Received: 4 November 2022 | Revised: 22 December 2022 | Accepted: 28 December 2022

DOI: 10.1111/jsr.13824

RESEARCH ARTICLE

Re-evaluating two popular EEG-based mobile sleep-monitoring devices for home use

Emily Wood | **James K. Westphal** | **Itamar Lerner**

Department of Psychology, The University of Texas at San Antonio, San Antonio, Texas, USA

Correspondence
Itamar Lerner, Department of Psychology, The University of Texas at San Antonio, San Antonio, TX 78249, USA.
Email: itamar.lerner@utsa.edu

Summary

Mobile sleep-monitoring devices for consumer use have been gaining traction as a possible replacement to traditional polysomnography recordings. Such devices potentially offer detailed sleep analysis without requiring the use of designated sleep labs operated by qualified technicians. However, the accuracy of these mobile devices is often not sufficiently evaluated by independent researchers. Here, we compared the performance of two popular mobile electroencephalogram-based systems, the DREEM 3 headband and the Zmachine Insight+. Both devices can be used by participants with minimal training, and provide detailed sleep scoring previously validated by the respective developers in comparison to the gold-standard of polysomnography. A total of 25 participants used both devices simultaneously to record their sleep for two consecutive nights while also keeping a sleep log. We compared the devices' performance, both with each other and in relation to the sleep logs, using several well-known sleep metrics. In addition, we developed a Bayesian lower limit for the devices' expected epoch-by-epoch sleep stage agreement based on their previously published agreement with polysomnography, and compared it with our empirical findings. Results suggest that the Zmachine tends to overestimate periods of wakefulness, likely at the expense of N1/N2 detection, whereas the DREEM tends to underestimate wakefulness and mistake it for N1/N2, with both results more pronounced than previously reported. In addition, we found that the agreement between the devices tends to increase from night 1 to night 2. We formulate several recommendations for how best to use these devices based on our results.

KEYWORDS
DREEM, polysomnography, sleep, validation, Zmachine

1 | INTRODUCTION

While sleep is widely recognized as imperative to proper human development and normal functioning throughout the day, accurate sleep monitoring for clinical and research purposes using home devices still leaves much to be desired. The gold-standard for sleep monitoring is polysomnography (PSG), which allows investigators to determine, among other things, the series of sleep stages that participants experience throughout the night (including N1, N2, N3, rapid eye movement [REM] sleep, and Wake; Abhang et al., 2016; Iber et al., 2007; Malhotra & Avidan, 2013). PSG devices, however, are not simple to use and must be applied by qualified sleep technicians in a lab, a procedure that places significant costs in money and time. Additionally, a known phenomenon affecting PSG monitoring is the "first-night-effects", or FNE, which is characterized by the difficulty to achieve sufficient quality sleep in the first night of testing due to the necessity to sleep in an unfamiliar environment connected to a cumbersome monitoring device (Agnew Jr et al., 1966; Byun et al., 2019).

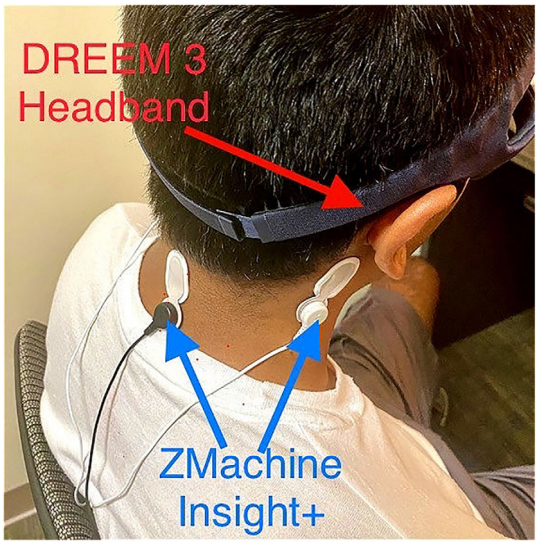

FIGURE 1 Image of the two devices (DREEM, Zmachine) worn simultaneously on a participant's scalp

While ambulatory PSG devices, used in participants' own homes, have been shown to decrease FNE (Coates et al., 1981), they still require the involvement of experienced technicians and are therefore not an ideal solution. Such challenges in obtaining easy and accurate sleep measurements may contribute to the fact that prevalent sleep disorders like insomnia and narcolepsy are underdiagnosed in the general population (Ohayon, 2011).

To tackle the aforementioned challenges, a promising approach gaining traction over the last decade is using mobile sleep-monitoring devices that are easily operated by individuals in their own homes for multiple nights. Critical to this approach is the use of easily-placeable physiological sensors (such as single or few-channel electroencephalogram [EEG] devices) that allow reliable, efficient and automatic sleep staging using internal algorithms. Several such devices have entered the market over the last decade, but few have had their sleep-staging performance validated in comparison to the gold-standard PSG. Two devices, the *Zmachine* and the *DREEM 3 headband*, stand out from the rest due to their comfort, current availability, relatively inexpensive price tag (~$1500 USD) and, most importantly, the fact that their sleep staging capabilities have been previously validated in comparison to PSG (Arnal et al., 2020; Wang et al., 2015), resulting in various research studies incorporating them as part of their protocol (Lerner et al., 2019; Pépin et al., 2021; Zambelli et al., 2022).

While both the Zmachine and the DREEM headband are considered viable tools for interested researchers and clinicians, they have never been directly compared with one another, or validated by independent researchers. Because each system utilizes different EEG channels occupying different areas of the scalp, they are readily available for simultaneous use with minimal interference (Figure 1). In the current study, we sought to compare the sleep staging performance of these devices, as well as several additional measures relevant to individual users. We particularly aimed to determine which, if either, is more suitable for multi-night recordings of sleep in college-aged students in their own homes. By applying several Bayesian analytical tools, we also aimed to exploit the agreement between the devices as an additional, independent validation of their consistency with the gold-standard PSG.

2 | METHODS

2.1 | Sleep-monitoring devices

The DREEM 3 headband is a reduced-montage EEG device that measures brain activity and movements during sleep. It is secured at the back of the head with velcro straps, which can be altered using various sizes of extenders. The DREEM consists of five dry electrodes (F7, F8, Fp2, O1, O2), and a 3D accelerometer that tracks the participants' head and body movement throughout the night. Raw data are automatically analysed by a proprietary algorithm that produces a typical hypnogram with a 30-s epoch resolution differentiating between four sleep stages (N1, N2, N3, REM) and wake (Arnal et al., 2020). The collected data from the DREEM can be uploaded directly from the headband to an online portal using a companion phone app (Alfin) and Bluetooth connection, and thus accessed remotely.

