[Title Page]

Performance of the Verily Study Watch for Measuring Sleep Compared to Polysomnography

Authors: Sohrab Saeb, PhD[1]*; Benjamin W. Nelson, PhD[1, 2]*; Poulami Barman, MS[1]; Nishant Verma, PhD[1]; Hannah Allen, BS[1]; Massimiliano de Zambotti, PhD[3]; Fiona C. Baker, PhD[3]; Nicole Arra, BA[3]; Niranjan Sridhar, PhD[1]; Shannon S. Sullivan, MD, MSc[1, 4]; Scooter Plowman, MD, MBA, MHSA, MSc[1]; Erin Rainaldi, MS[1]; Ritu Kapur, PhD[1, 5]; Sooyoon Shin, PhD[1]

Affiliations:

[1]Verily Life Sciences, South San Francisco, CA, United States

[2]Division of Digital Psychiatry, Department of Psychiatry, Harvard Medical School and Beth Israel Deaconess Medical Center, Boston, MA, United States

[3]Center for Health Sciences, SRI, Menlo Park, California, USA, United States

[4]Division of Pulmonary, Asthma, and Sleep Medicine, Department of Pediatrics, Stanford University School of Medicine, Palo Alto, CA, United States

[5]Department of Neurology, Radboud University Medical Center, Nijmegen, Netherlands

*These authors contributed equally to this work, first authorship shared

Corresponding Author:

Sohrab Saeb, PhD

Verily Life Sciences

279 E Grand Ave, South San Francisco, CA 94080

P. 650-495-7100

Email: sosata@verily.com

Keywords: Sleep-wake detection; sleep stage; digital health measures; polysomnography; free-living; sleep detection accuracy; wearable technology

Running title: Verily Study Watch Performance Against PSG

Article length: Abstract, 215 (limit 250). Main, ~2156 (limit 4K)

Disclosures/Acknowledgements page

Study funding statement: The study was sponsored by Verily Life Sciences

Data [/code/materials] sharing statement: Data from this study are not available due to the nature of this program. Participants did not consent for their data to be shared publicly.

Authors' disclosures:

S Saeb, BWN, PB, NV, HA, NS, S Sullivan, SP, ER, RK, S Shin report employment and equity ownership in Verily Life Sciences.

MDZ, FB and NA received research funding through their institution from Verily Life Sciences for study execution.

Authors' contributions:

Study concept and design: SSaeb, BWN, SSullivan, MdZ, RK, SShin

Data collection: Verily Life Sciences, FCB, NA

Data analysis and interpretation: PB, SSaeb

Draft writing and review: All

Draft approval for submission: All

Acknowledgements:

Authors wish to acknowledge participants and study personnel that made the study possible.

Abstract

Introduction: This study evaluated the performance of a wrist-worn wearable, Verily Study Watch (VSW), in detecting key sleep measures against polysomnography (PSG). **Methods:** We collected data from 41 adults without obstructive sleep apnea or insomnia during a single overnight laboratory visit. We evaluated epoch-by-epoch performance for sleep versus wake classification, sleep stage classification and duration, total sleep time (TST), wake after sleep onset (WASO), sleep onset latency (SOL), sleep efficiency (SE), and number of awakenings (NAWK). Performance metrics included sensitivity, specificity, Cohen's kappa, and Bland-Altman analyses. **Results:** Sensitivity and specificity (95% CIs) of sleep versus wake classification were 0.97 (0.96, 0.98) and 0.70 (0.66, 0.74), respectively. Cohen's kappa (95% CI) for 4-class stage detection was 0.64 (0.18, 0.82). Most VSW sleep measures had proportional bias. The mean bias values (95% CI) were 14.0 minutes (5.55, 23.20) for TST, -13.1 minutes (-21.33, -6.21) for WASO, 2.97% (1.25, 4.84) for SE, -1.34 minutes (-7.29, 4.81) for SOL, 1.91 minutes (-8.28, 11.98) for *light sleep* duration, 5.24 minutes (-3.35, 14.13) for *deep sleep* duration, and 6.39 minutes (-0.68, 13.18) for *REM sleep* duration. Mean and median NAWK count differences (95% CI) were 0.05 (-0.42, 0.53) and 0.0 (0.0, 0.0), respectively. **Discussion:** Results support applying the VSW to track overnight sleep measures in free-living settings. Registered at clinicaltrials.gov (NCT05276362).

