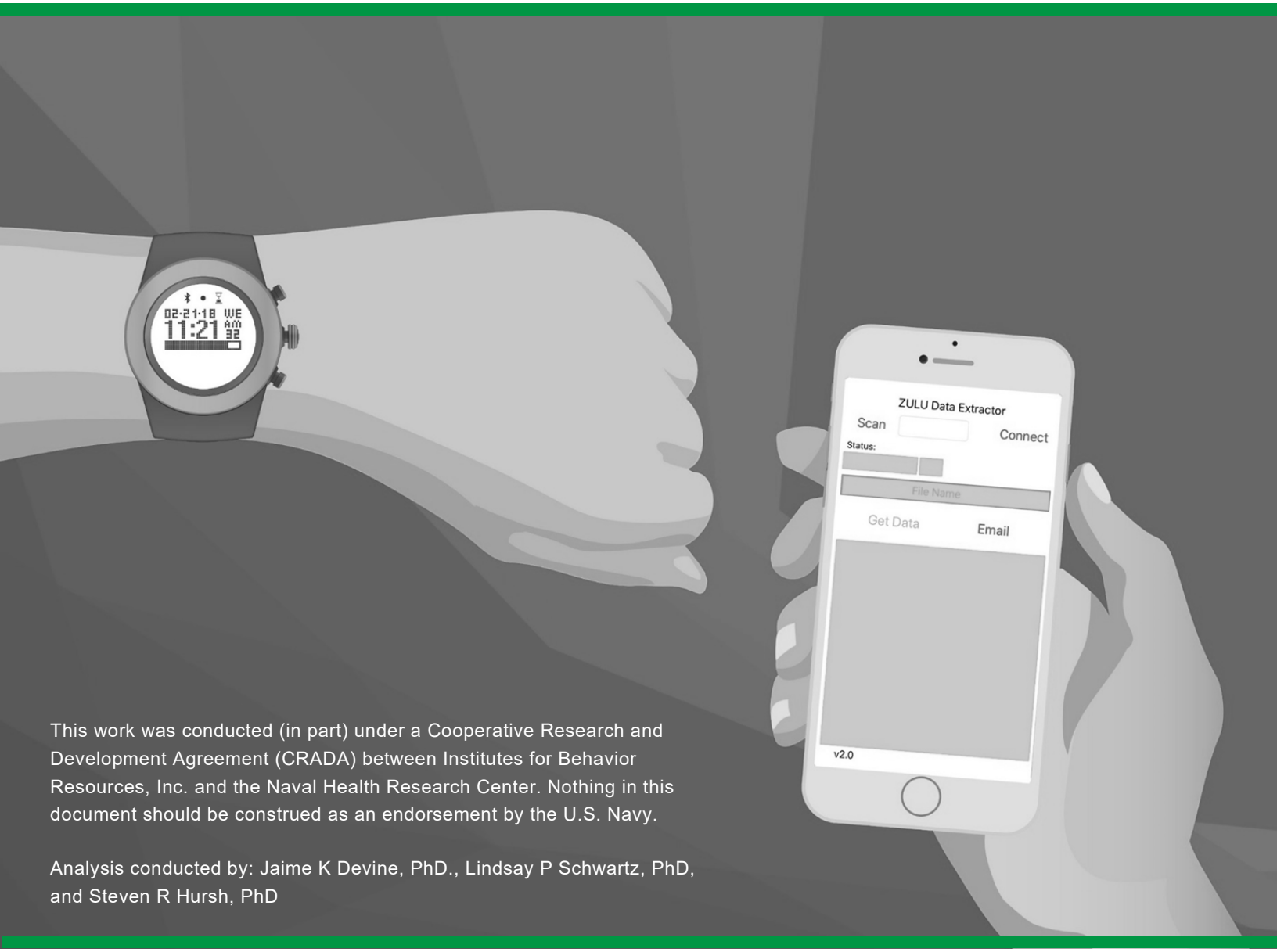
SAFTEFAST

The Science of Performance at Work

Zulu Watch Validation Against Polysomnography



This work was conducted (in part) under a Cooperative Research and Development Agreement (CRADA) between Institutes for Behavior Resources, Inc. and the Naval Health Research Center. Nothing in this document should be construed as an endorsement by the U.S. Navy.