The Zmachine is also a dry-montage EEG recorder that use three disposable, self-stick EEG sensors. One EEG sensor is placed behind each ear (M1, M2), and one on the back of the neck (ground). Sensors are placed outside of the hairline for easy application and quick removal. The Zmachine uses its own proprietary algorithm to track three sleep stages, Light (N1/N2 combined), Deep (N3), and REM, in addition to wake, producing a typical 30-s hypnogram. Additional software can also be downloaded for free to further visualize collected data. Zmachine data are stored locally on a microSD card. Though the storage capacity is sufficient for recording multiple nights in a row, the data must be physically removed from the device between each participant.

2.2 | Participants

A total of 25 participants (15 females) were recruited from the student population at The University of Texas at San Antonio through flyers and participant pools (ages 17–23 years, M = 19.8 ± 1.55 years; years of education: 12–20, M = 14.8 ± 1.8 years). All participants completed the study for either class credit or monetary compensation. Exclusion criteria included a history of sleep disorders, major neurological or psychiatric disorders, head injuries, unusual memory deficits, use of sleep aids (e.g. melatonin), and recreational drug use. All participants gave informed consent for their participation in the study. Seven additional participants began the study but were removed; two because they decided to withdraw from the study and five were discarded due to failure in obtaining usable sleep records for at least 1 night (this relatively high attrition rate stemmed from our design focusing on assessing participants at their natural environment at home with no researcher supervision; this resulted in some participants not tightening one of the devices sufficiently, leading it

to disconnect or slip out of place during the night; two participants were discarded due to poor recording of the Zmachine, and three were discarded due to poor recording of the DREEM).

2.3 | Procedure

Eligible participants arrived at the lab, and received detailed training on the simultaneous usage of the DREEM 3 headband and the Zmachine. Participants were asked to maintain their everyday routine, sleep schedule and caffeine/alcohol consumption throughout the study. Additionally, participants were asked to keep a basic sleep log during the nights of measurement, indicating the time they went to sleep and woke up, how long it took them to fall asleep, the number of awakenings occurring during each night, and what time they got out of bed. After training, the participants took the devices home and spent two consecutive nights monitoring their sleep in their natural environment (i.e. bedrooms). Upon completion, participants returned the devices to the lab and received compensation for their participation. Data for the DREEM 3 headband was then accessed and downloaded through the DREEM portal, and Zmachine data were extracted directly from the device using the micro secure digital card.

2.4 | Statistical analysis

All data analysis was performed using Excel, MATLAB 2021b and SPSS 27. The calculations below were conducted for the full dataset as well as for each of the two experimental nights separately. One participant failed to accurately record sleep data on night 1, and three participants failed to accurately record night 2, due to either battery or disconnection issues. Therefore, analysis of the data for these participants included only the night for which their monitoring was successful. Significance tests directly comparing night 1 and night 2 included only the participants with full datasets for both nights. In addition, two participants failed to report the total time spent awake after falling asleep, and these data points were discarded from analysis.

2.4.1 | Comparison of summary statistics with the sleep logs

We extracted four continuous metrics from each participants' sleep log: (1) TST (total sleep time); (2) SOL (sleep-onset latency); (3) WASO (wake after sleep onset); and (4) NoA (number of awakenings). Corresponding measures were extracted from the sleep data collected by the devices. TST, SOL and WASO are directly reported by each device. To compute the NoA, the epoch-by-epoch sleep staging data were manually scanned to detect any awakenings that occurred after sleep onset lasting four consecutive epochs or more (disregarding short awakenings, which are less likely to be consciously remembered by participants). The four metrics were then compared between the sleep logs and each EEG device to get an overall measure of consistency between each device and participants' self-evaluation of sleep. Results were compared in several fashions. First, Pearson's correlation coefficients were computed between each device and the sleep logs, as well as between the devices themselves for each of the four metrics. The correlation values between the devices and the sleep logs were then compared with each other with a test of correlated correlation coefficients using Fisher Z-transformation (Meng et al., 1992) to determine whether one of the devices was more predictive of the participants' self-evaluations. Comparisons of these correlations between night 1 and night 2 were conducted using Dann and Clark's Z (Raghunathan et al., 1996). Next, to evaluate if any of the devices introduced a consistent bias, mean scores for each metric were individually compared between each device and the sleep log and between the devices themselves, as well as between night 1 and night 2, using independent *t*-tests. In addition, Bland–Altman plots were compiled to qualitatively assess any trends in the differences. Finally, to estimate which device yielded larger absolute biases from the sleep logs, we used the Pitman–Morgan test to compare the variances of the differences between each device and the sleep logs (as well as the variance of the differences between night 1 and night 2 for each device by itself).

2.4.2 | Epoch-by-epoch comparisons of the devices

Epoch-by-epoch sleep staging data produced by each device were aligned and compared against one another to determine levels of sleep scoring agreement. Because the Zmachine does not differentiate between N1 and N2 sleep, these stages were combined in the DREEM output to allow for direct comparison. Epochs with incomplete data (due to either device not reporting sleep stage in the middle of recording for any reason), amounting to 3.81% of the entire dataset, were removed. A total of 20,638 and 16,519 epochs were compared for night 1 and night 2, respectively (37,157 epochs overall). First, we calculated stage-to-stage confusion matrices (including Wake, combined N1/N2, N3, and REM), once with the DREEM Headband serving as the reference (i.e. for each sleep stage reported by DREEM, we computed the percentage of epochs for each sleep stage reported by the Zmachine), and once with the Zmachine serving as reference. Second, Cohen's Kappa was computed to reflect the overall agreement between the devices. We then analysed these matrices and compared them with the previously reported confusion matrices of each device with PSG. Full details of these latter analyses are described in the Results.

3 | RESULTS

3.1 | Comparison of the DREEM and Zmachine with participants' sleep logs

Our first analysis focused on outlining the various differences between the devices, and between the devices and the sleep logs, for

TABLE 1 Summary sleep metrics across participants (standard deviations in parentheses)

Device	DREEM headband	Zmachine	Sleep logs
TST (min)			
Night 1	429.4 (± 140.0)	352.7 (± 88.9)	382 (± 100.1)
Night 2	385.4 (± 88.4)	345.8 (± 81.4)	375.7 (± 87.5)
Combined	407.0 (± 91.1)	345.9 (± 63.0)	377.9 (± 73.0)
SOL (min)			
Night 1	17.9 (± 15.5)	31.4 (± 20.2)	22.6 (± 16.8)
Night 2	18.7 (± 23.1)	25.5 (± 20.5)	17.9 (± 9.8)
Combined	19.4 (± 17.9)	30.5 (± 20.3)	20.0 (± 11.1)
WASO (min)			
Night 1	30.9 (± 40.9)	49.6 (± 68.8)	17.8 (± 21.0)
Night 2	15.9 (± 12.8)	24.8 (± 18.1)	12.2 (± 13.8)
Combined	24.7 (± 23.2)	37.8 (± 36.8)	16.9 (± 17.3)
NoA			
Night 1	1.4 (± 1.3)	3.9 (± 2.3)	2.2 (± 2.3)
Night 2	1.1 (± 1.3)	3.8 (± 2.0)	1.6 (± 1.4)
Combined	1.2 (± 0.9)	3.9 (± 1.5)	2.1 (± 2.1)
Time in N1/N2 (min)			
Night 1	218.4 (± 100.4)	173.6 (± 52.7)	–
Night 2	189.6 (± 47.7)	171.4 (± 42.7)	–
Combined	202.6 (± 56.8)	170.7 (± 35.3)	–
Time in N3 (min)			
Night 1	101.5 (± 36.0)	90.8 (± 25.3)	–
Night 2	99.4(± 40.3)	86.0 (± 22.6)	–
Combined	99.9 (± 35.4)	86.9 (± 21.9)	–
Time in REM (min)			
Night 1	109.5 (± 49.2)	88.3 (± 41.8)	–
Night 2	96.3 (± 35.9)	88.4 (± 45.2)	–
Combined	104.5 (± 40.6)	88.3 (± 33.8)	–