Introduction

Characterizing sleep in a free-living setting provides valuable insights into physical and mental health. Changes in sleep may be key in the diagnosis of sleep disorders like insomnia and hypersomnia, and are clinically meaningful components for tracking mental and cardiovascular health, as well as other conditions (Parish, 2009)(Freeman et al., 2020)(Tobaldini et al., 2019)(Young et al., 2008)(Ahmadi et al., 2009)(Hayashino et al., 2010). The gold standard for sleep assessment is lab-based polysomnography (PSG). However, PSG is resource intensive, challenging to administer and subject to intra- and inter-scorer variability, moreover, availability of PSG laboratories may be limited (Norman et al., 2000)(Deutsch et al., 2006). It is also impractical for long-term surveillance, and may be prone to artifacts that affect representativeness, such as altered sleep patterns due to the novelty of a laboratory, and/or the discomfort of the electrode setting (Toussaint et al., 1995). Furthermore, while portable PSG tools do exist, they still have limited application in free-living environments or routine clinical care.

Wearable sensors, particularly wrist-worn devices, provide a promising avenue for sleep assessment in free-living settings. These devices are widely available, relatively inexpensive, comfortable to wear during sleep and include physiological sensors, such as photoplethysmogram (PPG) and accelerometer, that can be used for sleep monitoring (Imtiaz, 2021)(de Zambotti et al., 2024). However, before utilizing wearable-based technology as a routine approach to monitor daily sleep, whether for care or for research purposes, it is important to conduct performance evaluation of devices and algorithms compared to a gold standard reference such as PSG. Furthermore, researchers now know the importance of conducting those analytical and clinical evaluations across diverse and representative populations, such as participants with different ages or skin tones, to increase confidence in the generalizability of the results (Colvonen et al., 2020)(Baumert et al., 2023)(Nelson et al., 2020).

This study evaluated the performance of the Verily Study Watch (VSW, a wrist-worn wearable) to monitor sleep in a diverse cohort of sleepers without obstructive sleep apnea (OSA) or elevated insomnia symptoms, by comparing VSW sleep measures against measures obtained from PSG-based labels. The VSW classifies every 30-second epoch into 4 sleep-related stages: wake, light sleep, deep sleep, and rapid eye movement (REM) sleep. These classifications enable the calculation of multiple sleep measures that provide information on the quantity and the quality of an individual's overnight sleep. In this study, the measures of interest were: total sleep time (TST), wake after sleep onset (WASO), sleep efficiency (SE), sleep onset latency (SOL), number of awakenings (NAWK), and duration of each sleep stage. Our main objective was to compare epoch-by-epoch VSW- against PSG- derived classification of sleep-versus-wake state and of sleep stages. Additionally, we wanted to assess the VSW's accuracy for all computed sleep measures (listed above). Finally, we wanted to evaluate any potential variability in the performance of the VSW's sleep algorithm across demographic factors such as age, sex, body mass index (BMI), skin tone, and arm hair density.

Methods

Participants

The basic setup and eligibility for the study have been described elsewhere (Nelson, 2024, submitted). Eligible participants were between 18-80 years old, agreed to abstain from any drugs or medications that may affect sleep or wakefulness prior to and during the lab visit, and did not have identified symptoms of sleep disorders, such as obstructive sleep apnea (OSA, defined by OSA 50 score ≥5), or elevated insomnia symptoms (defined by having an insomnia severity index (ISI) score ≥ 8). The study was approved by the WCG Institutional Review Board (20215892), and all participants provided informed consent.

This study was registered at clinicaltrials.gov (NCT05276362).

