Astro and Xiaomi over half the time. ActiGraph’s ability to detect both wake and sleep onset transitions

---

**Fig. 5.** Scatter plots showing the relationship between sleep efficiency (SE) and Sleep Period 1 Sleep Onset Time accuracy. Participants from the Young group are shown in Green, the Middle group in Gold, and the Older group in Red. The regression line is plotted in black with confidence intervals (dashed black lines). The vertical dashed blue line marks 85% SE.
when peri-sleep behavior was excluded suggests that it is sensitive to the movement associated with the sleep transitions in the protocol (e.g., putting the phone on the bedside table and finding a comfortable position to sleep in, or retrieving the phone and getting in position to do the activities). However, on either side of the movement, when the participant was still, ActiGraph reported the participant as asleep.

The marginally better sleep on Night 2 compared with Night 1 might be due to the first-night effect and participants wearing the Dreem headband which was not well-tolerated by some participants on Night 1. Both Oura and Fitbit significantly underestimated TST and SE on Night 2, but not Night 1. This might be because the wake interruption resulted in these CSTs missing the first SOT for some

---

**Fig. 6.** Detection accuracy for the Interruption Wake Time (IWT). The blue region in the timeline indicates the window that was considered correct detection. The table below shows color-coded accuracy for each device in detecting the IWT in the correct window. Lines indicate the extent and direction of errors.

**Fig. 7.** Interruption Wake Time (IWT) accuracy for all devices and participants when it was not missed. The inset shows a zoomed-in version of the graph on the left. Xiaomi is not shown because it missed all interruptions.

**Fig. 8.** Detection period for Sleep Period 2 Sleep Onset Time (SP2 SOT). The blue region indicates the window that was considered correct detection. All devices detected SP2 SOT in the correct window.
participants. This may, in turn, have resulted in a shorter average TST and lower SE. This explanation implies that an event which effects sleep at one point in time may have broader consequences for sleep measures at a different time.

There were significant differences in performance metrics depending on whether the peri-sleep periods were included in the analysis or not. Both specificity and Kappa increased for Oura and Fitbit because additional periods of predominantly correct wake classification were included in the analysis. For ActiGraph, specificity decreased and Kappa remained the same. This is because the classification accuracy of the periods added to the analysis were at chance levels (see Supplementary Table S2). The effect of this was to reduce specificity. However, the Kappa statistic was unaffected because it accounts for chance agreement, which was similar for the peri-sleep periods and the wake periods between bedtime and the morning alarm.

One reason that well-developed CSTs outperform “research-grade” actigraphy may be that they use data from multiple sensors, compared to actigraphs (typically including photoplethysmography and temperature sensors). Several recent studies have shown that including data from additional sources in classification algorithms improves performance; more broadly, research has also demonstrated the value of combining data sources to overcome limitations of individual measures.

Another possible reason for CSTs’ superior performance is that they use algorithms derived from larger training sets that include data from a
Fig. 11. Morning Wake Time accuracy for all devices and participants where a wake time was detected. The inset shows a zoomed-in version of the graph on the left.

wider age range, and from people of different ethnicities. Actigraphs typically use variants of the Cole-Kripke or Sadeh algorithms, both of which were developed approximately 30 years ago. Using newer algorithms on the same actigraphy data can improve classification performance, although the improvement afforded by new machine learning algorithms may not yet be as robust as originally envisioned.

While high-quality CSTs can be improved further, their relative affordability, reliable performance and already extensive deployment provide valuable opportunities for sleep research. Large-scale studies on cross-country differences in sleep patterns have already been reported, although they lack valuable phenotypic data that future studies should include.