Abbreviation: NoA, number of awakenings; REM, rapid eye movement; SOL, sleep-onset latency; TST, total sleep time; WASO, wake after sleep onset.

the four parameters of interests: TST, SOL, WASO and NoA. Any significant or marginally significant effect is reported (uncorrected *p*-values). The Discussion summarizes the main trends stemming from these effects. Summary statistics across all participants for the DREEM, Zmachine and sleep logs are presented in Table 1. We began by calculating Pearson's correlation coefficients across participants between each device and the sleep log, as well as between the devices themselves (please note that because sleep logs are subjective, they are prone to biases and cannot be taken at face value; however, the biases themselves are often consistent and can be used to assess the devices' accuracy; see Discussion). Results are presented in Figures 2 and 3 (upper panels), and in Table 2. For TST, both devices were highly correlated with the sleep logs ($r_{23} = 0.808$ and $r_{23} = 0.696$ for the DREEM and Zmachine, respectively, both $p < 0.001$; Figure 2, left upper panel), with the correlation for DREEM trending towards a higher value ($z = 1.311$, $p = 0.097$). As seen in the upper left panel of Figure 2, the line of best fit closely resembled a perfect agreement with the sleep logs, whereas the Zmachine tended to estimate lower TST values than the sleep logs. The other three summary metrics (SOL, WASO and NoA) showed a far lower agreement with the sleep logs, with the DREEM practically unable to predict participants' self-rating of SOL ($r_{23} = 0.044$; Figure 2, upper panel, second from left) and the Zmachine unable to predict WASO ($r_{21} = 0.066$; Figure 2, upper panel, second from right). The Zmachine did predict SOL slightly better ($r_{23} = 0.376$, $p = 0.064$) with its correlation being higher than the DREEM ($z = 1.714$, $p = 0.043$; Figure 2, upper panel, second from left), whereas the DREEM, despite its correlation with WASO not reaching statistical significance ($r_{21} = 0.342$, $p = 0.110$; Figure 2, upper panel, second from right), did predict it better than the Zmachine ($z = 1.714$, $p = 0.018$). Finally, DREEM was significantly correlated with participants' self-assessment of NoA ($r_{23} = 0.499$, $p = 0.011$; Figure 2, upper right panel), whereas the correlation was non-significant for the Zmachine ($p = 0.106$), but there was no significant difference between them ($p = 0.193$). Lines of best fit do not necessarily supply valuable information when the correlations are low, but it is notable that they mostly showed the

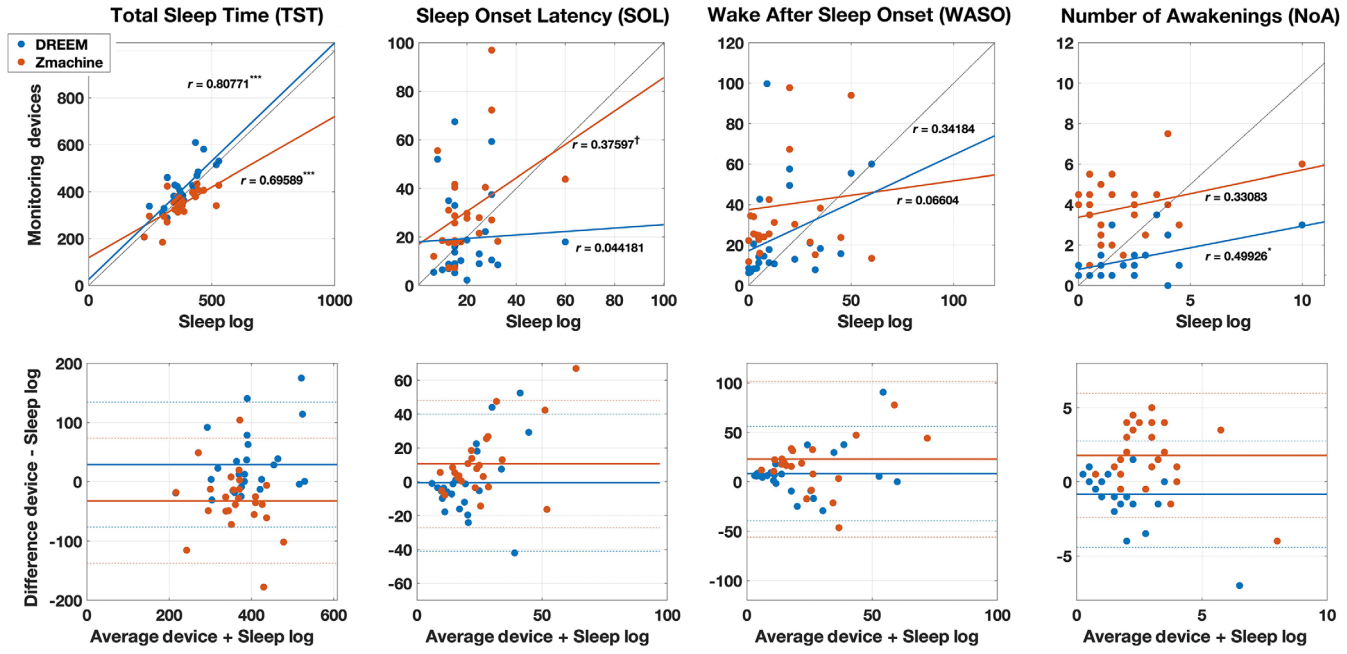

FIGURE 2 Correlation and Bland–Altman plots between each device (DREEM, Zmachine) and the sleep logs for the main four summary sleep metrics. The dashed diagonal line in the correlation plots represents perfect agreement. NoA, number of awakenings; SOL, sleep-onset latency; TST, total sleep time; WASO, wake after sleep onset; †$p < 0.08$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