Data Collection

For each participant, data were collected during a single overnight stay in a sleep laboratory at a single site (SRI; Menlo Park, California), between February 14th and September 1st, 2023. Participants slept in comfortable, sound-proof and temperature-controlled bedrooms. Standard PSG protocols were used for preparation, recording procedures, and instrument calibration (Nelson, 2024, submitted).

Study Watch Data

During their overnight visit, participants wore the VSW on their dominant wrist. This analysis was part of a larger study including two devices: the Verily Numetric Watch (VNW) (Nelson, 2024, submitted), in addition to the VSW. VSW is equipped with two sensors: a green-light PPG sensor, and a 3-axis accelerometer. Both sensors had a sampling frequency of 60 Hz (in the VNW, the PPG sensor consists of a green light emitter diode and two PPG signal channels and the sampling rate of the 3-axis accelerometer is 104 Hz). Using the PPG and accelerometer signals, the VSW classifies every 30-second epoch into one of the following 4 classes: *wake*, *light sleep*, *deep sleep*, and *REM sleep*.

The sleep stage classification algorithm consisted of a deep convolutional neural network that was initially trained using 10,000 nights of data from the Sleep Heart Health Study (SHHS) and Multi-Ethnic Study of Atherosclerosis (MESA) public datasets (Sridhar et al., 2020). The algorithm was fine-tuned using a smaller dataset collected at SRI, consisting of 30 nights of PSG-labeled data.

The overnight sleep measures, including TST, WASO, SE, SOL, NAWK, and sleep stage durations (Supplementary Table 1), for each participant were calculated using the VSW's predicted sleep stages, from the time the lights were turned off ("lights-off") to the time lights were turned back on ("lights-on"). VSW start time was synced to the Lights Off time recorded on PSG to ensure alignment for analysis of simultaneously recorded signals, using procedures described elsewhere (Nelson, 2024, submitted) (de Zambotti et al., 2019).

Reference Data

Standard laboratory PSG sleep assessment including electroencephalography (EEG), submental electromyography and bilateral electrooculography was performed according to the American Academy of Sleep Medicine (AASM) guidelines. Leg movement, electrocardiography (ECG), respiratory, and oxygen saturation signals were also collected and used to confirm the

absence of sleep disordered breathing. All recordings were performed using the Compumedics Grael® HD-PSG system (Compumedics, Abbotsford, Victoria, Australia). Two independent sleep scorers labeled every 30-second epoch of the PSG data by one of the following categories: *wake*, *N1*, *N2*, *N3*, *REM*. Inter-rater reliability (Kappa) between the two scorers was 91%, and discrepancies were resolved by a third scorer.

For this analysis, PSG stages *N1* and *N2* were combined into a single *light sleep* category, and PSG *N3* was termed *deep sleep*.

Similar to VSW, for each participant, the overnight sleep measures for PSG were calculated using the sleep scorer's stage labels from lights-off to lights-on.

Performance Evaluation

Performance evaluation was done based on an existing standardization framework (Menghini et al., 2021).

We evaluated the epoch-by-epoch performance of VSW's sleep stage classification against PSG in two ways: (1) *sleep* versus *wake* classification, using *sleep* as the positive class; and (2) 4-class (wake, light, deep, REM) sleep stage classification. For the evaluation of sleep vs wake classification, we estimated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). We calculated the 95% CI using cluster bootstrapping, and we accounted for the clustering of epochs within a participant using logistic mixed-effect regression models with the participant as random effect. For the 4-class stage classification, we used Cohen's kappa and accuracy along with their 95% bootstrapped CIs. Additionally we evaluated performance for each sleep stage by reporting Cohen's Kappa, accuracy, PPV and sensitivity using the average method (Menghini et al., 2021). To obtain performance metrics on each sleep stage, the outcomes were dichotomized to the sleep stage of interest against all others. The average method calculates kappa for each individual participant and then averages out the kappa across all participants with their associated bootstrapped 95% CIs. All analyses were confined to the lights-off to lights-on period.

For evaluating the performance of all overnight sleep measures except NAWK, we performed the Bland Altman analysis, estimating the mean bias and lower and upper limits of agreement, testing for the assumptions of proportional bias, heteroscedasticity, and normality. For NAWK, we estimated the mean and median count difference and linearly weighted Cohen's kappa with their 95% CIs.