Analysis conducted by: Jaime K Devine, PhD., Lindsay P Schwartz, PhD, and Steven R Hursh, PhD

Contents

Introduction

The gold-standard of sleep measurement is polysomnography (PSG), which uses a number of physiological measures in order to determine whether an individual is awake or asleep (sleep-wake determination) and whether they are experiencing light or deep sleep (sleep scoring). Despite the high quality of the data, sleep measurement through PSG requires a controlled laboratory environment and specially-trained technician and thus, is not practical for daily measurement of sleep in the real world. Many research-grade actigraphy devices reliably measure sleep outside the laboratory, but data must be downloaded and interpreted by researchers in order to provide any meaningful information.

Ideally, a sleep measurement device would be able to accurately collect sleep data during daily operations and provide direct feedback to the wearer in order to properly manage sleep in real time. The Zulu watch is a commercial sleep tracking device which is capable of on-wrist sleep-wake determination and sleep scoring. The purpose of this white paper is to demonstrate the validity of the sleep-tracking algorithm in the Zulu watch to accurately determine sleep and estimate sleep stages compared to PSG.

Automatic Sleep Determination and Sleep Scoring by the Zulu Watch

The Zulu hardware device collects activity data in two-minute epochs and automatically scores data on-wrist based on a proprietary algorithm for sleep-wake determination. Data can then be exported as scored sleep interval information for up to 80 sleep intervals and/or as raw two-minute epoch-by-epoch data from the previous 7 days. Devices were programmed to detect multiple sleep episodes per day; minimum sleep detection was set at 20 minutes. Files reported all sleep interval start and end times, sleep interval duration in minutes, and sleep efficiency (SE) as a percentage. Sleep interval duration was used to identify time in bed (TIB). Total sleep time (TST) was calculated as TIB*SE to determine the number of minutes during the sleep interval in which sleep was actually occurring. Epoch data are scored as on-wrist or off-wrist, and periods of wake are scored as “0”, restless or interrupted sleep is scored as “1”, light sleep is scored as “2”, and deep sleep is scored as “3”.

Study Design and Statistical Analysis

Eight healthy young adult participants wore Zulu watches continuously over the course of a three-day consecutive PSG sleep study. Participants arrived at the lab each evening, and were provided with an eight-hour sleep opportunity based on their habitual bed and wake times. Participants left the lab during the day but were instructed to keep wearing the Zulu continuously at home, except for showers or rigorous activities where the watches may become damaged. They were instructed to not nap at home during the study period. One participant’s third night of data was excluded because the Zulu watch indicated that it was off-wrist during the study night. Another participant’s second night of data was excluded because of technical issues collecting PSG data. PSG data were collected in 30-second epochs and scored by a trained technician in accordance with AASM guidelines (Berry et al., 2012).

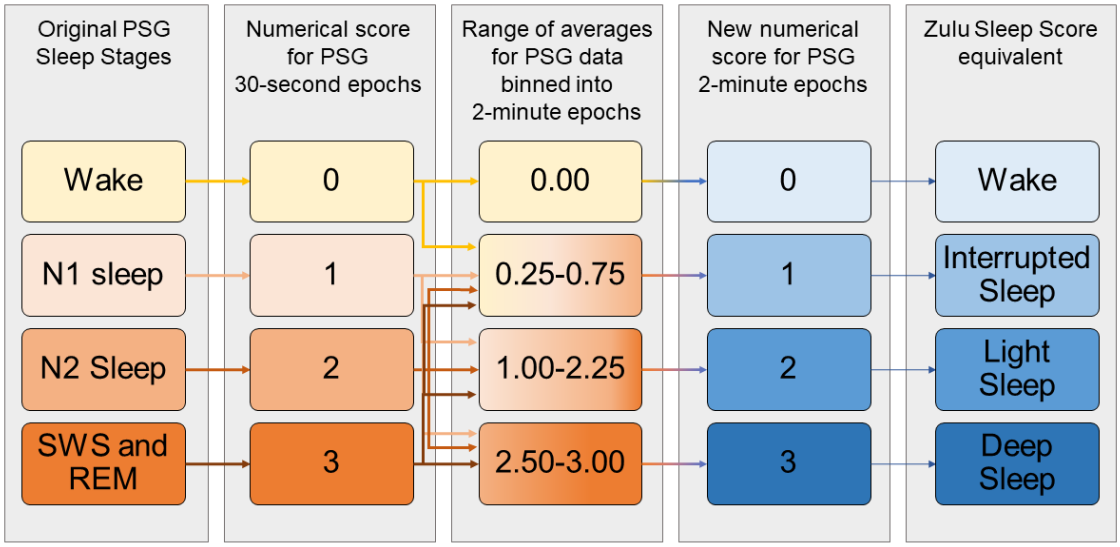
Zulu watches were compared against PSG in three separate domains: sleep-wake determination, sleep summary statistics, and sleep scoring. While for PSG, these domains are determined holistically based on visual inspection of a polysomnogram, the Zulu watch provides separate output files for epoch-by-epoch data with sleep-wake and sleep scores and sleep interval

