Received: 12 June 2019 | Revised: 28 August 2019 | Accepted: 14 October 2019

DOI: 10.1111/jsr.12944



Journal of Sleep Research | ESRS

REGULAR RESEARCH PAPER

# Validity, potential clinical utility and comparison of a consumer activity tracker and a research-grade activity tracker in insomnia disorder II: Outside the laboratory

Kellie Hamill[1] | Ria Jumabhoy[1] | Piyumi Kahawage[1] | Massimiliano de Zambotti[2] | Elizabeth M Walters[1] | Sean P. A. Drummond[1]

[1]Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton, Vic, Australia

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

**Correspondence**
Sean P. A. Drummond, Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, 18 Innovation Walk, Clayton, Vic. 3800, Australia.
Email: sean.drummond@monash.edu

**Funding information**
National Health and Medical Research Council Approval ID: 1105458.

## Abstract

Accurate assessment of sleep can be fundamental for monitoring, managing and evaluating treatment outcomes within diseases. A proliferation of consumer activity trackers gives easy access to objective sleep. We evaluated the performance of a commercial device (Fitbit Alta HR) relative to a research-grade actigraph (Actiwatch Spectrum Pro) in measuring sleep before and after a cognitive behavioural intervention in insomnia disorder. Twenty-five individuals with DSM-5 insomnia disorder ($M$ = 50.6 ± 15.9 years) wore Fitbit and Actiwatch and completed a sleep diary during an in-laboratory polysomnogram, and for 1 week preceding and following seven weekly sessions of cognitive-behavioural intervention for insomnia. Device performance was compared for sleep outcomes (total sleep time, sleep latency, sleep efficiency and wake after sleep onset). The analyses assessed (a) agreement between devices across days and pre- to post-treatment, and (b) whether pre- to post-treatment changes in sleep assessed by devices correlated with clinical measures of change. Devices generally did not significantly differ from each other on sleep variable estimates, either night to night, in response to sleep manipulation (pre- to post-treatment) or in response to changes in environment (in the laboratory versus at home). Change in sleep measures across time from each device showed some correlation with common clinical measures of change in insomnia, but not insomnia diagnosis as a categorical variable. Overall, the Fitbit provides similar estimates of sleep outside the laboratory to a research grade actigraph. Despite the similarity between Fitbit and Actiwatch performance, the use of consumer technology is still in its infancy and caution should be taken in its interpretation.

KEYWORDS
actigraphy, activity monitor, consumer sleep tracker, night-to-night variability, wearables

## 1 | INTRODUCTION

Sleep is a fundamental component of a person's well-being, with poor sleep being implicated in several disease conditions (e.g., cardiovascular and inflammatory disease, and dementia). Insomnia is the most commonly reported sleep disorder, with 20% of Australian adults experiencing difficulty getting to sleep, staying asleep or waking up to early to some degree (Adams et al., 2017). Sleep diaries are frequently used in the diagnosis, management and treatment of the disorder, allowing researchers and health professionals to calculate

**Clinical Trial**: Researching Effective Sleep Treatments (Project REST), ANZCTR Registration: ACTRN12616000586415.

  

subjective sleep variables such as sleep onset latency (SL), wake after sleep onset (WASO), total sleep time (TST) and sleep efficiency (SE) (Carney et al., 2012). Although this subjective snapshot is essential in insomnia, an objective assessment of sleep could prove a useful clinical adjunct. Moreover, the proliferation of consumer activity trackers measuring sleep means large numbers of individuals track their sleep "objectively" at home. Information on the accuracy and reliability of these devices is severely lacking and widely needed.

Polysomnography (PSG) is considered the reference standard objective measurement of sleep (Taibi, Landis, & Vitiello, 2013). However, PSG does have drawbacks: it is time consuming, requires technical input and patients can find it mildly uncomfortable (Ancoli-Israel et al., 2003; Cook, Prairie, & Plante, 2017; Sadeh, 2015). Actigraphy is a non-invasive method for sleep assessment and it is considered the "accepted" alternative to PSG for measuring sleep in non-laboratory settings. Actigraphy devices are less invasive and expensive than PSG and are particularly useful for tracking objective variations in individuals' night-to-night sleep–wake patterns (Montgomery-Downs, Insana, & Bond, 2012). Comparison between research grade actigraphs and PSG is well documented in the literature, with results typically demonstrating high levels of sensitivity (ability to accurately score sleep; >90%, Marino et al., 2013; Sivertsen et al., 2006; Toon et al., 2016) and low specificity (ability to accurately score wake; <40% Marino et al., 2013; Sivertsen et al., 2006; Toon et al., 2016) in sleep-disturbed populations. Commonly, actigraphy tends to overestimate TST and SE and underestimate SL and WASO in sleep disorders (Marino et al., 2013; Montgomery-Downs et al., 2012; Sivertsen et al., 2006; Taibi et al., 2013). Regardless, actigraphy is the accepted method for measuring sleep objectively outside the laboratory, especially over long periods of time.

In the recent decade, commercial activity trackers have become available at a much lower cost, tracking various statistics, including sleep, calorie intake, exercise and heart rate. Use of these devices has risen within the general public, as individuals are becoming increasingly interested in tracking their own sleep and health. In both 2017 and 2018, the Worldwide Survey of Fitness showed wearable activity trackers were in the top three of worldwide fitness trends (Thompson, 2016, 2017). At the time of writing, one such popular activity tracker is Fitbit. Few studies have compared this device against PSG in a sleep-disturbed population; however, our group recently compared validity of the Fitbit Alta HR (FBA) and the Actiwatch Spectrum Pro (Philips Respironics, AWS) against PSG, showing sensitivity, specificity and other typical validity measures generally do not differ between the two devices during one night in the laboratory when compared to PSG (Kahawage, Jumabhoy, Hamill, Zambotti, & Drummond, 2019).

Although it has become common in clinical practice for patients to provide their sleep data obtained from activity trackers, sleep professionals are still wary of integrating such data into treatment due to lack of validation (Vallières & Morin, 2003). Unreliable or non-valid data could be particularly problematic when they inform treatment decisions. The sleep research field and consumers alike would thus benefit from an inexpensive activity-monitoring device that is able to provide reliable and valid data and can be used in treatment (Montgomery-Downs et al., 2012). Brooks, Friedman, Bliwise, & Yesavage, 1993 suggested actigraphy combined with a subjective measure of sleep could be sensitive enough to detect treatment-based changes in behavioural interventions for insomnia. If a consumer sleep tracker could be utilized in a similar fashion, it would add further clinical value to such devices.

