

Alternatives to polysomnography (PSG): A validation of wrist actigraphy and a partial-PSG system

Anastasi Kosmadopoulos · Charli Sargent ·
David Darwent · Xuan Zhou · Gregory D. Roach

© Psychonomic Society, Inc. 2014

Abstract The objective of this study was to assess the validity of a sleep/wake activity monitor, an energy expenditure activity monitor, and a partial-polysomnography system at measuring sleep and wake under identical conditions. Secondary aims were to evaluate the sleep/wake thresholds for each activity monitor and to compare the three devices. To achieve these aims, two nights of sleep were recorded simultaneously with polysomnography (PSG), two activity monitors, and a partial-PSG system in a sleep laboratory. Agreement with PSG was evaluated epoch by epoch and with summary measures including total sleep time (TST) and wake after sleep onset (WASO). All of the devices had high agreement rates for identifying sleep and wake, but the partial-PSG system was the best, with an agreement of $91.6 \% \pm 5.1 \%$. At their best thresholds, the sleep/wake monitor (medium threshold, $87.7 \% \pm 7.6 \%$) and the energy expenditure monitor (very low threshold, $86.8 \% \pm 8.6 \%$) had similarly high rates of agreement. The summary measures were similar to those determined by PSG, but the partial-PSG system provided the most consistent estimates. Although the partial-PSG system was the most accurate device, both activity monitors were also valid for sleep estimation, provided that appropriate thresholds were selected. Each device has advantages, so the primary

consideration for researchers will be to determine which best suits a given research design.

Keywords Actigraphy · Accelerometry · Wireless sleep monitoring

In laboratory-based sleep research, polysomnography (PSG) is the gold standard for measuring sleep. PSG provides detailed information about sleep architecture, sleep duration, and sleep quality. However, in field-based settings, this technique is expensive, time-consuming, and impractical, because it requires the attendance of a sleep technologist to set up equipment and apply electrodes to multiple sites on the face and scalp (Van de Water, Holmes, & Hurley, 2011). Therefore, a variety of alternative options for objective sleep measurement have been developed. The most frequently employed alternative in the field involves wrist actigraphy (Van de Water et al., 2011). Other alternatives derive sleep stage estimates by approximating PSG and simplifying the means of recording electrophysiological brain signals (Shambroom, Fabregas, & Johnstone, 2012).

Wrist actigraphy is a method of recording rest and activity patterns using wristwatch-like devices known as *activity monitors*. The advantage of wrist actigraphy over PSG in the field is that it is simple, noninvasive, unobtrusive, and can be measured for weeks with less inconvenience to participants. Activity monitors contain accelerometers that sample and record movement at regular intervals (Van de Water et al., 2011). Validated algorithms can then be used to calculate sleep/wake estimates on the basis of the amount of activity recorded (low activity being associated with sleep, and high activity being associated with wakefulness; Cole, Kripke, Gruen, Mullaney, & Gillin, 1992; Sadeh, Sharkey, & Carskadon, 1994; Van de Water et al., 2011). Threshold-based algorithms estimate sleep and wakefulness on the basis

A. Kosmadopoulos (✉) · C. Sargent · D. Darwent · X. Zhou ·
G. D. Roach
Appleton Institute, Central Queensland University, P.O. Box 42,
Goodwood, South Australia, Australia 5034
e-mail: a.kosmadopoulos@cqu.edu.au

A. Kosmadopoulos
Bushfire Co-operative Research Centre, East Melbourne, Victoria,
Australia

of whether activity scores, calculated for each epoch (usually 30 or 60 s) during the monitoring period, fall above or below an activity threshold that is set post hoc. Epochs with scores above this threshold during sleep periods are classified as “wake” (Kushida et al., 2001; Paquet, Kawinska, & Carrier, 2007). Lowering the threshold increases their ability to detect wakefulness but compromises the ability to detect sleep (Hyde et al., 2007; Kushida et al., 2001; Signal, Gale, & Gander, 2005). Although activity monitors have an acknowledged tendency to overestimate sleep by misidentifying quiescent wakefulness as sleep, this has not inhibited their use in many studies with different populations (Darwent, Roach, & Dawson, 2012; Paterson, Dorrian, Pincombe, Grech, & Dawson, 2010; Roach, Petrilli, Dawson, & Lamond, 2012; Roach, Sargent, Darwent, & Dawson, 2012; Sargent, Halson, & Roach, 2012; Tremaine et al., 2013), and the bias has generally been deemed minimal, particularly when they are used in conjunction with sleep diaries (Carney, Lajos, & Waters, 2004; de Souza et al., 2003; Jean-Louis, Kripke, Mason, Elliott, & Youngstedt, 2001; Kanady, Drummond, & Mednick, 2011; Kushida et al., 2001; Paquet et al., 2007; Reid & Dawson, 1999; Rupp & Balkin, 2011; Signal et al., 2005).

Activity monitors are not solely developed for measuring sleep. For example, some devices are designed to produce estimates of energy expenditure from waist actigraphy, using different algorithms (Heil, 2006). Nonetheless, the component technology (i.e., accelerometers) and theoretical basis of these devices (i.e., that movement is positively associated with energy expenditure) are similar to those of activity monitors developed for sleep (Evans et al., 2011). The potential for these devices to produce concurrent estimates of both energy expenditure and sleep/wake patterns has recently been recognized (Evans et al., 2011; Weiss, Johnson, Berger, & Redline, 2010). Recent validation studies have supported the use of energy expenditure monitors to also measure sleep and wake (Evans et al., 2011; Galland, Kennedy, Mitchell, & Taylor, 2012; Kosmadopoulos, Sargent, Zhou, Darwent, & Roach, 2012; Robillard, Lambert, & Rogers, 2012; Weiss et al., 2010). However, for those studies that estimated sleep and wake with energy expenditure monitors using threshold-based algorithms designed for sleep/wake monitors, these recommendations have been conditional on selecting a threshold lower than the default for sleep/wake monitors (Kosmadopoulos et al., 2012; Weiss et al., 2010). Increased wake detection and better overall performance at a low threshold, relative to the default medium threshold, has led to speculation that an additional threshold reduction may improve performance further (Kosmadopoulos et al., 2012). However, the extent to which this may inhibit sleep detection is not clear, given that the increased wake detection is at the expense of sleep detection.