Finally, we evaluated all performance metrics across the participant subgroups, including age, sex, BMI, skin tone, arm hair index. For subgroups with insufficient number of samples (< 10), we did not evaluate the performance.

All analyses were performed with R version 4.3.1 (2023-06-16).

Results

There were 41 adult participants (18 male, age range: 18-78 years) in this study. Participants had a diverse range of skin tones, BMI, and arm hair density (Supplementary Table 2).

VSW estimated sleep stages for a total of 38,796 epochs with data collected between lights-off and lights-on for each participant.

The sensitivity (95% CI) of the VSW in classifying sleep vs wake was 0.97 (0.96, 0.98), specificity (95% CI) was 0.70 (0.66, 0.74), PPV (95% CI) was 0.93 (0.92, 0.95), and NPV (95% CI) was 0.83 (0.78, 0.88) (Table 1).

The accuracy (95% CI) of the VSW sleep algorithm in classifying all 4 sleep stages was 0.78 (0.58, 0.89), and the kappa (95% CI) was 0.64 (0.18, 0.82) (Table 2). There was variability in the performance across different sleep stages, with *light sleep* stage prediction having the lowest accuracy (Table 2), as there were instances of confusion between the *light sleep* stage and all other stages (Supplementary Table 3).

Mean bias and 95% CI values for all overnight sleep measures is shown in Table 3. Bland-Altman analyses (Figure 1) showed that all measures had significant proportional bias, with the VSW overestimating the measures at the lower end of the distribution, and underestimating them at the upper end, relative to the PSG. For all overnight sleep measures except the sleep stage durations, the assumption of normality was false, and for all measures except SE the assumption of homoscedasticity was true.

Performance of the VSW metrics across demographic subgroups of age, sex, BMI, skin tone, and arm hair density are reported (Supplementary Tables 4 and 5) without formal statistical testing, due to small subgroup sample size.

Discussion

The results of this study show the ability of the VSW to capture information related to sleep quantity and quality, as well as the distribution of sleep stages across overnight periods in individuals without OSA or elevated insomnia symptoms. The sensitivity and specificity of the VSW in classifying sleep vs wake were 0.97 and 0.70 respectively, and the Cohen's kappa for the 4-class stage classification was 0.64. This performance supports the application of the VSW to monitor overnight sleep in free-living settings.

As with other wearable sleep-wake detection devices (Pesonen and Kuula, 2018)(de Zambotti et al., 2016)(Miller et al., 2022), the sleep algorithm in this study was more likely to miss wake than sleep, as reflected in the higher sensitivity relative to specificity, and the positive and negative bias values for TST and WASO, respectively. When evaluating the performance of sleep monitoring devices, the AASM has established a range of 'allowable differences', based on actigraphy studies conducted in patients with specific sleep disorders (e.g. insomnia)(Smith et al., 2018). The 95% CIs of the mean bias estimates for TST, WASO, SOL, and SE measured by the VSW were within those allowable difference ranges. However, for the *proportional* mean bias estimates, which account for variations in bias over the range of measurement, 95% CIs exceeded these thresholds at lower and higher ends of the measurements (Figure 1). Nonetheless, applying the AASM standards to these results may require caution. Unlike the studies included in the AASM assessment, the present study excluded (via questionnaire) participants with symptoms of certain sleep disorders.

There are a few caveats to consider when interpreting our results. First, data collection for this study took place at a sleep laboratory, with standardized study boundaries and settings, such as lights-on/off to define the "in bed' time period when an individual is (in theory) set to sleep. Free-living environments are more organic and complex, and the generation of sleep measures in them may require additional layers of data. Following the prior example, defining "in bed" time may necessitate additional sensor readings, which then would be integrated into the derivation of the measures, particularly sleep stage classification and duration, or SOL.