Given the above, there is a need to evaluate the performance of commercial activity trackers in relation to research-grade actigraphs in clinical populations, outside the laboratory and over multiple nights. Outcomes for these studies would help better educate clinicians and the general public on the strengths and limitations of these consumer devices. There are several ways in which one can evaluate the utility of a consumer device, relative to research grade actigraphy. Prior research has typically focused on the discrepancy in measures of sleep between such devices within healthy sleepers by directly comparing devices' sleep outcomes based only on a single-night evaluation. The novelty of our study is in the longitudinal evaluation of the magnitude and stability of FBA–AWS consistency/discrepancy in a sample of patients with insomnia disorder, while recognizing that only PSG can provide the reference standard for direct comparison of performance (de Zambotti, Cellini, Goldstone, Colrain, & Baker, 2019). Specifically, we tested (a) the night-to-night variability in the discrepancies for the main sleep outcomes (TST, SE, WASO and SL) and sensitivity, specificity and accuracy and (b) how these discrepancies vary in response to a direct sleep manipulation (7 weeks of cognitive-behavioural intervention). By taking advantage of the fact that FBA and AWS data were also collected during a PSG night preceding the at-home assessments, we (c) evaluated whether the magnitude of FBA and AWS discrepancies outside the laboratory is similar to those found in the laboratory during a standard PSG and finally (d) examined the clinical utility of information provided by the devices by: (i) assessing whether changes in sleep measured by these devices correlate with subjective clinical symptoms from the insomnia severity index (ISI) and sleep diaries (the main metrics used to evaluate the clinical impact of insomnia) and (ii) examining whether each device can identify change (or lack thereof) in ''insomnia case'' status over time to the same degree as a sleep diary.

## 2 | METHODS

### 2.1 | Participants

Data were collected from participants during the first and last week of a larger randomized clinical trial (Researching Effective Sleep Treatments, CF16/276-2016000125) investigating the effectiveness of insomnia interventions (Mellor et al., 2019) from July 2017 onwards. All participants read explanatory statements and gave informed consent. Participant inclusion criteria included: (a) 18+ years old, (b) diagnosed with DSM-5 insomnia disorder and (c) fluent in English. Exclusion criteria were: (a) diagnosis of

schizophrenia or bipolar disorder, (b) current substance use disorder (SUD) or SUD in the past 90 days, (c) unmanaged sleep disorders (other than insomnia), (d) other conditions such as traumatic brain injury, dementia or stroke that would limit their ability to process and comprehend information, (e) engaged in shift work, (f) currently pregnant (must finish treatment phase before first trimester) or had a newborn (<1 year), (g) current, or a history of, domestic violence, or (h) currently receiving behavioural treatment for insomnia (deemed eligible if they had finished the behavioural treatment >1 month prior). Before starting treatment, those who had travelled across time zones were given an adjustment time of 1 week per time zone travelled. Data were collected from 25 participants ($F$ = 14; 56%), with mean age 50.6 ± 15.9 years; however, four participants are missing from the in-laboratory stay. Participants were reimbursed for their time for wearing both watches ($10 per week) and treatment was free.

## 2.2 | Procedure

Potential participants expressed interest in the study via phone or email and were briefly prescreened for basic eligibility criteria. All participants then underwent more intensive screening, with psychological health assessed via the Structured Clinical Interview for DSM-5 and endorsement of sleep disorders via The Duke Structured Interview for Sleep Disorders (Edinger et al., 2006), along with a battery of questionnaires as part of the larger clinical trial (Mellor et al., 2019). Eligible participants were then scheduled for an in-laboratory overnight sleep study assessment. If participants were found to have a sleep disorder other than insomnia during the in-laboratory assessment, they were excluded from the study and did not continue to the at-home stage. During the at-home stage, participants wore both the FBA and AWS on the same wrist for 1 week at the beginning and the end of treatment. Participants were asked to wear both devices "24/7", with the exception of being in contact with water (e.g., showering or swimming, etc).

### 2.2.1 | In the laboratory

During the one-night laboratory stay, a standard clinical polysomnography was recorded. Both the AWS and FBA were placed on the participant's non-dominant wrist. The order of activity trackers was counterbalanced to account for the possibility that accuracy of the devices may change according to the position on the arm. Before discharge in the morning, participants completed a sleep diary.

### 2.2.2 | At home

Following completion of the overnight laboratory stay, participants were randomized to one of three conditions: (a) individual cognitive behavioural therapy (ICBTI), (b) partner-assisted cognitive behavioural therapy for insomnia (PA-CBTI), or (c) partner-assisted sleep management therapy (PA-SMT). All therapies contained seven 1-hr treatment sessions over a minimum of 7 weeks and a maximum of 12 weeks. Data were collected from the AWS and FBA during the first ("pretreatment") and last ("post-treatment") therapy weeks. During this time, participants were asked to wear the AWS and FBA on their non-dominant wrist (with the order counterbalanced across participants). At pretreatment, participants were not asked to follow a sleep schedule; however, at post-treatment participants were prescribed a sleep schedule for PA-CBTI and ICBTI. To avoid any potential influence on the therapies, participants did not have access to their daily sleep data whilst wearing the devices and did not get any notifications or other information regarding their activities as measured by the devices. They could, however, see their step count and HR across the day. The devices were only synced by the investigators when returned by the participants. Participants also completed a daily sleep diary every morning.

## 2.3 | Materials and measures

### 2.3.1 | Activity trackers

*Research-grade actigraphy*

The Actiwatch Spectrum Pro by Philips Respironics (2013 release) was configured to collect data in 30-s epochs with a sensitive wake threshold (score wake at more than 40 activity counts per epoch) using the Cole-Kripke algorithm (Cole, Kripke, Gruen, Mullaney, & Gillin, 1992). Automatically scored sleep and wake times were manually adjusted based on the completed sleep diary for that week. When necessary, we used the +60 method of adjustment (Straus, Drummond, Nappi, Jenkins, & Norman, 2015), which consisted of manually extending the sleep/wake time beyond the sleep diary-defined time in bed if the actigraph showed obvious signs of sleep immediately before or after sleep diary-defined light off or lights on. In those cases, time in bed was extended until the start or end of the sleep episode but was never extended beyond 60 min. Actigraphy data were double scored by two independent raters and relevant measures were extracted from the software.