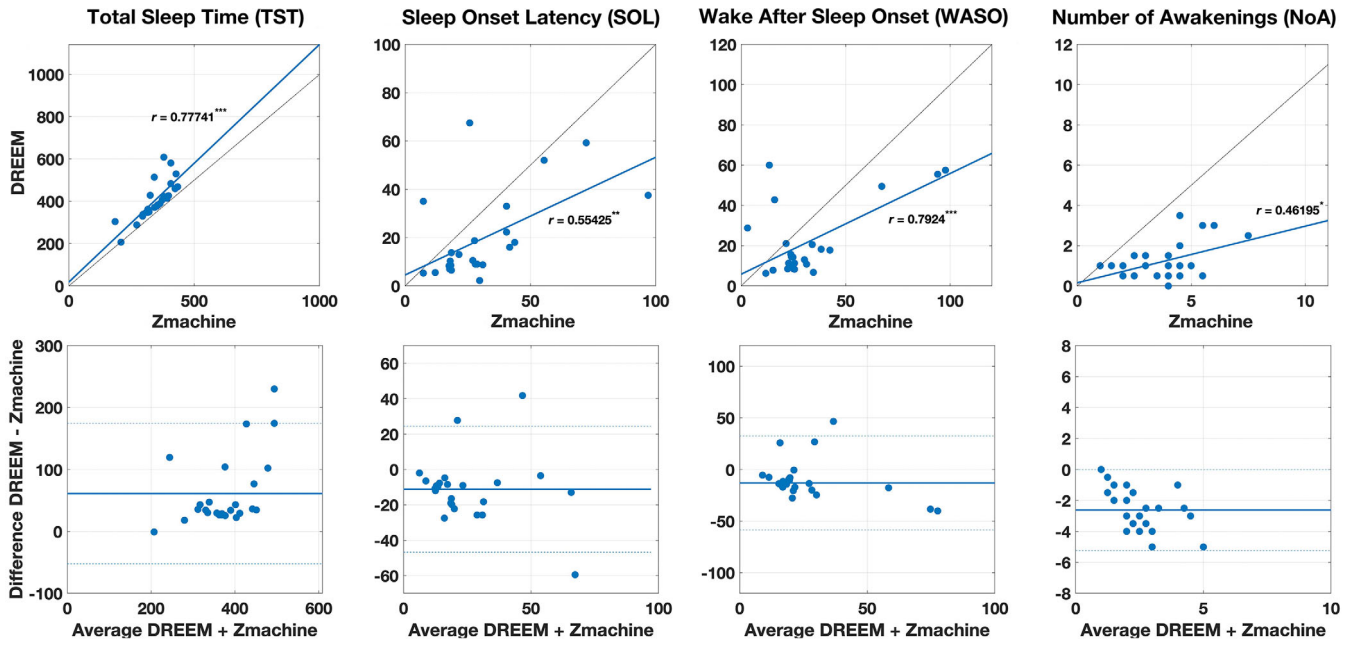

FIGURE 3 Correlation and Bland–Altman plots between the DREEM and Zmachine for the main four summary sleep metrics. The dashed diagonal line in the correlation plots represents perfect agreement. NoA, number of awakenings; SOL, sleep-onset latency; TST, total sleep time; WASO, wake after sleep onset. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

opposite tendency to TST, with the Zmachine almost uniformly predicting higher values than the DREEM for SOL, WASO and NoA, a trend seen throughout this analysis. Finally, the correlations between the two devices themselves (Figure 3) were high for TST, SOL and WASO (all $p < 0.01$), and moderately high for NoA ($p = 0.020$).

We subsequently examined Bland–Altman plots comparing each device with the sleep logs, as well as with each other (Figures 2 and 3, lower panels). Consistent with the main theme in the correlation results, the plots suggested that the DREEM tended to overestimate TST compared with the sleep logs, whereas the Zmachine tended to

TABLE 2 *t*-test and Pearson's correlations comparing each device against the sleep logs, and against each other (SDs in parentheses)

	TST			SOL			WASO			NoA		
***t*-tests**	**M (SD)**	***t***	***p***	**M (SD)**	***t***	***p***	**M (SD)**	***t***	***p***	**M (SD)**	***t***	***p***
DR-SL												
Night 1	47.5 (94.1)	2.47	0.021*	−4.7 (21.9)	−1.06	0.303	13.5 (43.1)	1.48	0.16	−0.8 (2.2)	−1.74	0.095
Night 2	9.7 (53.3)	0.85	0.404	0.8 (24.3)	−0.16	0.872	4 (15.9)	1.13	0.271	−0.5 (1.2)	−2.03	0.056†
Combined	29 (53.7)	2.7	0.012*	−0.58 (20.6	−0.14	0.889	8.3 (24.4)	1.63	0.118	−0.8 (1.8)	−2.3	0.031*
ZM-SL												
Night 1	−29.3 (94.8)	−1.51	0.144	8.8 (23.9)	1.8	0.085	35.7 (71.7)	2.34	0.029*	1.7 (2.6)	3.16	0.004**
Night 2	−29.9 (48)	−2.92	0.008**	7.7 (19.2)	1.87	0.075†	13 (18.6)	3.23	0.006**	2.1 (2)	5.05	< 0.001***
Combined	−32 (53.8)	−2.97	0.007**	10.6 (19.2)	2.75	0.011*	23 (40.3)	2.74	0.012*	1.8 (2.1)	4.17	< 0.001***
DR-ZM												
Night 1	76.7 (108.1)	3.48	0.002**	−13.5 (18.2)	−3.63	0.001**	−18.6 (40.1)	−2.28	0.032*	−2.5 (2.2)	−5.65	< 0.001***
Night 2	39.5 (38.5)	4.83	< 0.001***	−6.8 (18.2)	−1.75	0.094	−8.9 (9.8)	−4.3	< 0.001**	−2.7 (2)	−6.41	< 0.001***
Combined	61 (57.9)	5.27	< 0.001***	−11.14 (18.2)	−3.06	< 0.005**	−13.1 (23.2)	−2.82	0.01**	−2.6 (1.3)	−9.83	< 0.001***
Pearson's *r*	***r***		***p***	***r***		***p***	***r***		***p***	***r***		***p***
DR-SL												
Night 1	0.74		< 0.001***	0.08		0.711	0.216 (22)		0.335	0.345		0.099
Night 2	0.817		< 0.001***	0.09		0.689	0.32 (20)		0.17	0.569		0.006**
Combined	0.808		< 0.001***	0.044		0.834	0.342 (23)		0.11	0.499		0.011*
ZM-SL												
Night 1	0.508		0.011*	0.177		0.409	0.1 (22)		0.659	0.324		0.123
Night 2	0.841		< 0.001***	0.371		0.089	0.348 (20)		0.133	0.357		0.103
Combined	0.696		< 0.001***	0.376		0.068†	0.066 (23)		0.765	0.331		0.106
DR-ZM												
Night 1	0.635		< 0.001***	0.505		0.012*	0.854		< 0.001***	0.361		0.083
Night 2	0.900		< 0.001***	0.656		< 0.001***	0.853		< 0.001***	0.332		0.131
Combined	0.777		< 0.001***	0.554		0.004**	0.792		< 0.001***	0.462		0.02*

Abbreviation: DR, DREEM; NoA, number of awakenings; SL, sleep logs; SOL, sleep-onset latency; TST, total sleep time; WASO, wake after sleep onset; ZM, ZMachine.