Another caveat is that participants in this study were free of sleep-related diagnoses and symptoms (such as OSA or heightened insomnia symptoms). Participants with certain clinical conditions may manifest different patterns in their biological signals (e.g., pulse rate) and/or sleep architecture, which could complicate the sleep stage classification task. Future studies should evaluate the performance of VSW in real-world settings and in clinically relevant populations such as individuals with sleep disorders.

In summary, we evaluated the performance of the VSW and its algorithm to classify sleep versus wake state and the four different sleep stages in sleepers without OSA or heightened insomnia symptoms, as well as a series of measures that illustrate the quantity and quality of overnight sleep. The results demonstrate the potential of VSW to classify sleep vs wake states and sleep stages and compute overnight sleep measures when compared to gold-standard PSG measurements. These findings support further application of the VSW to tracking the overnight sleep behaviors in sleepers without OSA or heightened insomnia symptoms in free-living settings.

References

Ahmadi, N., Shapiro, G. K., Chung, S. A., and Shapiro, C. M. (2009). Clinical diagnosis of sleep apnea based on single night of polysomnography vs. two nights of polysomnography. *Sleep Breath.* 13, 221–226. doi: 10.1007/s11325-008-0234-2

Baumert, M., Hartmann, S., and Phan, H. (2023). Automatic sleep staging for the young and the old - Evaluating age bias in deep learning. *Sleep Med.* 107, 18–25. doi: 10.1016/j.sleep.2023.04.002

Colvonen, P. J., DeYoung, P. N., Bosompra, N.-O. A., and Owens, R. L. (2020). Limiting racial disparities and bias for wearable devices in health science research. *Sleep* 43. doi: 10.1093/sleep/zsaa159

Deutsch, P. A., Simmons, M. S., and Wallace, J. M. (2006). Cost-effectiveness of split-night polysomnography and home studies in the evaluation of obstructive sleep apnea syndrome. *J. Clin. Sleep Med.* 2, 145–153. Available at: https://www.ncbi.nlm.nih.gov/pubmed/17557487

de Zambotti, M., Baker, F. C., Willoughby, A. R., Godino, J. G., Wing, D., Patrick, K., et al. (2016). Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol. Behav.* 158, 143–149. doi: 10.1016/j.physbeh.2016.03.006

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

de Zambotti, M., Goldstein, C., Cook, J., Menghini, L., Altini, M., Cheng, P., et al. (2024). State of the science and recommendations for using wearable technology in sleep and circadian research. *Sleep* 47. doi: 10.1093/sleep/zsad325

Freeman, D., Sheaves, B., Waite, F., Harvey, A. G., and Harrison, P. J. (2020). Sleep disturbance and psychiatric disorders. *Lancet Psychiatry* 7, 628–637. doi: 10.1016/S2215-0366(20)30136-X

Hayashino, Y., Yamazaki, S., Takegami, M., Nakayama, T., Sokejima, S., and Fukuhara, S. (2010). Association between number of comorbid conditions, depression, and sleep quality using the Pittsburgh Sleep Quality Index: results from a population-based survey. *Sleep Med.* 11, 366–371. doi: 10.1016/j.sleep.2009.05.021

Imtiaz, S. A. (2021). A Systematic Review of Sensing Technologies for Wearable Sleep Staging. *Sensors* 21. doi: 10.3390/s21051562

Menghini, L., Cellini, N., Goldstone, A., Baker, F. C., and de Zambotti, M. (2021). A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. *Sleep* 44. doi: 10.1093/sleep/zsaa170

Miller, D. J., Sargent, C., and Roach, G. D. (2022). A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults. *Sensors* 22. doi: 10.3390/s22166317

Nelson, B.W., Low, C.A., Jacobson, N. Areán, P., Torous, J., Allen, N. B. (2020). Guidelines for

289 wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *NPJ Digit.*
290 *Med*. 3, 90. https://doi.org/10.1038/s41746-020-0297-4

291 Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A., and Rapoport, D. M. (2000). Interobserver
292 agreement among sleep scorers from different centers in a large dataset. *Sleep* 23, 901–908.
293 Available at: https://www.ncbi.nlm.nih.gov/pubmed/11083599