*Commercial device*

The commercially available FBA (Fitbit Inc. 2017, version 26.63.2) activity tracker uses a Bluetooth function to sync data to the Fitbit app. Information regarding the algorithms used to infer sleep–wake data is considered proprietary and not disclosed to the general public, but the manufacturer states the algorithm combines heart rate data with movement data to estimate the sleep stage. Epoch length was preset to 30-s intervals and the standard "normal" setting was employed, as opposed to "sensitive". Normal settings were used as (a) they are the default mode and commonly employed by FB users and (b) the few studies comparing "normal" versus "sensitive" settings indicated poor performance of FB devices used in "sensitive" mode (Kahawage et al., 2019). Epoch-by-epoch data for the FBA were not available as output files. Therefore, researchers manually examined

the actogram provided by the Fitbit software/app to extract the sleep/wake stage in 30-s increments. The FBA was aligned with PSG for the single in-laboratory night and to AWS lights off/on for the at-home nights. Twenty percent of the weeks were randomly selected for double scoring and of a total 107,578 epochs across all 116 double-scored nights, there were only eight epochs (0.01%) that were discrepant between scorers.

### 2.3.2 | Sleep measures

*Sleep diary*

A nine-item sleep diary was completed each morning the devices were worn. Participants were asked to record sleep habits such as bedtime, wake time, time in bed, number and duration of awakenings and naps. Weekly averages for sleep efficiency, total sleep time, sleep onset latency, early morning wake time and wake after sleep onset were calculated.

*Insomnia severity index (ISI)*

The ISI is a seven-item self-report questionnaire assessing the nature, impact and severity of insomnia over the past week. Dimensions evaluated include: severity of sleep onset, sleep maintenance, early morning awakening problems, sleep dissatisfaction, interference of sleep difficulties with daytime functioning, noticeability of sleep problems by others, and distress caused by sleep difficulties (Bastien, Vallières, & Morin, 2001).

## 2.4 | Data analysis

Sleep variables were calculated for each activity tracker: (a) total sleep time (TST); (b) sleep onset latency (SL), defined as the number of minutes (from sleep diary-defined lights out bedtime) before the first epoch scored as sleep; (c) sleep efficiency (SE), defined as the ratio of total sleep time compared to total amount of time spent in bed during the night period; and (d) wake after sleep onset (WASO), defined as the amount of wakefulness in minutes after sleep onset.

Discrepancy scores were calculated for each measure as AWS minus FBA (e.g., SE of AWS Night 1 minus SE of FBA Night 1). Thus, a positive value means the AWS overestimates a given variable, relative to the FBA. This calculation was performed for each individual night, both pre- and post-treatment.

Sensitivity, specificity and accuracy between the devices were calculated using a formula generated from Tilmanne, Urbain, Kothare, Wouwer, and Kothare (2009) whereby the AWS was considered the reference for at-home sleep recording; thus, the FBA was compared to the AWS. The formula consisted of a confusion matrix to calculate true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). TPs are considered when both actigraphy and the FBA score the 30-s epoch as sleep. TNs are considered when actigraphy and the FBA score the epoch as awake. FPs are considered when actigraphy scores the epoch as sleep, but the FBA scores the epoch as wake. FNs are considered when actigraphy scores the epoch as wake, but the FBA scores the epoch as sleep. Sensitivity is the capability of the FBA to accurately score sleep and is calculated using TP/(TP + FN). Specificity is the capability of the FBA to accurately score wake and is calculated using TN/(TN + FP). Accuracy is the total percentage of sleep and wake epochs correctly identified by the compared device and is calculated as (TP + TN)/(TP + TN+FN + FP). All calculations were performed for each individual night, both pre- and post-treatment. These comparisons, although not intended as a "validation" of the FBA against the AWS, do allow us to more deeply investigate the nature of the FBA–AWS discrepancies and compliment the analyses comparing discrepancies in sleep parameters.

Changes (Δ) in discrepancy scores were calculated to assess the change in the discrepancy between devices, either from pre- to post-treatment (Aim 2) or from in the laboratory to at home (Aim 3). That is, we wanted to know if the difference in, say, SE between devices pretreatment was the same as the difference between devices in that same measure post-treatment. Change scores were calculated as post-treatment minus pre-treatment (e.g., SE post-treatment Night 1 discrepancy minus SE pre-treatment Night 1 discrepancy) and at-home nights minus in-laboratory night (e.g., SE at-home Night 1 discrepancy minus SE in-laboratory discrepancy).

### Missing data

There was a maximum of 175 nights pretreatment and 175 nights post- treatment. Nights were excluded due to either FBA malfunction/non-compliance or AWS malfunction/non-compliance. At pre-treatment, 22 (12.6%) nights were excluded; all nights were a result of malfunction/non-compliance of the AWS, with seven of those nights also having malfunction/non-compliance for the FBA. At post-treatment, 24 (13.7%) nights were excluded; again, all were a result of malfunction/non-compliance of the AWS device, with 10 of those nights also having FBA malfunction/non-compliance. For the in-laboratory to at-home comparison there were four participants missing from the dataset (due to participants not wearing the devices on the PSG night), resulting in a further 28 nights missing for both pre- and post-treatment for Aim 3.

### Statistics

Data were analysed with the IBM Statistical Package for the Social Sciences (SPSS), version 25. Alpha was set to 0.05, two tailed.

To assess the discrepancy between devices night to night (Aim 1) and pre- to post-intervention (Aim 2), linear mixed models were conducted. The model structure was selected as the dataset was considered to be nested. This allowed analysis to incorporate differences between nights and timepoint. First-level units were nights of data, with a maximum of 14 nights per individual, resulting in 350 units. Second-level units were time (pretreatment and post-treatment). There were 152 nights at the pretreatment time-point and 153 at post-treatment. The third level was the 25 individuals. The model also included a two-way interaction between night and time.

Separate models were run to predict each outcome variable: SE, SL, WASO, TST, sensitivity, specificity and accuracy.

The same linear mixed models were run using the difference scores between the in-laboratory night and at-home nights to assess the night-to-night variability and the effect of time in the differences between in-laboratory and at-home data (Aim 3). First-level units were nights of data, with a maximum of 14 nights per individual, resulting in 264 units. Second-level units were time (pretreatment and post-treatment). There were 133 nights at the pretreatment time-point and 131 at post-treatment. The third-level was the 21 individuals. The model also included a two-way interaction between night and time. Separate models were run to predict each outcome variable: SE, SL, WASO and TST.