†$p < 0.08$.

*$p < 0.05$.**$p < 0.01$.***$p < 0.001$.

Z MACHINE	DREEM				
		WAKE	N2	N3	REM
	WAKE	35.9% (2152)	44.7% (2679)	8.8% (528)	10.6% (638)
	N2	4.4% (689)	70.9% (10998)	10.5% (1636)	14.2% (2201)
	N3	3.1% (245)	12.1% (949)	83.7% (6577)	1.1% (88)
	REM	1.5% (120)	30.2% (2351)	1.2% (93)	67.0% (5213)

PSG	DREEM				
		WAKE	N2	N3	REM
	WAKE	77.6% (2400)	18.0% (556)	0.8% (26)	3.5% (109)
	N2	3.3% (453)	84.0% (11612)	6.0% (839)	6.6% (918)
	N3	0.2% (5)	12.3% (379)	86.6% (2660)	0.9% (27)
	REM	3.2% (150)	10.8% (507)	0.0% (0)	86.0% (4021)

PSG	DREEM				
		WAKE	N2	N3	REM
	WAKE	57.2% [53.8-59.2] (1766)	33.5% [18.0-39.9] (1035)	0.7% [0.0-2.9] (21)	8.7% [3.5-26.0] (268)
	N2	5.1% [3.4-5.1] (706)	82.7% [82.2-84.0] (11423)	6.0% [4.8-7.0] (833)	6.2% [6.0-7.9] (859)
	N3	0.0% [0.0-1.6] (0)	12.3% [10.2-12.9] (379)	86.6% [85.6-88.9] (2660)	1.0% [0.1-1.2] (32)
	REM	3.8% [2.6-5.5] (178)	10.0% [8.7-11.4] (469)	0.2% [0.0-1.4] (10)	86.0% [84.3-86.8] (4021)

DREEM	Z MACHINE				
		WAKE	N2	N3	REM
	WAKE	67.1% (2152)	21.5% (689)	7.6% (245)	3.7% (120)
	N2	15.8% (2679)	64.8% (10998)	5.6% (949)	13.9% (2351)
	N3	6.0% (528)	18.5% (1636)	75.5% (6577)	1.1% (93)
	REM	7.8% (638)	27.0% (2201)	1.1% (88)	64.0% (5213)

PSG	Z MACHINE				
		WAKE	N2	N3	REM
	WAKE	90.8% (15353)	6.6% (1109)	0.0% (0)	2.6% (446)
	N2	5.9% (2721)	83.5% (38769)	3.0% (1385)	7.6% (3548)
	N3	0.9% (59)	25.2% (1700)	73.8% (4974)	0.0% (3)
	REM	0.5% (81)	27.3% (4136)	0.0% (1)	72.1% (10921)

PSG	Z MACHINE				
		WAKE	N2	N3	REM
	WAKE	90.8% [86.3-90.8] (15353)	6.6% [5.9-9.8] (1109)	0.0% [0.0-2.3] (0)	2.6% [2.4-6.2] (446)
	N2	22.5% [18.0-22.9] (10432)	67.4% [66.8-72.9] (31290)	2.6% [0.3-3.9] (1187)	7.6% [6.1-7.9] (3514)
	N3	0.9% [0.0-2.3] (58)	25.3% [24.2-26.3] (1702)	73.8% [72.7-74.6] (4973)	0.0% [0.0-0.9] (3)
	REM	0.7% [0.0-2.3] (98)	27.3% [26.2-27.8] (4139)	0.0% [0.0-1.1] (0)	72.0% [70.3-72.9] (10901)

FIGURE 4 Confusion matrices showing the level of agreement in sleep scoring (in percentages out of total number of epochs for that row; the number of epochs for each cell are displayed below) between the DREEM and Zmachine based on the current study (left column), previously published results for the agreement between each device and polysomnography (PSG; middle column; adapted from Arnal et al., 2020 and Wang et al., 2015; please see Section 3.2), and the corrected agreement between each device and PSG based on our analysis (right column; the values in the brackets display the range for the best 0.1% of results found in the analysis)

underestimate it (Figure 2, lower left panel). Paired *t*-tests comparing the mean difference between each device and sleep logs across participants confirmed this estimation ($t_{24} = 2.702$, $p = 0.013$ and $t_{24} = -2.969$, $p = 0.007$, for the DREEM and Zmachine, respectively). SOL, WASO and NoA all showed the opposite tendency, with Zmachine overestimating all three of these wake-related parameters compared with the corresponding sleep log values ($t_{24} = 2.754$, $p = 0.011$, $t_{21} = 2.736$, $p = 0.012$ and $t_{24} = 4.166$, $p < 0.001$, for SOL, WASO and NoA, respectively; Figure 2, three lower right panels), whereas the DREEM significantly underestimating NoA ($t_{24} = -2.295$, $p = 0.038$; Figure 2, right lower panel). Direct comparisons of the DREEM and Zmachine (Figure 3) showed a significant mean difference for all parameters, with the DREEM producing higher values for TST and Zmachine producing higher values for the remaining parameters (all $p < 0.01$). Finally, comparing, for each of the four metrics, the absolute biases of each device from the sleep logs, a test of variance showed that for TST, the Zmachine had larger discrepancies from the sleep logs compared with the DREEM ($\sigma_{ZMachine} = 40.25$ versus $\sigma_{DREEM} = 24.35$, $t_{21} = 4.806$, $p < 0.001$). Furthermore, graphical inspection of the plots reveals that the discrepancy between the devices and sleep logs for SOL, WASO and NoA tended to increase as their values increased (Figure 2, three lower right panels), evident by a smaller spread on the left side of the plot of each of these three parameters and a higher spread on the right side of the plot. In other words, the longer the SOL and WASO values and the higher the NoA value were, the less the devices reflected participants' self-evaluation of these metrics.

3.2 | Comparison to previously published validation results

We next assessed the epoch-by-epoch sleep scoring agreement between the DREEM and the Zmachine. We found that the agreement between the devices based on Cohen's Kappa was 0.5297, indicating moderate agreement. Confusion matrices for the sleep stage comparison are presented on the two left panels of Figure 4 (upper left panel with DREEM as reference, lower left panel with Zmachine as reference). Compared with the previously published agreement between each device and PSG (yielding Kappa values of 0.72 and 0.748 for Zmachine and DREEM, respectively; see also the corresponding confusion matrices in Figure 4, middle panels; note that values for the DREEM versus PSG reported here are calculated based on the number of epochs cited in Figure 3 of Arnal et al. (2020) rather than the percentages appearing in that figure, Those percentages reflected averages over participants rather than percentage out of the total data collected (Arnal, Personal communication). We chose to calculate the percentages over the full data to allow direct comparison with the corresponding data for the Zmachine in Wang et al., 2015, Table 1), the agreement of DREEM with Zmachine was considerably lower – a generally expected result given that the sleep scoring algorithms of each device were independently developed to fit PSG-based scoring. Nevertheless, we wanted to estimate if our findings are consistent with those previously published results.