Since movement is only an indirect gauge of sleep and wakefulness, actigraphy is limited to calculating summary

measures, such as “total sleep time” and “wake after sleep onset.” Therefore, other alternatives to PSG derive sleep-stage information from physiological data obtained using less expensive and time-consuming methods (Shambroom et al., 2012; Van de Water et al., 2011). Included among these are simplified approximations of PSG that calculate sleep/wake estimates from an amalgamation of electrophysiological signals in the forehead, taking the form of headbands. So far, this form of partial-PSG has been found to agree well with PSG in terms of detecting sleep and differentiating sleep stages (Griessenberger, Heib, Kunz, Hoedlmoser, & Schabus, 2013; Shambroom et al., 2012). When compared to results derived from sleep/wake monitors, this method seems better able to identify sleep and wake, and its estimations of sleep and wake are more frequently confirmed by PSG (Shambroom et al., 2012). However, selecting sleep-monitoring devices and associated activity thresholds requires knowledge of how well these systems compare against PSG under the same conditions. There has been no systematic evaluation of sleep/wake monitors and energy expenditure monitors at different sleep/wake thresholds, together in a single study. Similarly, the self-administered partial-PSG system has only been compared with sleep/wake monitors at their default threshold, and not with the energy expenditure monitor. Therefore, the aim of this study was to address these gaps by concurrently evaluating their capacities to distinguish sleep from wake, epoch by epoch, and in terms of total sleep time (TST), sleep efficiency, wake after sleep onset (WASO), and sleep onset latency (SOL).

Method and materials

Participants

The participants were 22 young adults (18 male, four female) who responded to advertisements placed online and on public noticeboards in Adelaide, South Australia. They had a mean (\pm SD) age of 23.9 (\pm 3.8) years and a body mass index of 22.4 (\pm 2.1) kg/m². Participants were screened by interview and general health questionnaire. The criteria for inclusion were habitual bedtimes between 22:00 and 00:00 h, sleep durations of 7–9 h per night, and a body mass index of 20–25 kg/m². Volunteers were excluded if they reported medical problems, sleep difficulties, smoking, regular excessive consumption of alcohol or caffeine, or transmeridian travel/shift work in the previous month.

This study was approved by the CQUniversity Human Research Ethics Committee following the guidelines of the National Health and Medical Research Council of Australia. Participants provided written informed consent and were given an honorarium for their involvement.

Procedure

Participants attended the sleep laboratory on two occasions, at least one week apart. On each occasion, participants were provided a 9.5-h sleep opportunity (22:00–07:30 h) in individual bedrooms. During this sleep period, each participant wore two activity monitors on the wrist and a wireless sleep-monitoring system on the forehead, and had electrodes attached to the face and scalp for PSG. All sleep-monitoring devices, including the PSG system, stored data in 30-s epochs and were time-synchronized to the same computer clock prior to each sleep period, so that corresponding epochs could be aligned (Kushida et al., 2001). Twenty participants attended the laboratory for two nonconsecutive night sleeps, and two participants attended for a single night sleep. In all, 42 sleep periods were assessed.

Polysomnography

The gold standard measure of sleep was PSG, conducted using the Siesta Portable EEG system (Compumedics, Melbourne, Victoria, Australia) and Grass gold-cup electrode leads (Astro-Med, Inc., West Warwick, RI). Brain activity was monitored with two channels of electroencephalography (C3–A2, C4–A1); eye movements were monitored with right and left electro-oculograms, and muscle tone was measured with two channels of chin electromyography. Sleep/wake stages were manually scored from PSG recordings by an experienced sleep technician, in 30-s epochs, following the Rechtschaffen and Kales (R&K) criteria (Rechtschaffen & Kales, 1968).

Activity monitors

The sleep/wake monitor used in this study was the Actiwatch-64 (Mini-Mitter Philips Respironics, Bend, OR), and the energy expenditure monitor was the Actical Z-series (Mini-Mitter Philips Respironics, Inc.). Devices were worn on the nondominant wrist, with the Actiwatch being placed closer to the hand. The activity monitors contain an omnidirectional piezoelectric accelerometer and sample movement at 32 Hz. The number of activity counts recorded each second reflects the intensity of movement at that time, with larger counts being indicative of increased movement. The total activity count for each user-defined epoch duration is calculated by summing the counts recorded during that epoch. Activity counts for the Actical are consistently lower than those for the Actiwatch, since it records the mean count per second, whereas the latter records the peak.

Sleep/wake estimation The sleep/wake monitor's accompanying software (Actiware version 3.4; Mini-Mitter Philips Respironics, Inc.) was used to derive sleep and wake estimates for all epochs recorded using both activity monitors. This software employs a validated algorithm to generate weighted

scores for each 30-s epoch on the basis of its activity count and the counts in the surrounding 2 min (Kushida et al., 2001; Oakley, 1997). The equation applied to each epoch was as follows:

$$A = 0.04E_{-4} + 0.04E_{-3} + 0.2E_{-2} + 0.2E_{-1} + 2E_0 + 0.2E_{+1} + 0.2E_{+2} + 0.04E_{+3} + 0.04E_{+4},$$

where A is the weighted activity score derived from counts of the current and surrounding epochs, E_0 is the activity count recorded during the current epoch, E_{-n} are the activity counts recorded during preceding epochs, and E_{+n} are the activity counts recorded during following epochs. Epochs during sleep periods are classified as wake if their weighted scores fall above the specified sleep/wake threshold. Thresholds for this study included three preset thresholds (low = 20, medium = 40, and high = 80) and a customized threshold (very low = 10).