To determine if changes in sleep parameters as measured by each device reflected changes in subjective clinical symptoms over time (Aim 4), we conducted two analyses. First, changes in sleep measures over time as assessed by each device were correlated with changes in ISI and the same sleep measures as assessed by sleep diaries. Data were checked for normality using Shapiro-Wilk statistics. Sleep diary SL ($D(25)$, $p < .001$), sleep diary WASO ($D(25)$, $p < .032$), Acti SL ($D(25)$, $p < .005$), Acti SE ($D(25)$, $p < .006$) and Acti WASO ($D(25)$, $p < .022$) were significantly non-normal; however, all other variables were normally distributed ($D(25)$, $p > .100$). Pearson's correlations were used for normally distributed variables to assess the relationship between changes (post- minus pre-treatment) in subjective sleep variables and changes in device-assessed sleep variables. For non-normally distributed sleep variables, Spearman's Rho correlations were performed. Second, we determined whether each device can identify an "insomnia case" as measured with sleep diaries at pretreatment and/or whether the change (or not) in "insomnia case" status over time was reflected by each device. Here, we defined "cases" as SL > 30 min, WASO > 30 (Buysse, Ancoli-Israel, Edinger, Lichstein, & Morin, 2006; Lichstein, Durrence, Taylor, Bush, & Riedel, 2003; Lineberger, Carney, Edinger, & Means, 2006) and SE < 85% (Perlis et al., 2004). For each measure, we first determined whether an individual was classified as a case or not pretreatment (dichotomous yes/no variable) and whether an individual changed case status over time or not (dichotomous yes/no variable) for sleep diaries, the FBA and the AWS, separately. We then conducted sensitivity/specificity analyses on each device for (a) the ability to identify a case pretreatment and (b) the ability to identify a change in case status over time. Each analysis used the sleep diary designation of case status as the reference point, because diaries are the primary tool used both in research and clinical practice when diagnosing insomnia and assessing treatment outcome.

## 3 | RESULTS

### 3.1 | Sleep characteristics

Figure 1 displays descriptive statistics of the mean and standard deviation for each sleep variable at pre- and post-treatment, as well as the discrepancy between devices at each time-point.

### 3.2 | Device discrepancy at home (Aims 1–2)

#### 3.2.1 | Sleep parameters

For all sleep variables except SL, models including night, time (i.e., pre- versus post-treatment) and the interaction between time*night did not improve predictive ability above that of models without predictors (WASO $\chi^2(3) = 1.69$, $p = .64$; TST $\chi^2(3) = 5.93$, $p = .16$; SE $\chi^2(3) = 4.12$, $p = .25$). For SL, the full model did improve predictive ability above the model with no predictors ($\chi^2(3) = 10.48$, $p = .02$) (Table 1). Examination of individual parameters showed discrepancy between devices in SL differed from pre- to post-treatment. Thus, the differences in estimates of sleep variables between devices did not differ either across days or across treatment for WASO, TST or SE. SL estimates, on the other hand, became closer after treatment (small effect size, semi-partial $r^2 = 2\%$).

#### 3.2.2 | Sensitivity, specificity and accuracy

For specificity, the models including night, time and night*time interaction did not improve the predictive ability above that of the model with the predictor ($\chi^2(3) = 1.18$, $p = .76$). For both sensitivity and accuracy, the full models including night, time and the interaction between night*time improved the predictive ability above the models with no predictors (sensitivity $\chi^2(3) = 11.13$, $p = .01$ and accuracy $\chi^2(3) = 24.78$, $p = .0001$) and thus these models were adopted as the final models (Table 2). Examination of the individual parameters in each model showed both sensitivity and accuracy varied across nights, but not over time (small effects size, each semi-partial $r^2 < 2\%$).

### 3.3 | Device discrepancy in the laboratory versus at-home differences (Aim 3)

The discrepancy between devices in the laboratory (AWS–FBA) was as follows: WASO, 1.85 ± 50.77 min; TST, 6.67 ± 55.74 min; SE, 1.42 ± 12.08%; SL, −5.35 ± 34.84 min. Models revealed that neither night nor time significantly predicted change in discrepancy scores from in the laboratory to at home for any sleep parameter estimate from either device. For all sleep variables except SL, full models did not improve predictive ability above that of models without predictors (WASO $\chi^2(3) = 1.5$, $p = .68$; TST $\chi^2(3) = 2.72$, $p = .44$; SE $\chi^2(3) = 1.23$, $p = .75$). For SL, the full model did improve predictive ability ($\chi^2(3) = 16.29$, $p = .001$), but only the parameter for "individual" was significant (Table 3).

### 3.4 | Device change score correlates with clinical symptoms (Aim 4)

#### 3.4.1 | Device correlates with ISI

There were no significant relationships between ISI change scores and the FBA or AWS sleep variable change scores (Table 4). There