While the expected agreement between the devices cannot be accurately deduced solely from their individual agreement with PSG,

it is possible to devise a minimal threshold for that agreement under reasonable assumptions. Specifically, it is expected that the detection of each sleep stage and wake by each device is based on at least somewhat similar EEG markers, given that the very definition of sleep stages is based on such markers (e.g. typical amplitudes of brain waves for REM sleep and N3, sleep spindles and k-complexes for stage N2, etc.; Iber et al., 2007). Therefore, it is expected that the agreement between the devices would be higher than what would have been obtained had their algorithms been completely independent. This allows to set a lower limit for the agreement between the devices based on each device's respective confusion matrix with PSG. In mathematical terms, to compute the lower limit we assume conditional independence between the devices given the PSG sleep scoring, such that: $p(Zm=S_i|\mathrm{DR}=S_j,\mathrm{PSG}=S_k)=p(Zm=S_i|\,\mathrm{PSG}=S_k)$, with *Zm* standing for the Zmachine, DR for the DREEM, and S_i representing any of the four sleep/wake stages (Wake, N1/N2, N3 or REM). In other words, we assume that to compute the probability that the Zmachine classified an epoch as belonging to a certain sleep stage, we only need to know the PSG determined for that epoch, with any additional information given by the DREEM being redundant (and vice versa). We use this mathematical identity and Bayes' theorem to derive an equation for the expected sleep stage classification by the Zmachine given the sleep stage classification of DREEM as follows:

$$\begin{aligned} &p(Zm=S_i|\mathrm{DR}=S_j)\\ &=\sum_k p(Zm=S_i|\mathrm{DR}=S_j,\mathrm{PSG}=S_k)\cdot p(\mathrm{PSG}=S_k|\mathrm{DR}=S_j)\\ &=\sum_k p(Zm=S_i|\mathrm{PSG}=S_k)\cdot p(\mathrm{DR}=S_j|\mathrm{PSG}=S_k)\cdot\frac{p(\mathrm{PSG}=S_k)}{p(\mathrm{DR}=S_j)} \end{aligned} \quad (1)$$

Each element in the final product can be taken directly from the confusion matrices published in the previous validation studies with $p(\mathrm{PSG}=S_k)$ and $p(\mathrm{DR}=S_j)$ calculated based on the total number of epochs in a sleep stage for the respective device out of the total number of epochs (i.e. row sum and column sum in the confusion matrices in the upper middle panel of Figure 4). In a similar vein, the expected sleep stage classification of the DREEM given the sleep stage classification of Zmachine would be:

$$p(\mathrm{DR}=S_i|Zm=S_j)=\sum_k p(\mathrm{DR}=S_i|\mathrm{PSG}=S_k)\cdot p(Zm=S_j|\mathrm{PSG}=S_k)\cdot\frac{p(\mathrm{PSG}=S_k)}{p(Zm=S_j)} \quad (2)$$

Using the equations above, we calculated the minimal expected agreement between the devices for each sleep stage and compared it with the actual agreement calculated based on the data we collected. Results are presented in Figure 5.

As can be seen, for both directions of comparison (DREEM given Zmachine and Zmachine given DREEM), the empirical agreement between the devices was above the lower limit for N3 and REM, but below it for Wake and N1/N2. This suggests that the agreement we found between the devices for N1/N2 and Wake is inconsistent with the previously reported agreement of each device with PSG.

We next asked what changes needed to be applied to the confusion matrices of each device with PSG such that they will be consistent with our own data. Specifically, we attempted to find the minimal correction that could be applied to those previous confusion matrices (middle panels of Figure 4) such that for all stages, the agreement values we found would be higher than the lower limit derived by assuming conditional independence.

To accomplish that, we expressed the confusion matrices in the middle panels of Figure 4 as 24 variables (p_1, ..., p_{24}), reflecting the probabilities of each cell in the first three columns and all four rows of both matrices (the last REM column of each matrix is not included because it is constrained by the requirement that all rows sum up to 1). We then implemented an iterative search algorithm (the Nelder–Mead simplex algorithm, implemented in MATLAB through the *fminsearch* command) to look for values of p_1–p_{24} that are as close as possible to the values of the original confusion matrices (in terms of sum of absolute differences) while still yielding, for each sleep stage, lower-limit agreement values between the devices that are equal or lower than the empirical ones found in our data. The algorithm was initialized to the original values with a small added gaussian noise ($\mu=0$, $\sigma=0.01$), under the constraints that all values cannot be lower than 0 or higher than 1, and that the sum for each row cannot exceed 1 (given that they reflect probabilities). The algorithm was then run and allowed to find a local minimum. Because the algorithm's output is sensitive to initial conditions, we repeated the optimization procedure 100,000 times, each run starting with a slightly different gaussian noise (additional runs did not change the results much further, nor did initializing the algorithm with completely random starting states). We then picked the output that yielded the smallest change from the original values. This result is displayed in the right panels of Figure 4 (with values in brackets reflecting the range of values found for the top 0.1% results). Compared with the original confusion matrices in the middle panels of Figure 4, it is evident that the algorithm suggested two main changes: (a) for the DREEM, a higher percentage of reporting N1/N2 when the PSG determines the participant is awake; (b) for the Zmachine, a higher percentage of reporting Wake when the PSG determines the participant is at N1/N2. The remaining values in the confusion matrices are largely similar to the original ones. Thus, our analysis suggests that the previously reported accuracy in the ability of DREEM to detect wake and the ability of the Zmachine to detect N1/N2 may be exaggerated, while their reported abilities in detecting N3 and REM may be more trustworthy.

3.3 | Differences between night 1 and night 2

To conclude our analyses, we repeated each comparison described above separately for night 1 and night 2 to determine whether the quality of sleep monitoring of each device changed over time and usage, and if indications of FNE could be detected. Effects of the

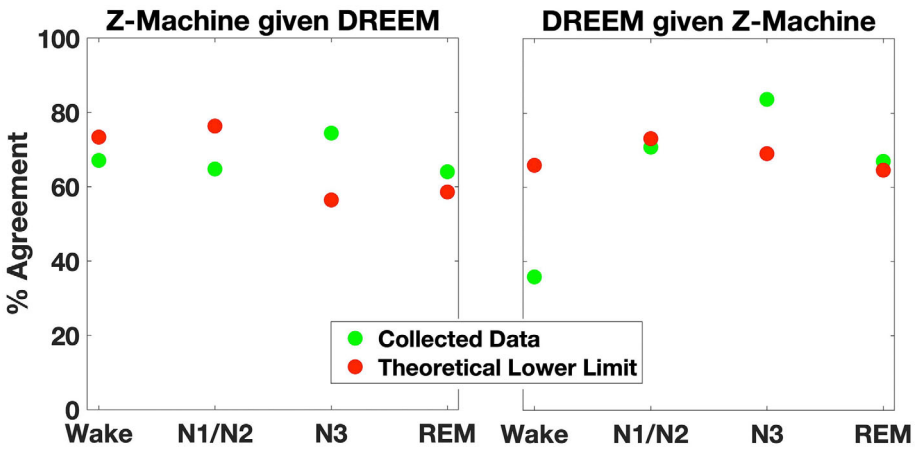

FIGURE 5 Empirical agreement between the Zmachine and the DREEM based on our collected data compared with a lower limit computed based on theoretical considerations from published polysomnography (PSG)-validation studies (Arnal et al., 2020; Wang et al., 2015)

individual nights are displayed in Table 2. Consistent with Section 3.1, only instances where significant or marginally significant changes between the two nights (uncorrected *p*-values) occurred are reported. A summary of the trends stemming from these effects is offered in the Discussion.