Partial-PSG system

The partial-PSG system used in this study was the Zeo (Newton, MA). The Zeo is a portable, automatic sleep-staging tool that includes a headband containing three dry electrodes, which record and transmit electrophysiological signals from the forehead, and a base station that wirelessly receives these signals for analysis (Shambroom et al., 2012). The signals recorded via the headband approximate an amalgamation of the brain activity, eye movement, and muscle tone recorded with PSG; the Zeo automatically scores these signals into sleep and wake stages using proprietary algorithms based on the R&K criteria (Shambroom et al., 2012). The sleep stages reported by the Zeo system are condensed, such that Stage 1 and 2 sleep are reported together as "light sleep," and Stage 3 and 4 sleep are reported together as "deep sleep." Wake and REM sleep are reported separately according to the standard definitions. Staff placed the partial-PSG system on participants' foreheads prior to the start of the sleep period.

Epoch-by-epoch comparisons

After being aligned with PSG, corresponding epochs from all of the devices at all thresholds were classified into one of four categories—True Sleep (TS), False Sleep (FS), True Wake (TW), and False Wake (FW)—on the basis of their agreement with PSG (see Table 1).

For all three of the devices, five statistical measures of epoch-by-epoch concordance were then calculated from the number of epochs in each category:

- $Agreement = [(TS + TW)/(TS + FW + TW + FS)] * 100 =$ percentage of all sleep and wake epochs correctly detected by the device;

Table 1 Sleep/wake agreement matrix

Device	Polysomnography	
	Sleep	Wake
Sleep	True Sleep (TS)	False Sleep (FS)
Wake	False Wake (FW)	True Wake (TW)

- *Sensitivity* = $[TS/(TS + FW)] * 100$ = percentage of sleep epochs correctly detected by the device;
- *Specificity* = $[TW/(TW + FS)] * 100$ = percentage of wake epochs correctly detected by the device;
- *Positive predictive value (PPV)* = $[TS/(TS + FS)] * 100$ = percentage of epochs correctly estimated by the device to be sleep; and,
- *Negative predictive value (NPV)* = $[TW/(TW + FW)] * 100$ = percentage of epochs correctly estimated by the device to be wake.

Agreement, sensitivity, and specificity indicate the proportions of true sleep and wake epochs (as determined by PSG) with which a device agrees. In contrast, PPV or NPV indicates the likelihood that an estimation of sleep or wake, respectively, using an alternative to PSG is true.

Cohen's kappa Cohen's kappa (κ) was calculated on the basis of epoch-by-epoch device comparisons, to evaluate agreement with PSG beyond what could be expected by chance alone (Sim & Wright, 2005). Agreement was interpreted against the guidelines of Landis and Koch (1977), where 0–.20 indicates *slight agreement*, .21–.40 is *fair*, .41–.60 is *moderate*, .61–.80 is *substantial*, and .81–1.0 is *almost perfect*.

Summary sleep measures

Four summary sleep measures were calculated from the outputs of PSG, the partial-PSG, and both activity monitors:

- *Total sleep time (TST)* = the number of minutes asleep in bed;
- *Sleep efficiency* = the percentage of time asleep between lights off and lights on;
- *Wake after sleep onset (WASO)* = the number of minutes spent awake after sleep onset and before final awakening; and,
- *Sleep onset latency (SOL)* = the number of minutes between lights off and sleep onset.

For PSG, sleep onset was defined as the first of three consecutive epochs of Stage 1 sleep, or one epoch of any other stage of sleep (i.e., Stages 2, 3, 4, or REM; Carskadon et al., 1986). Since the partial-PSG system collapses Stages 1

and 2 together, sleep onset for this device was defined as the first of two consecutive epochs of light sleep, or one epoch of deep sleep or REM sleep. Actiware sleep software calculated sleep onset for activity monitors as the first epoch of 10 min of immobility (CamNTEch Ltd., 2008).

Statistical tests of agreement were used to determine how accurate the devices were at estimating the summary measures. Since the four summary measures were not normally distributed, Wilcoxon signed-rank tests were conducted to identify which devices and/or thresholds significantly differed from PSG on each of the four measures.

Bland–Altman technique The Bland–Altman technique is a way of comparing alternative measurement methods with a gold standard by plotting the differences between them and analyzing the distribution of differences (Bland & Altman, 1986). The differences from TST and WASO, as determined by PSG, were plotted for the partial-PSG system and both activity monitors at their optimal thresholds (as determined by post hoc analyses). TST and WASO were chosen because they are measures that separately describe large proportions of sleep and wake during time in bed and allow for a clear comparison of devices against PSG. Differences in TST and WASO, between devices and PSG, were calculated such that a positive score indicated an overestimation by the device/threshold, and a negative score indicated an underestimation. Pearson correlations were calculated to determine the consistency of differences with PSG—that is, whether there was a relationship between TST or WASO duration and how much each device differed from PSG.

Results

Of the 42 nights recorded, one night each of PSG and sleep/wake monitor data and two nights of data for the wireless system had to be excluded from the analysis due to technical failure. The results from the remaining 38 complete data sets are reported here.