294 Parish, J. M. (2009). Sleep-related problems in common medical conditions. *Chest* 135, 563–
295 572. doi: 10.1378/chest.08-0934

296 Pesonen, A.-K., and Kuula, L. (2018). The Validity of a New Consumer-Targeted Wrist Device in
297 Sleep Measurement: An Overnight Comparison Against Polysomnography in Children and
298 Adolescents. *J. Clin. Sleep Med.* 14, 585–591. doi: 10.5664/jcsm.7050

299 Smith, M. T., McCrae, C. S., Cheung, J., Martin, J. L., Harrod, C. G., Heald, J. L., et al. (2018).
300 Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake
301 Disorders: An American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and
302 GRADE Assessment. *J. Clin. Sleep Med.* 14, 1209–1230. doi: 10.5664/jcsm.7228

303 Sridhar, N., Shoeb, A., Stephens, P., Kharbouch, A., Shimol, D. B., Burkart, J., et al. (2020).
304 Deep learning for automated sleep staging using instantaneous heart rate. *NPJ Digit Med* 3,
305 106. doi: 10.1038/s41746-020-0291-x

306 Tobaldini, E., Fiorelli, E. M., Solbiati, M., Costantino, G., Nobili, L., and Montano, N. (2019).
307 Short sleep duration and cardiometabolic risk: from pathophysiology to clinical evidence. *Nat.*
308 *Rev. Cardiol.* 16, 213–224. doi: 10.1038/s41569-018-0109-6

309 Toussaint, M., Luthringer, R., Schaltenbrand, N., Carelli, G., Lainey, E., Jacqmin, A., et al.
310 (1995). First-night effect in normal subjects and psychiatric inpatients. *Sleep* 18, 463–469. doi:
311 10.1093/sleep/18.6.463

312 Young, J. S., Bourgeois, J. A., Hilty, D. M., and Hardin, K. A. (2008). Sleep in hospitalized
313 medical patients, part 1: factors affecting sleep. *J. Hosp. Med.* 3, 473–482. doi: 10.1002/jhm.372

Tables

Table 1. Performance of VSW's sleep vs wake classification against PSG reference.

	Sensitivity (95% CI)	**Specificity (95% CI)**	**NPV (95% CI)**	**PPV (95% CI)**
Sleep vs Wake	0.97 (0.96, 0.98)	0.70 (0.66, 0.74)	0.83 (0.78, 0.88)	0.93 (0.92, 0.95)

CI: Confidence Interval; NPV: Negative Predictive Value; PPV: Positive Predictive Value

318 Table 2. VSW's performance in 4-class sleep stage detection against the PSG reference.

319

Sleep Stage	Kappa (95% CI)	Accuracy (95% CI)	PPV (95% CI)	Sensitivity (95% CI)
Overall	0.64 (0.18, 0.82)	0.78 (0.58, 0.89)	NA	NA
Wake	0.70 (0.43, 0.90)	0.92 (0.76, 0.98)	0.82 (0.51, 0.98)	0.71 (0.45, 0.94)
Light	0.60 (0.29, 0.78)	0.80 (0.66, 0.89)	0.80 (0.55, 0.91)	0.81 (0.59, 0.94)
Deep	0.66 (0.17, 0.91)	0.92 (0.84, 0.98)	0.69 (0.09, 0.97)	0.77 (0.37, 0.98)
REM	0.74 (0.38, 0.90)	0.92 (0.82, 0.98)	0.76 (0.44, 0.96)	0.84 (0.47, 0.99)

CI: confidence interval; PPV: Positive Predictive Value; REM: Rapid Eye Movement

320

Table 3. Performance of VSW overnight sleep measures against PSG reference.