Analysing each of the four summary metrics, we found that the Zmachine showed a significant increase in its correlation with the sleep log from night 1 to night 2 for TST ($Z = 2.132$, $p = 0.033$). High correlations were found on both nights, but to a much lower degree for night 1 ($r = 0.508$, $p = 0.011$) compared with night 2 ($r = 0.841$, $p < 0.001$). In comparison, the change in TST for DREEM from night 1 ($r = 0.74$) to night 2 (and $r = 0.817$) was not significant ($p = 0.459$). Fischer Z-transformation confirmed that the difference in correlations between the devices and the sleep log was significant for night 1, but not for night 2 ($z = 1.74$, $p = 0.041$ and $z = -0.44$, $p = 0.329$ for night 1 and night 2, respectively). Likewise, the correlation for TST between the devices themselves significantly increased from night 1 to night 2 ($Z = 2.447$, $p = 0.014$), with *r*-values increasing from 0.635 to 0.9 (both $p < 0.001$). Lastly, the change in direct correlation between the devices for SOL showed a trend ($Z = 1.84$, $p = 0.065$), again reflecting an increase from night 1 ($r = 0.505$, $p < 0.012$) to night 2 ($r = 0.656$, $p < 0.001$).

Comparisons of the change from night 1 to night 2 in the average biases between the devices and the sleep logs showed that there were no significant differences, with only the DREEM showing a trend towards lower discrepancies for SOL on night 2 ($t_{20} = 2.028$, $p = 0.056$). A significant reduction in average bias from night 1 to night 2 for SOL was also evident when comparing directly between the devices themselves ($t_{20} = 2.138$, $p = 0.045$). Comparing the change from night 1 to night 2 in the magnitude of the absolute biases of each device from the sleep log, we found that both devices showed significantly smaller discrepancies in night 2 for both TST and WASO, and the discrepancies between the devices themselves were also significantly lower for night 2 for TST and WASO (all $p < 0.02$). Finally, comparing the epoch-by-epoch agreements between the devices, we found only a slight increase from night 1 to night 2 (Cohen's kappa = 0.50 and 0.56 for night 1 and night 2, respectively).

4 | DISCUSSION

Overall, our results comparing the DREEM headband and Zmachine showed several consistent patterns: the devices were highly correlated among themselves and with participants' sleep logs for detection of sleep, but correlated to a far lesser extent in detection of wake. This discrepancy stemmed from the Zmachine detecting significantly more wake than the DREEM, strongly affecting parameters like WASO, SOL and NoA that constitute a relatively small percentage of total monitoring time, but less affecting TST, which constitutes most of the monitoring time (Table 1). The discrepancy in detecting wake was also evident in the epoch-by-epoch comparison, yielding high levels of confusion between wake and N2 for both devices, and to a lesser degree also between wake and REM (Figure 4, left column).

We used two methods to evaluate which device is more accurate. First, we compared the devices' scoring with the sleep logs. Naively, this comparison seems to show some advantage to the DREEM, yielding smaller average biases from the sleep logs for all the wake-related parameters (SOL, WASO, NoA) and smaller absolute biases for TST. However, it is well established that subjective self-reports of sleep are often biased compared with PSG, with typical underestimation of WASO, overestimation of TST, and either overestimation or accurate estimation of SOL (Kaplan et al., 2012; Lehrer et al., 2022; McCall et al., 1995). Our data in Table 1 are consistent with this pattern for WASO but exhibit the typical overestimation of TST only when comparing the sleep logs with the Zmachine, and only show accurate estimation of SOL when comparing the sleep logs with the DREEM (when compared with the Zmachine, SOL was actually underestimated). These discrepancies are consistent with our conclusion that the DREEM overestimates sleep, and the Zmachine overestimates wake.

The second method we applied to evaluate the devices' accuracy was a re-estimation of their agreement with PSG based on a combination of their direct agreement with each other and previously published validation studies. Results suggested that previous agreement values with PSG were over-optimistic, particularly concerning the degree to which both devices confuse Wake with N1/N2. The DREEM, according to our results, is more susceptible to mistake Wake for N1/N2 than previously reported, whereas the Zmachine tends to

make the opposite error (Figure 4, right column). Note, however, that these results stem from a minimal correction applied to the previous PSG validation findings such that they will fit our data; it is possible that still a bigger correction is warranted, in which case the real agreement might even be lower.

Finally, when comparing differences between night 1 and night 2, the general trend emerging from the various significant results was that the devices mostly improved their agreement for night 2. Because there are only 2 nights of data to assess, we cannot definitively account for what caused this increased agreement; however, several reasonable hypotheses could be suggested. First, this finding could result from participants becoming more comfortable with the monitoring devices, and thus reducing FNE (a hypothesis that is supported by an evident reduction in SOL, WASO and NoA in night 2; Table 1); second, participants' accuracy filling up the sleep logs might have improved, reducing the overall measurement noise; and third, the increased agreement could have resulted from an improved accuracy of the Zmachine's algorithm, which, according to the device's manufacturers, benefits from continued monitoring by allowing the detection of individuals' "sleep signature". Nevertheless, the contribution of this last feature to the improved agreement on night 2 is likely low because even a short period of sleep is noted to produce an accurate sleep signature – making it probable that a signature has already been established early on during the first night.