Epoch-by-epoch comparisons

When compared to PSG epoch by epoch, the sleep/wake monitor correctly identified 83.5 % \pm 6.0 % to 88.0 % \pm 9.0 % of epochs as sleep or wake, depending on the threshold employed (Table 2). Although utilizing the high threshold for the sleep/wake monitor accurately identified the most sleep epochs overall, it had the worst specificity and positive predictive value (PPV) relative to the other thresholds for this device; that is, it misidentified the highest proportion of wake epochs as sleep. The customized very low threshold resulted in the best specificity and the highest PPV for the sleep/wake monitor, but had the worst overall agreement, sensitivity to sleep epochs, and negative predictive value (NPV). In contrast

Table 2 Mean epoch-by-epoch sleep/wake percentage agreement with polysomnography for each device and activity threshold (with standard deviations in parentheses)

Device	Agreement (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Partial PSG	91.6 (5.1)	97.4 (4.5)	42.2 (23.5)	93.1 (4.4)	79.6 (23.6)
Sleep/Wake Monitor					
Very low	83.5 (6.0)	87.6 (5.6)	61.5 (16.2)	92.8 (7.9)	39.2 (18.1)
Low	86.3 (6.6)	92.5 (4.2)	49.7 (15.5)	91.6 (8.7)	46.4 (19.2)
Medium	87.7 (7.6)	95.8 (2.7)	37.7 (14.5)	90.4 (9.3)	52.8 (19.4)
High	88.0 (9.0)	97.8 (1.6)	26.9 (13.5)	89.3 (10.1)	59.1 (20.4)
Energy Expenditure Monitor					
Very low	86.8 (8.6)	95.0 (4.4)	36.3 (16.3)	90.1 (9.9)	50.9 (20.3)
Low	87.6 (9.1)	97.1 (2.9)	28.1 (14.1)	89.3 (10.1)	58.7 (21.8)
Medium	87.8 (9.7)	98.5 (1.8)	19.5 (11.3)	88.6 (10.4)	64.5 (21.9)
High	87.6 (10.3)	99.2 (1.1)	12.4 (8.6)	88.0 (10.6)	69.4 (22.3)

$N=38$, *PPV* positive predictive value, *NPV* negative predictive value

to these extremes, the medium and low thresholds both maintained a high overall agreement, identified more wake epochs than the high threshold and identified more sleep epochs than the very low threshold. The medium threshold was more accurate than the low threshold.

In contrast to the sleep/wake monitor, altering the thresholds did not affect the overall agreement of the energy expenditure monitor. However, this did influence its sensitivity, specificity, PPV, and NPV. The high and medium thresholds were the most sensitive to sleep, but this was at the expense of identifying epochs of wake. In contrast, the low and very low thresholds were slightly less sensitive but had much greater specificity. Of these two thresholds, the very low threshold had the greater specificity and the best PPV.

On average, the partial-PSG system did not define $9 (\pm 40.9)$ epochs with a sleep or wake stage per night. Of these undefined epochs, 96.3 % were aligned with an epoch classified as sleep by PSG. Despite this, it correctly identified 91.6 % of epochs as either sleep or wake (Table 2) and had a sensitivity, specificity, PPV, and NPV that were higher than for the sleep/wake monitor at a medium threshold, and than for the energy expenditure monitor at a very low threshold.

Cohen's kappa Following the criteria of Landis and Koch (1977), Cohen's kappa coefficients for the activity monitors described slight to fair sleep/wake epoch agreement with PSG beyond what could be expected by chance (Table 3). For the partial-PSG system, it described a moderate agreement. The results are consistent with the overall agreement of the devices.

Summary sleep measure comparisons

Mean ($\pm SD$) sleep and wake durations determined by each monitoring device are summarized in Table 4. For the activity monitors, higher thresholds for wake resulted in longer TST,

decreased sleep efficiency, and less WASO (Table 4). Since SOL for activity monitors is calculated as the duration from bedtime until 10 min of consecutive immobility, it remained unchanged, irrespective of the sleep/wake threshold. Estimations of TST, sleep efficiency, and WASO using the partial-PSG system were closest to the estimates of the sleep/wake monitor and the energy expenditure monitor when they were calculated using the medium and very low thresholds, respectively.

Significant differences from PSG, as determined by Wilcoxon signed-rank tests (Bonferroni correction, $p < .05$), are indicated in Table 4. The SOL as determined by PSG was significantly longer than the SOLs estimated by the three alternative devices. Even where summary measure estimates were not significantly different from PSG, the standard deviations for each device and threshold were large.

Table 3 Mean Cohen's kappa for each device and activity threshold (with standard deviations in parentheses)

Device	Cohen's κ
Partial PSG	.47 (.24)
Sleep/Wake Monitor	
Very low	.35 (.12)
Low	.37 (.12)
Medium	.35 (.12)
High	.30 (.13)
Energy Expenditure Monitor	
Very low	.31 (.12)
Low	.29 (.12)
Medium	.24 (.12)
High	.17 (.11)

$N=38$

Table 4 Mean summary sleep statistics for each device and activity threshold (with standard deviations in parentheses)