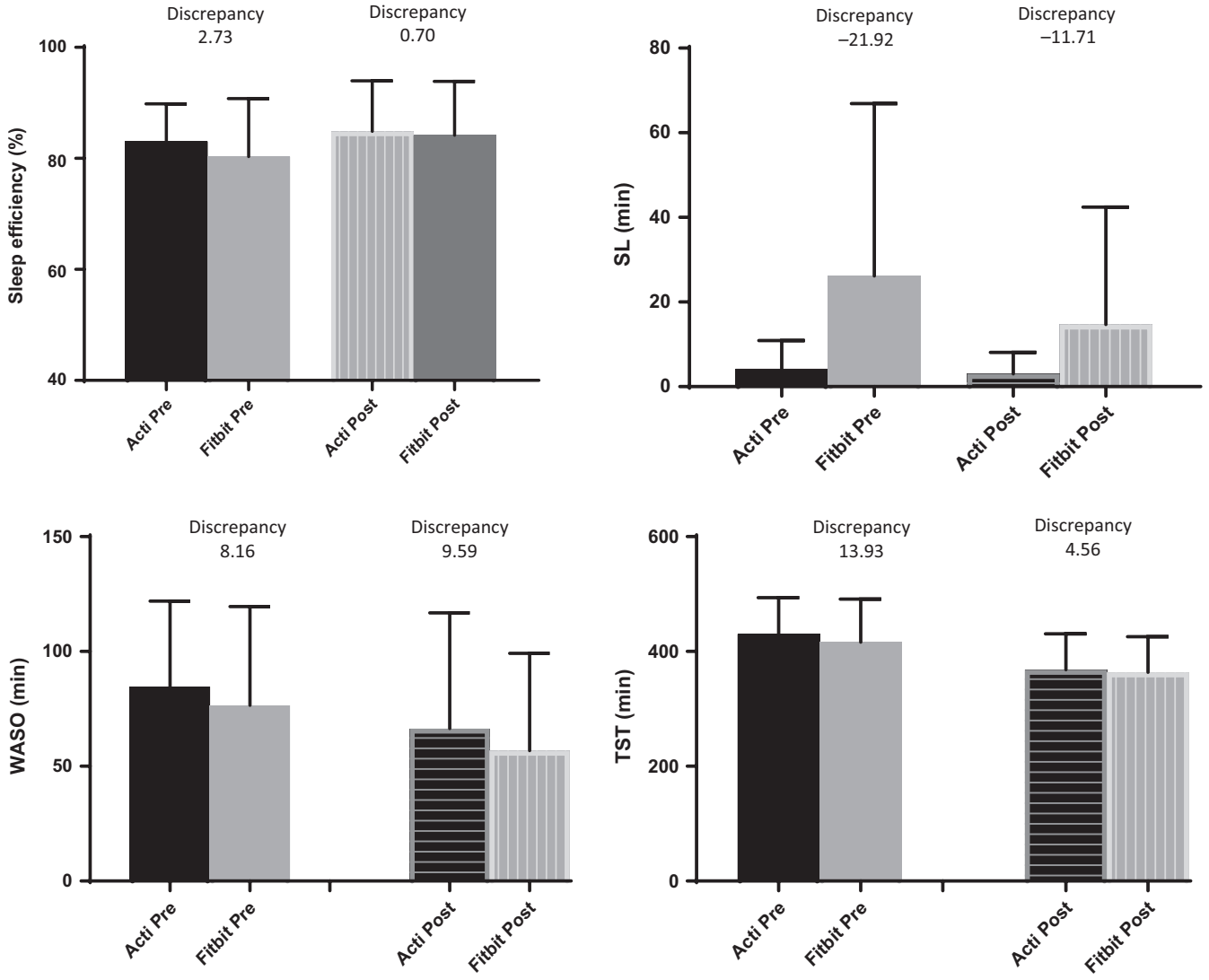


**FIGURE 1** Sleep variables for each device at both pre- and post-treatment. Note: bars represent weekly mean and error bars represent standard deviation. Discrepancy scores calculated as actigraphy minus Fitbit. SL, sleep latency; TST, total sleep time; WASO, wake after sleep onset

was a trend towards a negative correlation between ISI change scores and FBA SE change scores.

### 3.4.2 | Device correlates with sleep diary

There was a significant positive correlation between sleep diary WASO change scores and FBA TST and AWS TST change scores (Table 4). Sleep diary SE change was positively correlated with AWS SE and negatively correlated with AWS WASO. Sleep diary TST change was positively correlated with AWS TST.

### 3.4.3 | Insomnia case status, relative to sleep diaries

At pretreatment, the AWS identified all participants as "insomnia" cases based on WASO, and no participants as cases based on SL. Thus, sensitivity/specificity values showed the AWS was completely unable to discriminate cases from non-cases for those measures. For SE, the AWS showed moderate sensitivity (0.55) and moderate specificity (0.67). At pretreatment, the FBA showed the same WASO result as the AWS. For SL, the FBA showed low sensitivity (0.55) and high specificity (0.86). For SE, the FBA showed high sensitivity (0.86) and moderate specificity (0.67). When assessing the ability to identify change in case status over time, both devices showed low to very low sensitivity and high specificity for WASO and SL, because the majority of cases were identified as not changing status (whether the diaries identified a change or not). The devices diverged for SE. The AWS showed very low sensitivity (0.23) and specificity (0.42), whereas the FBA showed moderate sensitivity (0.69) and specificity (0.67). Thus, the only place the devices showed an even modest ability to identify or track change in insomnia case status was when the FBA measured SE.

**TABLE 1** Predicting device discrepancy

| Parameters | Sleep latency B | (95% CI) | *p* | $r^2$ |
|---|---|---|---|---|
| Fixed effects | | | | |
| Pretreatment[a] | −20.32 | [−36.04, −2.23] | .011 | 0.021 |
| Night | 0.55 | [−2.01, 3.12] | .671 | 0.000 |
| Pretreatment[a] by Night | 2.50 | [−1.09, 6.09] | .172 | 0.001 |
| Random effects | | | | |
| Intercept (client) | 196.02 | [87.98, 436.75] | .014* | |
| Residual | 930.48 | [787.18, 1,099.85] | <.001*** | |
| −2 log likelihood | 2,943.146 | | | |
| AIC | 2,995.146 | | | |
| No. of observations | 305 | | | |
| No. of participants | 25 | | | |

[a]Reference group:post-treatment.AIC, Akaike information criterion.
*$p$ <.05.
***$p$<.001.

**TABLE 3** Predicting change in discrepancy scores at home relative to in the laboratory

| Parameters | Sleep latency B | (95% CI) | *p* | $r^2$ |
|---|---|---|---|---|
| Fixed effects | | | | |
| Pretreatment [a] | −14.194 | [−28.70, 0.315] | .055 | 0.015 |
| Night | 0.994 | [−1.38, 3.37] | .411 | 0.003 |
| Pretreatment [a] by Night | 2.04 | [−1.28, 5.37] | .228 | 0.006 |
| Random effects | | | | |
| Intercept (client) | 511.089 | [258.84, 840.63] | <.001*** | |
| Residual | 703.575 | [588.87, 840.63] | .004** | |
| −2 log likelihood | 2,528.439 | | | |
| AIC | 2,540.439 | | | |
| No. of observations | 264 | | | |
| No. of participants | 21 | | | |

[a]Reference group:post-treatment.
**$p$ <.01.
***$p$<.001.