Table 5
Discrepancy analyses comparing Oura, Fitbit, and ActiGraph with PSG across both nights of the protocol and including/excluding peri-sleep periods (N = 50)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Dataset</th>
<th>PSG</th>
<th>Oura</th>
<th>Fitbit</th>
<th>ActiGraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>TST (min)</td>
<td>Night 1</td>
<td>365.06 (55.77)</td>
<td>2.77 (37.40)^X</td>
<td>9.97 (43.18)^XY</td>
<td>12.60 (46.82)^X</td>
</tr>
<tr>
<td></td>
<td>Night 2 BT</td>
<td>377.51 (42.62)</td>
<td>-10.68 (26.70)^XY</td>
<td>-12.65 (33.69)^XY</td>
<td>-2.88 (41.50)^Y</td>
</tr>
<tr>
<td></td>
<td>Night 2 IB</td>
<td>377.51 (42.62)</td>
<td>-9.99 (25.98)^XY</td>
<td>-10.52 (35.64)^XY</td>
<td>21.36 (47.51)^b</td>
</tr>
<tr>
<td>SE (%)</td>
<td>Night 1</td>
<td>84.42 (10.14)^X</td>
<td>0.60 (8.91)^X</td>
<td>2.04 (9.84)^Y</td>
<td>2.76 (10.84)^X</td>
</tr>
<tr>
<td></td>
<td>Night 2 BT</td>
<td>87.57 (8.33)^X</td>
<td>-2.59 (6.14)^Y</td>
<td>-3.12 (7.57)^Y</td>
<td>-0.95 (9.39)^Y</td>
</tr>
<tr>
<td></td>
<td>Night 2 IB</td>
<td>78.48 (8.05)^Y</td>
<td>-2.15 (5.19)^Y</td>
<td>-2.36 (7.09)^Z</td>
<td>4.11 (9.46)^X</td>
</tr>
<tr>
<td>SOL (min)</td>
<td>Night 1</td>
<td>15.52 (14.55)^X</td>
<td>10.99 (25.80)^Y</td>
<td>-1.39 (23.20)^b</td>
<td>-9.97 (15.22)^c</td>
</tr>
<tr>
<td></td>
<td>Night 2 BT/IB</td>
<td>9.92 (12.15)^Y</td>
<td>16.07 (23.99)^b</td>
<td>8.60 (21.98)^b</td>
<td>3.70 (12.39)^ab</td>
</tr>
<tr>
<td>WASO (min)</td>
<td>Night 1</td>
<td>52.81 (41.61)^XY</td>
<td>-13.76 (32.63)^ab</td>
<td>-8.58 (36.78)^ab</td>
<td>-2.64 (42.93)^b</td>
</tr>
<tr>
<td></td>
<td>Night 2 BT</td>
<td>45.32 (32.78)^X</td>
<td>-5.45 (21.99)^Y</td>
<td>3.99 (27.00)^b</td>
<td>6.58 (34.62)^b</td>
</tr>
<tr>
<td></td>
<td>Night 2 IB</td>
<td>60.13 (35.25)^Y</td>
<td>-3.80 (25.00)^b</td>
<td>3.80 (27.71)^b</td>
<td>15.91 (37.92)^b</td>
</tr>
</tbody>
</table>

Measure of accuracy with asterisks (*) indicate bias is significantly different from 0 (p < .05); a,b,c letters indicate significant differences between devices (read across rows; p < .05); X,Y,Z letters indicate significant differences between data sets (read down column for each measure; p < .05). All p-values are Bonfmerroni corrected for multiple comparisons. Comparisons discussed in the text are indicated in bold font.
Limitations and future perspectives

The technological ecosystem in which today’s CSTs are being developed is significantly different from that in which the original actigraphs were created. This means that algorithm and software improvements can be deployed more quickly and at a much larger scale than before. The change in marker position for the ActiGraph data resulted in a significant improvement in performance, suggesting that CSTs could also exploit user input to improve sleep-tracking performance. For example, allowing users to indicate when they get into and out of bed, and when they try to fall asleep and wake up, would help disambiguate periods that are challenging for CSTs to classify accurately. While consistent user input of this kind is unlikely to be sustained in the long term, it could be used to give dedicated users more accurate sleep measures when they need it most.

Furthermore, with the wealth of data currently being collected, algorithms could be tuned to specific populations to optimize performance.52 One feature that many research actigraphs have, which CSTs are yet to implement, is the ability to change sensitivity threshold levels. This may be particularly important in maintaining accuracy across different populations (e.g., children and adolescents, particularly boys, move more than adults during sleep, which requires different sensitivity thresholds to analyze appropriately53,54). Tuning sleep detection algorithms to accommodate aspects of physiology reliably associated with different sleep conditions as well as better handling of age effects are improvements to look forward to as more data is collected from increasingly diverse populations. However, this will require high-fidelity data labeling to fully realize, and it remains to be seen if this aspiration receives the necessary cooperation from device users and manufacturers.

Different devices provide sleep measures at different temporal resolutions. In order to compare these to PSG sleep measures, they need to be resampled to PSG epoch length (typically 30 seconds). This may, however, affect discrepancy measures, particularly for people with fragmented sleep who experience many stage transitions. While this is unlikely to have a significant impact on large effects found in performance evaluation studies in healthy sleepers, it should be considered when comparing device performance.