Device	Total Sleep Time (min)	Sleep Efficiency (%)	Wake After Sleep Onset (min)	Sleep Onset Latency (min)
PSG	496.2 (61.8)	86.9 (10.8)	41.3 (47.1)	27.1 (28.5)
Partial PSG	519.0 (61.5)*	90.7 (10.8)*	36.0 (58.6)	11.0 (11.0)*
Sleep/Wake Monitor				
Very low	461.1 (42.7)*	80.9 (7.5)*	92.9 (36.2)*	11.1 (12.3)*
Low	492.0 (36.5)	86.3 (6.4)	62.0 (29.1)*	11.1 (12.3)*
Medium	514.3 (30.1)	90.3 (5.3)	39.7 (21.5)	11.1 (12.3)*
High	530.2 (23.7)*	93.1 (4.1)*	23.7 (13.3)	11.1 (12.3)*
Energy Expenditure Monitor				
Very low	519.2 (31.7)	91.1 (5.5)	43.2 (28.2)	5.1 (7.6)*
Low	533.9 (24.6)*	93.7 (4.3)*	28.4 (20.6)	5.1 (7.6)*
Medium	545.2 (17.8)*	95.7 (3.1)*	17.2 (13.7)*	5.1 (7.6)*
High	552.6 (13.8)*	97.0 (2.4)*	9.7 (8.9)*	5.1 (7.6)*

$N = 38$

*Significant difference from PSG (Bonferroni correction, $p < .05$)

At low and medium thresholds, we found no significant differences from PSG in terms of TST and sleep efficiency for the sleep/wake monitor (Table 4). The sleep/wake monitor also did not differ from PSG in terms of WASO at medium and high thresholds. For the energy expenditure monitor, low and very low thresholds did not result in significant differences from PSG in WASO, and the very low threshold did not result in significant differences from PSG in TST and sleep efficiency. The partial-PSG system was not significantly different from PSG regarding WASO, but it did significantly overestimate TST and sleep efficiency.

Bland–Altman plots The distributions of differences in TST and WASO from PSG are depicted, for all devices, in Fig. 1. Plots for each activity monitor were created using its least biased threshold—that is, the threshold that resulted in the fewest significant mean differences in TST and WASO from PSG (Table 4). The medium threshold had the fewest significant differences for the sleep/wake monitor, and the very low threshold had the fewest significant differences for the energy expenditure monitor.

The partial-PSG system overestimated TST by a mean of 22.7 (± 33.7)min, and underestimated WASO by a mean of 5.4 (± 47.9)min (Fig. 1). In contrast, the sleep/wake and energy expenditure monitors overestimated TST, on average, by 18 (± 51.9)min and 22.9 (± 56.9)min, and underestimated WASO on average by 1.6 (± 42.4)min and 1.9 (± 50.0) min, respectively. The distribution of differences from PSG varied greatly for all devices and thresholds (Fig. 1). However, the partial-PSG system's differences from PSG were more consistent than those for the activity monitors. For partial-PSG, no significant ($p > .10$, two-tailed) correlations emerged between the size of the difference from PSG and either TST ($r < .01$) or

WASO ($r < .26$) duration (Fig. 1). In contrast, for both activity monitors, differences from PSG were significantly ($p < .05$) negatively correlated with the amount of TST (sleep/wake monitor, $r = -.69$; energy expenditure monitor, $r = -.62$) and WASO (sleep/wake monitor, $r = -.69$; energy expenditure monitor, $r = -.48$) obtained. That is, they progressively overestimated sleep for shorter durations of TST, and underestimated wake for longer durations of WASO.

Discussion

In the present study, the abilities of three devices to distinguish sleep from wake were evaluated in healthy young adults during night-time sleep periods. Our results confirmed previous findings that sleep/wake monitors, energy expenditure monitors, and partial-PSG systems have high predictive value and agreement with PSG regarding sleep, but have relatively poor predictive value and agreement with PSG regarding wake (de Souza et al., 2003; Galland et al., 2012; Griessenberger et al., 2013; Kosmadopoulos et al., 2012; Kushida et al., 2001; Paquet et al., 2007; Robillard et al., 2012; Shambroom et al., 2012; Signal et al., 2005). Our results also confirmed the large variability in differences from PSG previously found for individual sleep/wake estimates. Since these alternatives to PSG were all used simultaneously in our study, we were also able to compare the sleep/wake monitor, the energy expenditure monitor, and the portable partial-PSG system under the same conditions.

Evaluation and comparison of activity monitors

Due to the different methods used by the activity monitors to record activity counts (peak vs. mean count/s), counts per