## 4 | DISCUSSION

This study aimed to compare the commercially available activity-tracking device Fitbit Alta HR (FBA) to the research-grade Actiwatch Spectrum Pro (AWS) in an insomnia sample, over the course of two non-consecutive weeks. To our knowledge, this is the first study to compare these models within this population, and the first to examine the concordance of a consumer device and a research-grade actigraph device over 2 weeks at home. In addition, this is the first study to examine how well a consumer device tracks change in clinically relevant self-report measures with treatment. Overall, we found no significant differences between devices in their estimates of sleep continuity parameters (TST, SE, WASO and SL) over two non-consecutive weeks. Devices provided similar estimates both night to night and over time (pre- to post-treatment) for all sleep parameters, except SL,

**TABLE 2** Predicting change in sensitivity and accuracy of the Fitbit Alta HR (FBA) relative to the Actiwatch Spectrum Pro

| Parameters | Sensitivity B | (95% CI) | *p* | $r^2$ | Accuracy B | (95% CI) | *p* | $r^2$ |
|---|---|---|---|---|---|---|---|---|
| Fixed effects | | | | | | | | |
| Pretreatment [a] | −0.002 | [−0.046, 0.043] | .947 | 1.593 | −0.004 | [−0.046, 0.038] | .849 | 0.000 |
| Night | 0.008 | [0.001, 0.016] | .021 | 0.0187 | 0.007 | [0.000, 0.014] | .042 | 0.015 |
| Pretreatment [a] by Night | −0.005 | [−0.016, 0.005] | .329 | 0.004 | −0.005 | [−0.015, 0.005] | .305 | 0.004 |
| Random effects | | | | | | | | |
| Intercept (client) | 0.001 | [0.001, 0.003] | <.001*** | | 0.003 | [0.001, 0.005] | .<.001*** | |
| Residual | 0.008 | [0.006, 0.009] | .015** | | 0.007 | [0.006, 0.008] | .004*** | |
| −2 log likelihood | −589.882 | | | | −612.487 | | | |
| AIC | −577.882 | | | | −600.487 | | | |
| No. of observations | 305 | | | | 229 | | | |
| No. of participants | 25 | | | | 19 | | | |

[a]Reference group:post-treatment.
**$p$ <.01.
***$p$<.001.

**TABLE 4** Correlations among the sleep diary, the ISI and each device

| | ΔSE | | ΔTST | | ΔSL | | ΔWASO | |
|---|---|---|---|---|---|---|---|---|
| Variable | FBA | AWS | FBA | AWS | FBA | AWS | FBA | AWS |
| ΔISI | −0.382 (0.060) | [a]−0.104 (0.619) | −0.006 (0.978) | 0.185 (0.376) | 0.335 (0.102) | [a].185 (0.376) | 0.246 (0.235) | [a].026 (0.901) |
| ΔSleep diary TST | 0.328 (0.109) | [a]0.214 (0.305) | 0.356 (0.081) | **0.458** (0.021) | −0.284 (0.169) | [a]**0.**111 (0.598) | −0.025 (0.904) | [a] −0.030 (0.887) |
| ΔSleep diary SE | 0.170 (0.416) | [a]**0.577** (0.003) | −0.295 (0.152) | −0.213 (0.307) | 0.002 (0.991) | [a]−0.250 (0.228) | −0.302 (0.142) | [a]**−0.545** (0.004) |
| ΔSleep diary SL | −0.052 (0.804) | [a]−0.332 (0.105) | 0.048 (0.818) | −0.046 (0.827) | 0.249 (0.230) | [a]0.241 (0.246) | −0.006 (0.977) | [a]0.271 (0.190) |
| ΔSleep diary WASO | −0.025 (0.904) | [a]−0.335 (0.112) | **0.435** (0.030) | **0.406** (0.044) | −0.195 (0.351) | [a].301 (0.144) | 0.273 (0.187) | [a]0.324 (0.114) |

*Notes:* Significant correlations printed in bold.
Change in discrepancy scores are presented here as week 6 minus week 1 for all sleep variables for both devices. Results presented are *r* (*p*-value).
Abbreviations: AWS, Actiwatch spectrum Pro; FBA, Fitbit Alta Hr; ISI, Insomnia Severity Index; SE, sleep efficiency; SL, sleep latency; TST, total sleep time; WASO, wake after sleep onset.
[a]Spearman's correlation coefficient.

which showed small, but significant, night-to-night variability in the discrepancy between devices. It is especially noteworthy that device discrepancies did not change over the course of treatment, given TIB decreased post-treatment and sleep schedules were ad-lib pretreatment and largely prescribed by CBT-I post-treatment. This suggests agreement in device measurement is robust to changes in such things as length and regularity of TIB. Studies have demonstrated actigraphy can detect changes in sleep after a clinical intervention (Vallières & Morin, 2003). Thus, if sleep parameter estimates of the FBA relative to the AWS are stable over time, as we found here, it suggests the FBA may be as useful as research-grade actigraphy in aiding the measurement of treatment response (Vallières & Morin, 2003). Such a notion is also supported by the fact changes in sleep as measured by both the FBA and AWS reflected changes in specific self-report measures commonly used in insomnia interventions, although, critically, both devices were generally unable to identify sleep diary-assessed change in clinical case status over time. Finally, if one assumes the AWS is the reference (gold) standard for at-home sleep–wake assessment over time (or perhaps "silver" standard, because PSG is the true gold standard), the FBA showed stable sensitivity (0.89, 0.91 pre- and post-treatment, respectively), specificity (0.62, 0.62 pre- and post-treatment, respectively) and accuracy (0.85, 0.87 pre- and post-treatment, respectively) across time. Specificity was also stable night to night, although both sensitivity and accuracy significantly varied night to night, again to a small but statistically significant extent.

The present study also examined differences between devices for each night at home compared to differences between devices in an overnight sleep study under laboratory conditions. For all measures, the devices showed the same discrepancy outside the laboratory as inside the laboratory. This is critical, because performance in the laboratory could reflect the ideal, well-controlled environment of the laboratory. If the concordance between devices is the same in an at-home setting as in the laboratory, users can have more confidence that findings from in-laboratory validation studies carry over to the home environment. Nonetheless, one must keep in mind that, in the absence of PSG at home, it is impossible to establish which device is more "accurate". We have only established that the discrepancy between devices is stable when moving from in-laboratory to at-home environments.

In addition to investigating whether a device can truly reflect objective sleep–wake patterns, the clinical utility of consumer devices is boosted if they can reflect commonly accepted self-report measures of insomnia. Here, changes in sleep as captured by either device were unable to reflect post-treatment changes in ISI, although FBA-assessed changes in SE came close ($p = .06$). In contrast to the ISI, changes in both FBA and AWS-assessed sleep correlated with changes in subjective sleep assessed by sleep diaries. For example, an increase in sleep diary-assessed TST was reflected in an increase in TST as measured by the AWS. As sleep diary SE increased, this was reflected in SE increases and WASO decreases measured by the AWS. Additionally, decreases in sleep diary WASO over treatment were reflected in decreases in TST as measured by both the AWS and FBA. This is consistent with sleep restriction therapy reducing time in bed, and thus TST, while simultaneously increasing sleep consolidation (Perlis et al., 2004). Although both the FBA and AWS showed some accuracy in reflecting change in diary-based improvement, caution should be taken in interpreting these findings, as we did not control for multiple comparisons in this analysis. Therefore, it is not clear if one device is superior to the other, or even if these devices could be potentially useful in tracking clinically meaningful change in insomnia over the course of treatment.