Core conclusions

Our results indicate that higher-quality CSTs provide robust sleep measurement in the face of a suboptimal bedtime routine and moderate sleep disruption. However, as accuracy decreased with lower sleep quality, these devices can be expected to be less accurate in those with insomnia or disturbed sleep. Future research, involving performance evaluation in different settings, during different user behaviors and on different populations will prove invaluable in delineating the boundaries within which CSTs provide acceptable sleep measures. The paradigm described in the present work should prove an important methodological advance for that purpose.

Funding

This work was supported by funds from Oura Health Oy (Finland), the Yong Loo Lin School of Medicine (National University of Singapore, Singapore), the Lee Foundation (Singapore), and the National Medical Research Council Singapore (StAR May2019-001) awarded to Dr Michael W.L. Chee.

Author contributions

A.W.: Conceptualization, Formal analysis, Visualization, Writing – original draft, Writing – review and editing. H.G. and S.G.: Formal analysis, Data curation, Visualization. K.W. and N.C.: Investigation, Data curation. J.O.: Conceptualization, Project administration, Supervision, Writing – review and editing. M.C.: Conceptualization, Supervision, Funding acquisition, Writing – original draft, Writing – review and editing.

Declaration of conflicts of interest

Oura Health Oy funded the data collection for the evaluation of its new sleep staging algorithm (OSSA 2.0), but the company did not influence the design of the study, analyses, its interpretation, or data presentation. All other equipment was contributed by the Sleep and Cognition Laboratory.

Acknowledgments

The authors would like to thank Yashmit Lepcha, Liang Tian, and Nicha Turton for their assistance in data collection and Kyra Chong for figure preparation.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.sleh.2024.02.007.

References

1. de Zambotti M, Goldstein C, Cook J, et al. State of the science and recommendations for using wearable technology in sleep and circadian research. Published online Sleep. 2023;zasd325. https://doi.org/10.1093/sleep/zzasd325

A.R. Willoughby et al. Sleep Health: Journal of the National Sleep Foundation 10 (2024) 356–368
Ohayon MM, Carskadon MA, Guilleminault C, Vitiello MV. Meta-analysis of

Lefkowitz C. Staff Perspective: Why We Don’t Recommend Wearable Sleep
Trackers for the Assessment and Treatment of Insomnia Disorder. Published 2021.
Available at: https://deploymentspsych.org/blog/staff-perspective-why-we-dont-
Accessed January 10, 2024.

Chung SJ, An H, Suh S. What do people do before going to bed? A study of bedtime
10.1093/sleep/zsz267

Exelmann L, Gradisar M, Van den Bulck J. Sleep latency versus shutey latency:
prevalence, predictors and relation to insomnia symptoms in a representative
12737

Exelmann L, Van den Bulck J. Bedtime, shutey time and electronic media: sleep
org/10.1111/jsr.12510

Kang SC, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial
wearable sleep tracker in adult insomnia disorder patients and good sleepers. J
Pyschosom Res. 2017;97:38–44. https://doi.org/10.1016/j.jpsychires.2017.03.009

Paquet J, Kawniska A, Carrier J. Wake detection capacity of actigraphy during

Johns MW. A new method for measuring daytime sleepiness: the Epworth
540

Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin
Questionnaire to identify patients at risk for the sleep apnea syndrome. Ann

Siciliani F, LaRoque J, Postle B, Tontini G. Assessing sleep consciousness
doi.org/10.3389/fpsyg.2013.00452

Stephan AM, Lecci S, Cataldi J, Siciliani F. Conscious experiences and high-density

ASoFT Technology Inc. App Usage. Published online 2023. Available at: https://play.
google.com/store/apps/details?id=com.a0soft.gphone.uninstaller&hl=en

1038/s41776-021-00440-5

Iber C, Ancoli-Israel S., Chossan A, Quan SF. The AASM Manual for the Scoring of

Rechtschaffen A, Kales A. A Manual of Standardized Terminology, Techniques and
Scoring System for Sleep Stages of Human Subjects. Bethesda, MD: US Department
Printing Office; 1968.

The MathWorks Inc. MATLAB version: 9.13.0 (R2022b). Published online 2022.
Available at: https://www.mathworks.com/

R Core Team. R: A Language and Environment for Statistical Computing. R

IBM Corp. IBM SPSS Statistics for Macintosh. Published online 2020.

Ghayan MM, Carskadon MA, Guillenmann C, Vitello MV. Meta-analysis of
quantitative sleep parameters from childhood to old age in healthy individuals:
developing normative sleep values across the human lifespan. Sleep.

Scott H, Lack L, Lovato N. A systematic review of the accuracy of sleep
doi.org/10.1016/j.smrv.2019.10.1227

Bechtold B. Violin Plots for Matlab, Github Project. Published online 2016. doi: 10.
5281/zenodo.4559847.