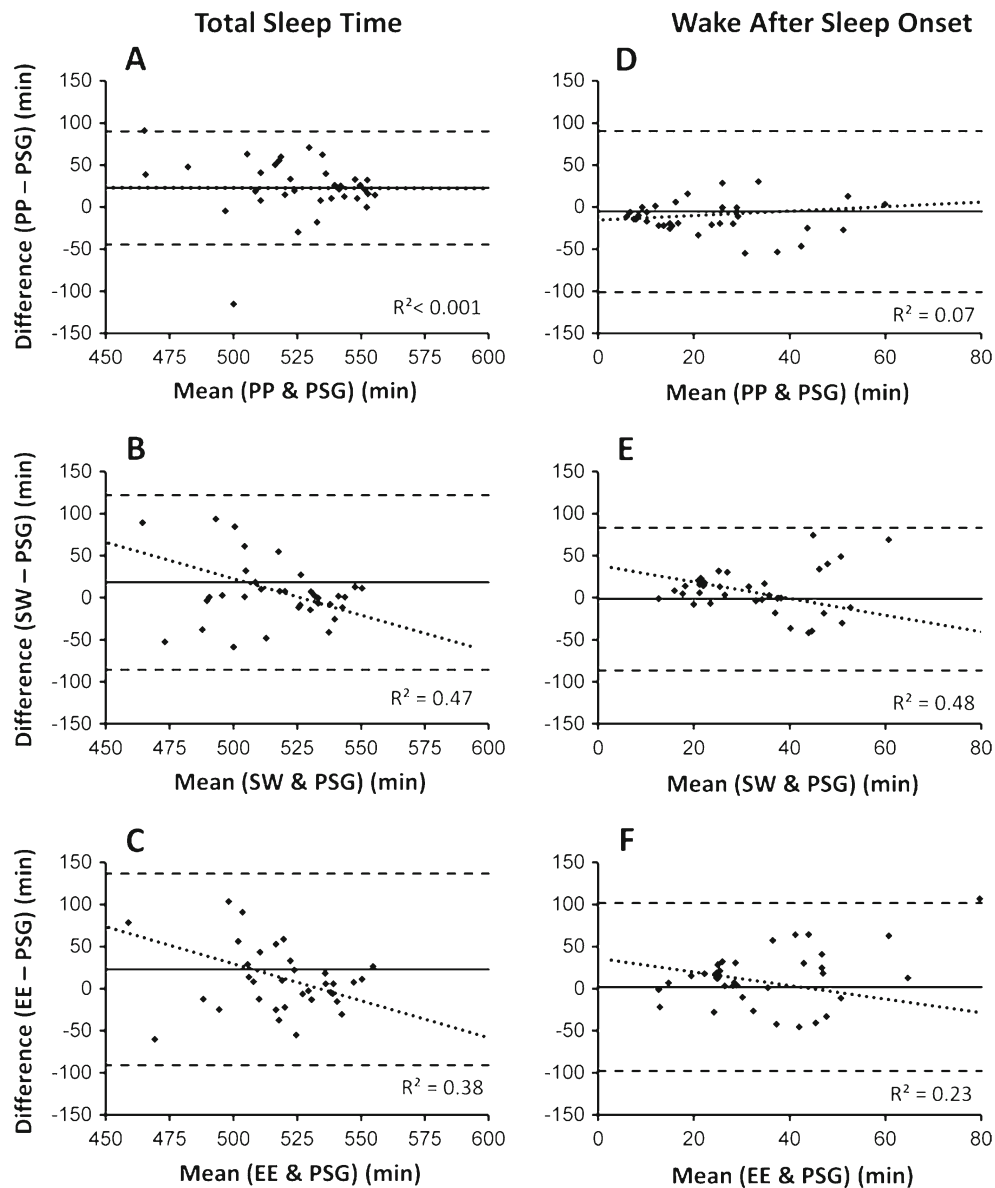


Fig. 1 Bland–Altman plots of TST (panels A–C) and WASO (panels D–F), comparing PSG with (1) the partial-PSG (PP) system (A and D), (2) the sleep/wake (SW) monitor at the medium threshold (B and E), and (3) the energy expenditure (EE) monitor at the very low threshold (C and F). The x-axes indicate the average TST or WASO of each device and PSG. The y-axes describe the difference between each device and PSG in minutes, such that positive values indicate an overestimation of TST or

WASO by the device relative to PSG, and negative values indicate an underestimation of the device relative to PSG. Solid horizontal lines indicate the mean bias from PSG, and broken horizontal lines indicate the limits of agreement (± 2 SDs). Dotted lines and R^2 values indicate the slope and strength of the relationship between TST or WASO and the difference from PSG

epoch for the sleep/wake monitor are higher than for the energy expenditure monitor. Consequently, at each threshold the former consistently detected more wake than the latter. However, consistent with previous validations, higher thresholds captured more sleep and lower thresholds captured more wake, irrespective of the monitor used (Kosmadopoulos et al., 2012; Kushida et al., 2001; Signal et al., 2005). On the basis of the epoch-by-epoch statistics and the comparisons of summary sleep measures with PSG, the default medium threshold for the sleep/wake monitor was found to have the optimum balance of

agreement with PSG, relative to higher and lower thresholds. For the energy expenditure monitor, the optimum threshold was very low, suggesting that the preset low threshold previously used when validating it may not have fully reflected its capacity for sleep measurement (Kosmadopoulos et al., 2012; Weiss et al., 2010). At their optimum thresholds, the sleep/wake monitor performed only marginally better than did the energy expenditure monitor in terms of epoch-by-epoch comparisons and summary measurement biases.

Comparison of partial-PSG with actigraphy

The partial-PSG system agreed with PSG determinations of sleep and wake more often than did wrist actigraphy, even when both were set at their optimum thresholds. This superiority ranged from being only marginally better at identifying sleep and wake, to being far better at predicting wake (see Table 2). Cohen's kappa coefficients suggest that the partial-PSG system had a moderate agreement with PSG beyond chance, whereas both activity monitors had a fair agreement with PSG.

The mean biases of the partial-PSG system for estimating TST and WASO only differed from those of both activity monitors at their best thresholds by a couple of minutes. However, the distribution of differences from PSG was smaller and more consistent for the partial-PSG system. Activity monitors progressively overestimated sleep for shorter sleep durations and underestimated wake for longer wake durations. In contrast, the biases of the partial-PSG system were not affected by these durations. The activity monitors and the partial-PSG system were equally very poor at estimating SOL.

Implications for field research

Since there is large variability in the differences of all device estimates from PSG, summary measures estimated for any one person might not be accurate, even when the mean results of many people are not significantly different from PSG. Therefore, these devices may be more appropriate for monitoring the average sleep/wake behaviors of groups, rather than individuals (Signal et al., 2005). Nonetheless, all of the devices satisfactorily distinguished sleep from wake on an epoch-by-epoch basis, having agreement rates within the accepted range of past validations (Galland et al., 2012; Kosmadopoulos et al., 2012; Paquet et al., 2007; Rupp & Balkin, 2011; Shambroom et al., 2012). Additionally, whereas the partial-PSG system performed better than the activity monitors on several measures in this study, the accuracy of actigraphic sleep/wake assessments in other studies has been shown to improve when analyzed in conjunction with sleep diaries (Acebo et al., 2005; Kushida et al., 2001). Activity monitors and the partial-PSG system are better suited for use in different circumstances. Therefore, when PSG is not possible or practicable, the primary consideration of which system to utilize is their theoretical and applied compatibility with the research question and the proposed research design.