We also explored the clinical utility of the information provided by the devices by asking how accurately they can identify, based on specific sleep parameters, whether or not an individual would

be classified as an insomnia "case". Here, we used the sleep diary as the standard and defined "cases" based on common clinical criteria. Essentially, the only device and sleep parameter showing evidence of clinical utility in this regard was FBA measurement of SE. Here, the FBA identified 86% of insomnia cases pretreatment and 69% of changes in case status over time using SE, showing moderate specificity at both time-points. This is important information, as clinicians need to know whether consumer device data supplied by their clients can guide diagnostic or treatment decisions. These findings suggest SE as provided by the FBA provides a fairly good reflection of what would have been reported by sleep diaries, especially at baseline. Thus, although by no means diagnostic, a low SE recorded from the FBA can provide one additional piece of data suggesting possible insomnia at baseline and an indication of treatment response at post-treatment. The relatively poor performance of the devices overall in classifying case status in this study may be due to the common objective–subjective sleep discrepancy in insomnia (Perlis et al., 2004).

A few limitations of our study are worth noting. Firstly, for Aim 3, we compared at-home data from several nights to only one night of in-laboratory data (i.e., we subtracted the difference score of the laboratory night from the difference score of several nights of at-home stage data). Thus, we do not know if the discrepancy between devices would differ night to night in the laboratory. However, our at-home data showed no night-to-night variability in discrepancy scores for any measure of sleep, thus the in-laboratory data are likely to show stability. Secondly, findings from Kahawage et al., 2019 demonstrated the mean difference between each device and PSG-recorded SL was <15 min, but accuracy of each device worsens with increasing SL. Here, SL showed a significantly larger discrepancy between devices pretreatment, relative to post-treatment. Therefore, individuals with insomnia, and those treating insomnia disorder, should use either device with caution, as those with severe early insomnia may obtain particularly inaccurate values of the exact variable with which they are most concerned (i.e., SL). In the case of the FBA, SL needs to be taken with an extra dose of caution, especially if the user does not manually indicate lights-out time. Future work would need to determine if using this manual feature improves the detection of SL. Finally, all consumer activity trackers suffer from the proprietary (and non-public) nature of their algorithms, as well as the fact manufacturers can update software without alerting consumers (de Zambotti et al., 2019). Given such a limitation, we examined whether any of our AWS–FBA discrepancy scores varied systematically over the course of the study, under the assumption that such a time effect would implicate changes in FBA firmware. None of our measures varied over the course of the study (all $p$'s > .500). Although this is encouraging, these data should still be generalized beyond the specific model tested here only with extreme caution.

In conclusion, data reported here show the FBA provided similar estimates to a traditional research-grade actigraph in a "real-world" setting. Differences in estimates are not significant between devices night to night, and do not change pre- to post-insomnia intervention (with the exception of SL, where devices become more similar after treatment). Furthermore, changes in sleep measured with both devices reflected changes in commonly utilized self-report measures. However, the devices were generally not good at identifying an insomnia "case" as assessed by sleep diaries (the possible exception being FBA assessment of SE). Overall, this suggests data from the FBA can be used to track sleep, and changes in sleep over a period of time, at least to the same degree as the AWS. The differences between devices are the same at home when compared to an in-laboratory experience. This is important to note as it suggests results from studies validating this device in laboratory settings may be applied at home. Both researchers and health professionals may consider the use of this FBA device as an alternative to the more expensive research-grade actigraphs. Further work is needed in research settings, clinical settings and consumer settings to ensure appropriate data interpretation.

## ACKNOWLEDGEMENTS

The authors thank Mr Luis Mascaro, Mr Elliot Brooker and Mrs Karin Quadros for assistance with actigraphy scoring, Ms Elle Nguyen for assistance with FB scoring and Mr Luke Thomson and Ms Belinda Foote for assistance with PSG scoring. The authors thank all other staff and students at the Monash Sleep and Circadian Medicine Laboratory for their assistance throughout the project.

## CONFLICTS OF INTEREST

The author(s) have nothing to disclose on receipt of financial support for the research, authorship and/or publication of this article. KH, RJ, PK and EMW have no conflicts of interest to declare. SPAD serves on the DSMB for an insomnia trial sponsored by Zelda Therapeutics. MdZ received research funding unrelated to this work from Ebb Therapeutics Inc., Fitbit Inc., International Flavors & Fragrances Inc. and Noctrix Health, Inc. MdZ is an advisor at Snooze, LLC.

## AUTHOR CONTRIBUTIONS

This project was conducted under the supervision of SPAD of the Monash University Sleep and Circadian Medicine Laboratory. KH formulated the hypothesis in collaboration with SPAD, PY and RJ. Recruitment and data collection were conducted by KH and PK as part of a larger study (Project REST (Researching Effective Sleep Treatments; MUHREC approval number: CF16/276-2016000125)) and data were extracted by KH, PK and RJ using both Actigraphy and Fitbit programs. The data were then prepared and analysed by KH, PK and RJ with EMW providing input in the analysis phase. KH prepared the manuscript with input from RJ. The study's aims, hypotheses and statistical analyses were formulated by the authors based on a literature review and consultation with SPAD.

## ORCID

*Kellie Hamill* https://orcid.org/0000-0002-3195-4419

*Sean P. A. Drummond* https://orcid.org/0000-0002-9815-626X

## REFERENCES

Adams, R. J., Appleton, S. L., Taylor, A. W., Gill, T. L., Lang, C., McEvoy, R. D., & Antic, N. A. (2017). Sleep health of Australian adults in 2016: Results of the 2016 Sleep Health Foundation national survey. *Sleep Health*, *3*, 35–42. https://doi.org/10.1016/j.sleh.2016.11.005

Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W., & Pollak, C. P. (2003). The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, *26*, 342–392. https://doi.org/10.1093/sleep/26.3.342

Bastien, C. H., Vallières, A., & Morin, C. M. (2001). Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Medicine*, *2*, 297–307. https://doi.org/10.1016/S1389-9457(00)00065-4

Brooks, J. O. 3rd, Friedman, L., Bliwise, D. L., & Yesavage, J. A. (1993). Use of the wrist actigraph to study insomnia in older adults. *Sleep*, *16*, 151–155. https://doi.org/10.1093/sleep/16.2.151