4.1 | Additional considerations and recommendations

Beyond accuracy in sleep detection and scoring, additional aspects should be considered when deciding which mobile sleep-monitoring device to use in research or clinical settings. Both the DREEM and Zmachine are user friendly, take only a few minutes to set up, and can be used autonomously after a short demonstration. However, the DREEM is slightly more prone to falling out of position during the night compared with the Zmachines' stick-on electrodes; the DREEM user may therefore be required to take additional measures, such as wearing a sweatband over the device, to keep it in place overnight (which, admittedly, is not a perfect solution for active sleepers who tend to toss and turn a lot during the night). On the other hand, due to the Zmachine's placement of electrodes on the back of the neck, it could potentially be at a disadvantage in recording important sleep events like spindles and k-complexes – events that some investigations may find useful. In addition, because the Zmachine uses disposable sensors, their regular replenishment may add significant costs in the long term. Taking both sleep scoring accuracy and practical use considerations into account, it seems to us that the DREEM headband should be preferred when sleep staging is central to the investigators, especially when the target population is young, healthy students – a typical setting of basic research studies. To compensate for the relative weakness of the DREEM in differentiating Wake from N1/N2, future studies could employ simultaneous use of wrist actigraphs, which have previously been used alongside EEG-based sleep devices and shown positive results (Lerner et al., 2016; Martin & Hakim, 2011), In contrast, when the population of interest may be more prone to face monitoring difficulties (e.g. elderly population or people with sleep deficiencies), the Zmachine may be preferred due to its lower sensitivity to movements during sleep, especially if the differentiation between sleep and wake is of higher importance than detailed sleep staging. This latter scenario may be more common in clinical investigations. Nevertheless, please note that our cohort of participants was composed of college-aged students without sleep abnormalities; therefore, any generalizations from this cohort to other populations should be done with caution.

4.2 | Limitations and future directions

Given that our study aimed to test the DREEM and Zmachine devices in natural settings, it unavoidably limited the level of control we could exert over participants' behaviour. For example, participants had varying sleep and wake times, and some had more trouble keeping the devices well attached to their scalp than others. An alternative approach could have had the experiment be fully conducted in the lab – at the expense of preserving a natural sleep environment and potentially increasing FNE effects. In addition, our study measured sleep for only two consecutive nights, while the developers of the Zmachine suggest that its algorithm improves its accuracy with time (a claim that our night 1–night 2 comparisons offer some support for); it is therefore possible that the agreement between the devices would have improved with longer monitoring, and further studies with these devices could look to extend the monitoring period, perhaps to a week, to get a more robust assessment. Nevertheless, the published validation of the Zmachine's (as well as the DREEM's) algorithm against PSG only used a single experimental night (Wang et al., 2015), informing our decision to restrict the monitoring time. Finally, our method of utilizing direct comparisons between two mobile sleep-monitoring devices to draw conclusions on their agreement with PSG (providing that previous PSG validation results are available for each) could be utilized by future studies to allow relatively easy (if limited) re-evaluation of the accuracy of such devices against the gold-standard in the industry.

AUTHOR CONTRIBUTIONS

Emily Wood collected the data, assisted in data analysis, and wrote parts of the manuscript; James K. Westphal collected the data, assisted in data analysis, and wrote parts of the manuscript; Itamar Lerner conceptualized the study, oversaw all statistical analyses, and wrote parts of the manuscript.

CONFLICT OF INTEREST

All authors declare no conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Itamar Lerner https://orcid.org/0000-0003-2525-2741

REFERENCES

Abhang, P. A., Gawali, B., & Mehrotra, S. (2016). *Introduction to EEG-and speech-based emotion recognition*. Academic Press.

Agnew, H. W., Jr., Webb, W. B., & Williams, R. L. (1966). The first night effect: An EEG study of sleep. *Psychophysiology*, *2*(3), 263–266.

Arnal, P. J., Thorey, V., Debellemaniere, E., Ballard, M. E., Bou Hernandez, A., Guillot, A., Jourde, H., Harris, M., Guillard, M., Van Beers, P., Chennaoui, M., & Sauvet, F. (2020). The Dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging. *Sleep*, *43*(11), zsaa097.

Byun, J. H., Kim, K. T., Moon, H. J., Motamedi, G. K., & Cho, Y. W. (2019). The first night effect during polysomnography, and patients' estimates of sleep quality. *Psychiatry Research*, *274*, 27–29.

Coates, T. J., George, J. M., Killen, J. D., Marchini, E., Hamilton, S., & Thorensen, C. E. (1981). First night effects in good sleepers and sleep-maintenance insomniacs when recorded at home. *Sleep*, *4*(3), 293–298.

Iber, C., Ancoli-Israel, S., Chesson, A. L., & Quan, S. F. (2007). The new sleep scoring manual–the evidence behind the rules. *Journal of Clinical Sleep Medicine*, *3*(2), 107.

Kaplan, K. A., Talbot, L. S., Gruber, J., & Harvey, A. G. (2012). Evaluating sleep in bipolar disorder: Comparison between actigraphy, polysomnography, and sleep diary. *Bipolar Disorders*, *14*(8), 870–879.

Lehrer, H. M., Yao, Z., Krafty, R. T., Evans, M. A., Buysse, D. J., Kravitz, H. M., … Hall, M. H. (2022). Comparing polysomnography, actigraphy, and sleep diary in the home environment: The study of Women's health across the nation (SWAN) sleep study. *Sleep Advances*, *3*(1), zpac001.

Lerner, I., Kerbaj, T., & Gluck, M. A. (2019). When sleep-dependent gist extraction goes awry: False composite memories are facilitated by slow wave sleep. *CogSci* (pp. 2119–2124).

Lerner, I., Lupkin, S. M., Corter, J. E., Peters, S. E., Cannella, L. A., & Gluck, M. A. (2016). The influence of sleep on emotional and cognitive processing is primarily trait-(but not state-) dependent. *Neurobiology of Learning and Memory*, *134*, 275–286.

Malhotra, R. K., & Avidan, A. Y. (2013). Sleep stages and scoring technique. In S. Chokroverty, & R. J. Thomas (Eds.), *Atlas of Sleep Medicine* (2nd ed., p 77-99). Elsevier.

Martin, J. L., & Hakim, A. D. (2011). Wrist actigraphy. *Chest*, *139*(6), 1514–1527.

McCall, W. V., Turpin, E., Reboussin, D., Edinger, J. D., & Haponik, E. F. (1995). Subjective estimates of sleep differ from polysomnographic measurements in obstructive sleep apnea patients. *Sleep*, *18*(8), 646–650.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175.

Ohayon, M. M. (2011). Epidemiological overview of sleep disorders in the general population. *Sleep Medicine Research*, *2*(1), 1–9.

Pépin, J. L., Bailly, S., Mordret, E., Gaucher, J., Tamisier, R., Ben Messaoud, R., Arnal, P. J., & Mignot, E. (2021). Greatest changes in objective sleep architecture during COVID-19 lockdown in night owls with increased REM sleep. *Sleep*, *44*(9), zsab075.

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*(2), 178–183.

Wang, Y., Loparo, K. A., Kelly, M. R., & Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. *Nature and Science of Sleep*, *7*, 101–111.

Zambelli, Z., Jakobsson, C. E., Threadgold, L., Fidalgo, A. R., Halstead, E. J., & Dimitriou, D. (2022). Exploring the feasibility and acceptability of a sleep wearable headband among a community sample of chronic pain individuals: An at-home observational study. *Digital Health*, *8*. https://doi.org/10.1177/20552076221097504

How to cite this article: Wood, E., Westphal, J. K., & Lerner, I. (2023). Re-evaluating two popular EEG-based mobile sleep-monitoring devices for home use. *Journal of Sleep Research*, e13824. https://doi.org/10.1111/jsr.13824