Activity monitors are unobtrusive and can be worn both day and night for weeks and months, making actigraphy, together with sleep diaries, more conducive than self-administered partial-PSG to long-term monitoring of patterns in sleep/wake behavior (Roach, Petrilli, et al., 2012; Sargent et al., 2012). Since the sleep/wake monitor that we evaluated is specifically developed for estimating sleep/wake states, threshold changes can be used to accommodate research

populations with different sleep behaviors. For example, the default threshold, which is the most appropriate in a healthy population, could be lowered for sleep-disordered populations as this has been found to increase wake detection (Hyde et al., 2007; Kushida et al., 2001). Given that the threshold we found best for the energy expenditure monitor is already very low, it is effectively limited to use with nonclinical populations. However, in research involving the concurrent analysis of sleep and other behaviors, such as exercise, using an energy expenditure monitor to measure both sleep/wake patterns and energy expenditure has been found to be both practical and cost effective (Evans et al., 2011).

Unlike actigraphy, the partial-PSG system employed in this study requires participants to wear a headband within close range of the base-station, making it more suited to shorter research protocols focused primarily on planned night-time sleeps. Its ability to correctly predict wake suggests that partial-PSG may be better than actigraphy for use with participants who frequently wake up due to sleeping disorders. Therefore, in research in which sleep staging information is required but PSG is not possible, validated systems like this would be more suitable than actigraphy.

Limitations

Activity monitors are typically used over multiple 24-h periods. Given their relatively low rates of wake detection and overestimation of sleep, a limitation of this study is that activity monitors were only validated for sleep/wake estimation during sleep periods. Future validation studies may consider evaluating the sleep/wake agreement of activity monitors, with a portable PSG system, across longer periods of time.

We were able to concurrently evaluate two types of activity monitors and a partial-PSG system. However, many other products, including a variety of different types of activity monitors and algorithms, are also used as alternatives to PSG (Van de Water et al., 2011). Due to poor standardization across these alternatives, brand-specific differences in design may reduce the generalizability of our findings beyond the devices validated here (te Lindert & Van Someren, 2013).

Conclusion

In conclusion, although the partial-PSG system performed better than sleep/wake and energy expenditure monitors for some measures of agreement with PSG, the alternatives to PSG all performed well. The energy expenditure monitor was not developed for sleep monitoring, but when a very low threshold was applied its capacity for capturing sleep and wake was similar to devices that were. Provided that the optimum thresholds are applied for the activity monitors, we

can make no definitive recommendations that any one of the three validated devices should be used over the others. Instead, when selecting a sleep-monitoring alternative, the primary consideration for researchers is which one will best address their research aim whilst being suitable for their research design.

Author note The authors wish to express appreciation to Michele Lastella for his assistance in data collection.