Buysse, D. J., Ancoli-Israel, S., Edinger, J. D., Lichstein, K. L., & Morin, C. M. (2006). Recommendations for a standard research assessment of insomnia. *Sleep*, *29*, 1155–1173. https://doi.org/10.1093/sleep/29.9.1155

Carney, C. E., Buysse, D. J., Ancoli-Israel, S., Edinger, J. D., Krystal, A. D., Lichsetin, K. L., & Morin, C. M. (2012). The consensus sleep diary: Standardizing prospective sleep self-monitoring. *Sleep*, *35*, 287–302. https://doi.org/10.5665/sleep.1642

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J., & Gillin, J. C. (1992). Automatic sleep/wake identification from wrist activity. *Sleep*, *15*, 461–469. https://doi.org/10.1093/sleep/15.5.461

Cook, J. D., Prairie, M. L., & Plante, D. T. (2017). Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *Journal of Affective Disorder*, *217*, 299–305. https://doi.org/10.1016/j.jad.2017.04.030

de Zambotti, M., Cellini, N., Goldstone, A., Colrain, I. M., & Baker, F. C. (2019). Wearable sleep technology in clinical and research settings. *Medicine & Science in Sports & Exercise*, *51*, 1538–1557. https://doi.org/10.1249/MSS.0000000000001947

Edinger, J. D., Wyatt, J. K., Olsen, M. K., Stechuchak, K. M., Carney, C. E., Chiang, A., … Radtke, R. A. (2006). *Duke structured interview schedule for DSM-IV-TR and international classification of sleep disorders*, 2nd ed. Durham, NC: Veterans Affairs and Duke University Medical Center.

Kahawage, P., Jumabhoy, R., Hamill, K., de Zambotti, M., & Drummond, S. P. A. (2019). Validity and potential clinical utility of a consumer and research-grade activity trackers in insomnia disorder I: In-lab validation against polysomnography. *Journal of Sleep Research*, e12931. https://doi.org/10.111/jsr.12931

Lichstein, K. L., Durrence, H. H., Taylor, D. J., Bush, A. J., & Riedel, B. W. (2003). Quantitative criteria for insomnia. *Behaviour Research and Therapy*, *41*, 427–445. https://doi.org/10.1016/S0005-7967(02)00023-2

Lineberger, M. D., Carney, C. E., Edinger, J. D., & Means, M. K. (2006). Defining insomnia: Quantitative criteria for insomnia severity and frequency. *Sleep*, *29*, 479–485. https://doi.org/10.1093/sleep/29.4.479

Marino, M., Li, Y. I., Rueschman, M. N., Winkelman, J. W., Ellenbogen, J. M., Solet, J. M., … Buxton, O. M. (2013). Measuring sleep: Accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep*, *36*, 1747–1755. https://doi.org/10.5665/sleep.3142

Mellor, A., Hamill, K., Jenkins, M. M., Baucom, D. H., Norton, P. J., & Drummond, S. P. (2019). Partner-assisted cognitive behavioural therapy for insomnia versus cognitive behavioural therapy for insomnia: A randomised controlled trial. *Trials*, *17*, 157.

Montgomery-Downs, H. E., Insana, S. P., & Bond, J. A. (2012). Movement toward a novel activity monitoring device. *Sleep Breath*, *16*, 913–917. https://doi.org/10.1007/s11325-011-0585-y

Perlis, M. L., Smith, M. T., Jungquist, C. R., Nowakowski, S., Orff, H., & Soeffing, J.-B. (2004). Cognitive-behavioral therapy for insomnia. In H. P. Attarian (Ed.), *Clinical handbook of insomnia* (1st edn, pp. 281–296). Totowa, NJ: Humana Press.

Sadeh, A. (2015). III. Sleep assessment methods. Monogr. Soc. Res. *Child Development*, *80*, 33–48. https://doi.org/10.1093/sleep/34.5.601

Sivertsen, B., Omvik, S., Havik, O. E., Pallesen, S., Bjorvatn, B., Nielsen, G. H., … Nordhus, I. H. (2006). A comparison of actigraphy and polysomnography in older adults treated for chronic primary insomnia. *Sleep*, *29*, 1353–1358. https://doi.org/10.1093/sleep/29.10.1353

Straus, L. D., Drummond, S. P. A., Nappi, C. M., Jenkins, M. M., & Norman, S. B. (2015). Sleep variability in military-related PTSD: A comparison to primary insomnia and healthy controls. *Journal of Traumatic Stress*, *28*, 8–16. https://doi.org/10.1002/jts.21982

Taibi, D. M., Landis, C. A., & Vitiello, M. V. (2013). Concordance of polysomnographic and actigraphic measurement of sleep and wake in older women with insomnia. *Journal of Clinical Sleep Medicine*, *9*, 217–225. https://doi.org/10.5664/jcsm.2482

Thompson, W. R. (2016). (2016) Worldwide survey of fitness trends for 2017. *ACSMS Health & Fitness Journal*, *20*, 8–17. https://doi.org/10.1249/FIT.0000000000000252

Thompson, W. R. (2017). Worldwide survey of fitness trends for 2018: The CREP Edition. *ACSMS Health & Fitness Journal*, *21*, 10–19. https://doi.org/10.1249/FIT.0000000000000341

Tilmanne, J., Urbain, J., Kothare, M. V., Wouwer, A. V., & Kothare, S. V. (2009). Algorithms for sleep-wake identification using actigraphy: A comparative study and new results. *Journal of Sleep Research*, *18*, 85–98. https://doi.org/10.1111/j.1365-2869.2008.00706.x

Toon, E., Davey, M. J., Hollis, S. L., Nixon, G. M., Horne, R. S. C., & Biggs, S. N. (2016). Comparison of commercial wrist-based and smartphone accelerometers, actigraphy, and PSG in a clinical cohort of children and adolescents. *Journal of Clinical Sleep Medicine*, *12*, 343–350. https://doi.org/10.5664/jcsm.5580

Vallières, A., & Morin, C. M. (2003). Actigraphy in the assessment of insomnia. *Sleep*, *26*, 902–906. https://doi.org/10.1093/sleep/26.7.902

**How to cite this article:** Hamill K, Jumabhoy R, Kahawage P, de Zambotti M, Walters EM, Drummond SPA. Validity, potential clinical utility and comparison of a consumer activity tracker and a research-grade activity tracker in insomnia disorder II: Outside the laboratory. *J Sleep Res*. 2020;29:e12944. https://doi.org/10.1111/jsr.12944