Measure	Mean			Assumptions	Proport. Bias		Lower LOA		Upper LOA	
	PSG (SD)	VSW (SD)	Bias (95% CI)		Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
TST (min)	384.98 (60.85)	398.98 (49.04)	14.00 (5.55, 23.20)	Prop Bias = T Normality = F Heteroscedastic = F	125.51 + -0.29 x PSG	Intercept = [65.50, 186.28] Slope = [-0.45, -0.14]	-9.04	[-87.36,18.83]	105.04	[26.76,133.36]
WASO (min)	62.72 (49.97)	49.60 (38.73)	-13.12 (-21.33, -6.21)	Prop Bias = T Normality = F Heteroscedastic = F	7.11 + -0.32 x PSG	Intercept = [-3.68, 15.8] Slope = [-0.51, -0.10]	bias - 2.46(1.32 + 0.18 x PSG)	Intercept = [-2.91, 6.58] Slope = [0.06, 0.27]	bias + 2.46(1.32 + 0.18 x PSG	Intercept = [-2.91, 6.58], Slope = [0.06, 0.27]
SE (%)	81.69 (11.71)	84.67 (9.01)	2.97 (1.25, 4.84)	Prop Bias = T Normality = F Heteroscedastic = T	30.04 + -0.33 x PSG	Intercept = [14.81, 42.28] Slope = [-0.47, -0.16]	bias - 2.46(11.98 + -0.11 x PSG)	Intercept = [4.19, 21.56], Slope = [-0.22, -0.02]	bias + 2.46(11.98 + -0.11 x PSG	Intercept = [4.19, 21.56], Slope = [-0.22, -0.02]
SOL (min)	25.43 (20.37)	24.09 (19.73)	-1.34 (-7.29, 4.81)	Prop Bias = T Normality = F Heteroscedastic = F	11.7 + -0.51 x PSG	Intercept = [3.28, 21.81] Slope = [-0.86, -0.15]	-44.21	[-84.21,-7.38]	34.21	[-5.59,70.51]
Light (min)	240.65 (49.27)	242.56 (43.83)	1.91 (-8.28, 11.98)	Prop Bias = T Normality = T Heteroscedastic = F	83.34 + -0.34 x PSG	Intercept = [40.42, 123.65] Slope = [-0.51, -0.17]	-25.06	[-119.66,-10.10]	107.06	[12.69,122.06]
Deep (min)	63.39 (27.19)	68.62 (20.12)	5.24 (-3.35, 14.13)	Prop Bias = T Normality = T	54.33 + -0.77 x PSG	Intercept = [40.81, 67.44] Slope = [-0.95, -	-72.30	[-97.21,3.10]	39.31	[14.51,114.76]

				Heteroscedastic = F		0.60]				
REM (min)	82.49 (25.46)	88.88 (23.60)	6.39 (-0.68, 13.18)	Prop Bias = T Normality = T Heteroscedastic = F	45.69 + -0.48 x PSG	Intercept = [26.79, 69.87] Slope = [-0.78, -0.24]	-21.48	[-84.16,-3.21]	68.48	[5.91,86.55]

Measure	**PSG Mean (SD)**	**VSW Mean (SD)**	**Mean Difference (95% CI)**	**PSG Median**	**VSW Median**	**Median Difference (95% CI)**	**Linear Weighted Kappa (95% CI)**
NAWK (count)	2.17 (1.96)	1.88 (2.31)	0.05 (-0.42, 0.53)	1	1	0.0 (0.0, 0.0)	0.58 (0.41, 0.71)

CI: confidence interval; LOA: Limits of Agreement; NAWK: Night Awakenings; PPV: Positive Predictive Value; PSG: Polysomnography; REM: Rapid Eye Movement; SD: Standard Deviation; SE: Sleep Efficiency; SOL: Sleep Onset Latency; TST: Total Sleep Time; VSW: Verily Study Watch; WASO: Wake After Sleep Onset

Figure 1.

Bland-Altman plots of overnight sleep measures for the device (VSW) against the reference (PSG). Solid red lines indicate mean bias, dotted red lines indicate 95% CI of mean bias, solid gray lines indicate the 95% LOAs, and dotted gray lines indicate 95% CI of LOAs. Black dots are observations.
(CI: confidence interval; REM: Rapid Eye Movement; SD: Standard Deviation; SE: Sleep Efficiency; SOL: Sleep Onset Latency; TST: Total Sleep Time; WASO: Wake After Sleep Onset)