References

- Acebo, C., Sadeh, A., Seifer, R., Tzischinsky, O., Hafer, A., & Carskadon, M. A. (2005). Sleep/wake patterns derived from activity monitoring and maternal report for healthy 1- to 5-year-old children. *Sleep, 28*, 1568–1577.
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327*, 307–310. doi:10.1016/S0140-6736(86)90837-8
- CamNTEch Ltd. (2008). *The Actiwatch user manual v7.2*. Cambridge, UK: Author.
- Carney, C. E., Lajos, L. E., & Waters, W. F. (2004). Wrist actigraph versus self-report in normal sleepers: Sleep schedule adherence and self-report validity. *Behavioral Sleep Medicine, 2*, 134–143. doi:10.1207/s15402010bsm0203_2
- Carskadon, M. A., Dement, W. C., Mitler, M. M., Roth, T., Westbrook, P. R., & Keenan, S. (1986). Guidelines for the multiple sleep latency test (MSLT): A standard measure of sleepiness. *Sleep, 9*, 519–524.
- 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.
- Darwent, D., Roach, G., & Dawson, D. (2012). How well do truck drivers sleep in cabin sleeper berths? *Applied Ergonomics, 43*, 442–446. doi:10.1016/j.apergo.2011.06.019
- de Souza, L., Benedito-Silva, A. A., Pires, M. L. N., Poyares, D., Tufik, S., & Calil, H. M. (2003). Further validation of actigraphy for sleep studies. *Sleep, 1*, 81–85.
- Evans, D. S., Snitker, S., Wu, S.-H., Mody, A., Njajou, O. T., Perlis, M. L., ... Hsueh, W.-C. (2011). Habitual sleep/wake patterns in the Old Order Amish: Heritability and association with non-genetic factors. *Sleep, 34*, 661–669.
- Galland, B. C., Kennedy, G. J., Mitchell, E. A., & Taylor, B. J. (2012). Algorithms for using an activity-based accelerometer for identification of infant sleep–wake states during nap studies. *Sleep Medicine, 13*, 743–751. doi:10.1016/j.sleep.2012.01.018
- Griessenberger, H., Heib, D. P. J., Kunz, A. B., Hoedlmoser, K., & Schabus, M. (2013). Assessment of a wireless headband for automatic sleep scoring. *Sleep and Breathing, 17*, 747–752. doi:10.1007/s11325-012-0757-4
- Heil, D. P. (2006). Predicting activity energy expenditure using the Actical activity monitor. *Research Quarterly for Exercise and Sport, 77*, 64–80. doi:10.1080/02701367.2006.10599333
- Hyde, M., O'Driscoll, D. M., Binette, S., Galang, C., Tan, S. K., Verginis, N., ... Horne, R. S. C. (2007). Validation of actigraphy for determining sleep and wake in children with sleep disordered breathing. *Journal of Sleep Research, 16*, 213–216. doi:10.1111/j.1365-2869.2007.00588.x
- Jean-Louis, G., Kripke, D. F., Mason, W. J., Elliott, J. A., & Youngstedt, S. D. (2001). Sleep estimation from wrist movement quantified by different actigraphic modalities. *Journal of Neuroscience Methods, 105*, 185–191. doi:10.1016/S0165-0270(00)00364-2
- Kanady, J. C., Drummond, S. P. A., & Mednick, S. C. (2011). Actigraphic assessment of a polysomnographic-recorded nap: A validation study. *Journal of Sleep Research, 20*, 214–222. doi:10.1111/j.1365-2869.2010.00858.x
- Kosmadopoulos, A., Sargent, C., Zhou, X., Darwent, D., & Roach, G. D. (2012). Two for one? Validating an energy expenditure monitor against polysomnography for sleep/wake measurement. In X. Zhou & C. Sargent (Eds.), *Sleep of different populations* (pp. 16–20). Adelaide, Australia: Australasian Chronobiology Society.
- Kushida, C. A., Chang, A., Gadkary, C., Guilleminault, C., Carrillo, O., & Dement, W. C. (2001). Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Medicine, 2*, 389–396. doi:10.1016/S1389-9457(00)00098-8
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. Retrieved from www.jstor.org/stable/2529310
- Oakley, N. R. (1997). *Validation with polysomnography of the Sleepwatch sleep/wake scoring algorithm used by the Actiwatch activity monitoring system (Technical Report)*. Bend, OR: Mini-Mitter Co.
- Paquet, J., Kawinska, A., & Carrier, J. (2007). Wake detection capacity of actigraphy during sleep. *Sleep, 30*, 1362–1369.
- Paterson, J., Dorrian, J., Pincombe, J., Grech, C., & Dawson, D. (2010). Mood change and perception of workload in Australian midwives. *Industrial Health, 48*, 381–389. doi:10.2486/indhealth.MSSW-01
- Rechtschaffen, A., & Kales, A. (1968). *A manual of standardized terminology, techniques, and scoring system for sleep stages of human subjects*. Los Angeles, CA: University of California, Brain Information Service/Brain Research Institute.
- Reid, K., & Dawson, D. (1999). Correlation between wrist activity monitor and electrophysiological measures of sleep in a simulated shiftwork environment for younger and older subjects. *Sleep, 22*, 378–385.
- Roach, G. D., Petrilli, R. M. A., Dawson, D., & Lamond, N. (2012). Impact of layover length on sleep, subjective fatigue levels, and sustained attention of long-haul airline pilots. *Chronobiology International, 29*, 580–586. doi:10.3109/07420528.2012.675222
- Roach, G. D., Sargent, C., Darwent, D., & Dawson, D. (2012). Duty periods with early start times restrict the amount of sleep obtained by short-haul airline pilots. *Accident Analysis and Prevention, 45S*, 22–26. doi:10.1016/j.apergo.2011.06.019
- Robillard, R., Lambert, T. J. R., & Rogers, N. L. (2012). Measuring sleep–wake patterns with physical activity and energy expenditure monitors. *Biological Rhythm Research, 43*, 555–562. doi:10.1080/09291016.2011.614794
- Rupp, T. L., & Balkin, T. J. (2011). Comparison of Motionlogger Watch and Actiwatch actigraphs to polysomnography for sleep/wake estimation in healthy young adults. *Behavior Research Methods, 43*, 1152–1160. doi:10.3758/s13428-011-0098-4
- Sadeh, A., Sharkey, K. M., & Carskadon, M. A. (1994). Activity-based sleep–wake identification: An empirical test of methodological issues. *Sleep, 17*, 201–207.
- Sargent, C., Halson, S., & Roach, G. D. (2012). Sleep or swim? Early-morning training severely restricts the amount of sleep obtained by elite swimmers. *European Journal of Sport Science*. doi:10.1080/17461391.2012.696711
- Shambroom, J. R., Fabregas, S. E., & Johnstone, J. (2012). Validation of an automated wireless system to monitor sleep in healthy adults. *Journal of Sleep Research, 21*, 221–230. doi:10.1111/j.1365-2869.2011.00944.x

- Signal, L. T., Gale, J., & Gander, P. (2005). Sleep measurement in flight crew: Comparing actigraphic and subjective estimates to polysomnography. *Aviation, Space, and Environmental Medicine*, *76*, 1058–1063.
- Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*, 257–268. Retrieved from www.physicaltherapyjournal.com/
- te Lindert, B. H., & Van Someren, E. J. (2013). Sleep estimates using microelectromechanical systems (MEMS). *Sleep*, *36*, 781–789. doi: [10.5665/sleep.2648](https://doi.org/10.5665/sleep.2648)
- Tremaine, R., Dorrian, J., Paterson, J., Neall, A., Piggott, E., Grech, C., & Pincombe, J. (2013). Actigraph estimates of the sleep of Australian midwives: The impact of shift work. *Biological Research for Nursing*, *15*, 191–199. doi:[10.1177/1099800411422249](https://doi.org/10.1177/1099800411422249)
- Van de Water, A. T. M., Holmes, A., & Hurley, D. A. (2011). Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography—a systematic review. *Journal of Sleep Research*, *20*, 183–200. doi:[10.1111/j.1365-2869.2009.00814.x](https://doi.org/10.1111/j.1365-2869.2009.00814.x)
- Weiss, A. R., Johnson, N. L., Berger, N. A., & Redline, S. (2010). Validity of activity-based devices to estimate sleep. *Journal of Clinical Sleep Medicine*, *6*, 336–342.