summary statistics. It was therefore necessary to analyze PSG and Zulu data independently for each domain.

For sleep-wake determination, PSG and Zulu watch data were recoded so that all epochs that were scored as sleep were considered a 1, and all wake epochs were scored as 0. PSG scores were recalculated into two-minute epochs by averaging binary sleep scores from 30 second intervals into two-minute bins and rounding up to the nearest integer. Using binary sleep-wake scores, Zulu watch actigraphy data were again compared epoch-by-epoch with PSG by subtracting Zulu sleep scores from the corresponding PSG score to determine agreement for individual epochs. Sensitivity, specificity, and accuracy were then computed for the total recording period. Accuracy was calculated as the proportion of all epochs wherein Zulu scoring was in agreement with PSG over total PSG-recorded epochs. Sensitivity was calculated as the proportion of all epochs identified as sleep by the Zulu watch over all epochs identified as sleep by PSG, and specificity was calculated as the proportion of all epochs identified as wake by the Zulu watch over all epochs identified as wake by PSG.

To compare PSG and Zulu sleep scoring, PSG scores were recalculated into two-minute epochs by averaging sleep staging scores from 30 second intervals into two-minute bins. Zulu sleep scoring divides sleep periods into “interrupted sleep”, “light sleep” and “deep sleep”, which do not directly correlate to the PSG scoring system. Previous studies have compared commercial wearable or mobile app sleep scoring of light and deep sleep against PSG under the assumptions that sleep stages N1 and N2 are comparable to “light sleep” while N3, also called slow wave sleep (SWS), is thought of as “deep sleep”, and REM is its own category (Bhat et al., 2015; de Zambotti, Cellini, Goldstone, Colrain, & Baker, 2019; Z. Liang & Chapa-Martell, 2019; Zilu Liang & Martell, 2018). However, the Zulu watch does not provide a separate category score for REM sleep, and provides the category of “interrupted sleep” which has not been compared against PSG in previous studies. Without further guidance from the literature, we have attempted to achieve equivalence by recoding PSG sleep stages into Zulu sleep scores through the logic summarized in Figure 1.

Figure 1: Logic for rescoring of PSG sleep staging data into 2-minute epochs for comparison against Zulu watch sleep scores



Two-minute bins with an average of 0.0 indicated wakefulness for the entire period and were considered equivalent to a Zulu score of 0 (wake). Two-minute bins with averages between 0.25 to 0.75 were considered to contain enough epochs scored as wake (0) to be considered equivalent to interrupted sleep, or a Zulu score of 1. Light sleep (Zulu score of 2) was defined to be any two-minute bins which resulted in an average between 1 and 2.25. Deep sleep (Zulu score of 3) was defined as any two-minute bins predominated by PSG epochs scored as SWS or REM, resulting in an average greater than or equal to 2.5. Zulu watch actigraphy data were then compared bin-by-bin with PSG by subtracting Zulu sleep scores from the corresponding PSG score to determine agreement for individual epochs. Accuracy for each sleep stage was calculated as the proportion of all bins wherein Zulu scoring was in agreement with PSG bins for that stage (0, 1, 2, or 3).

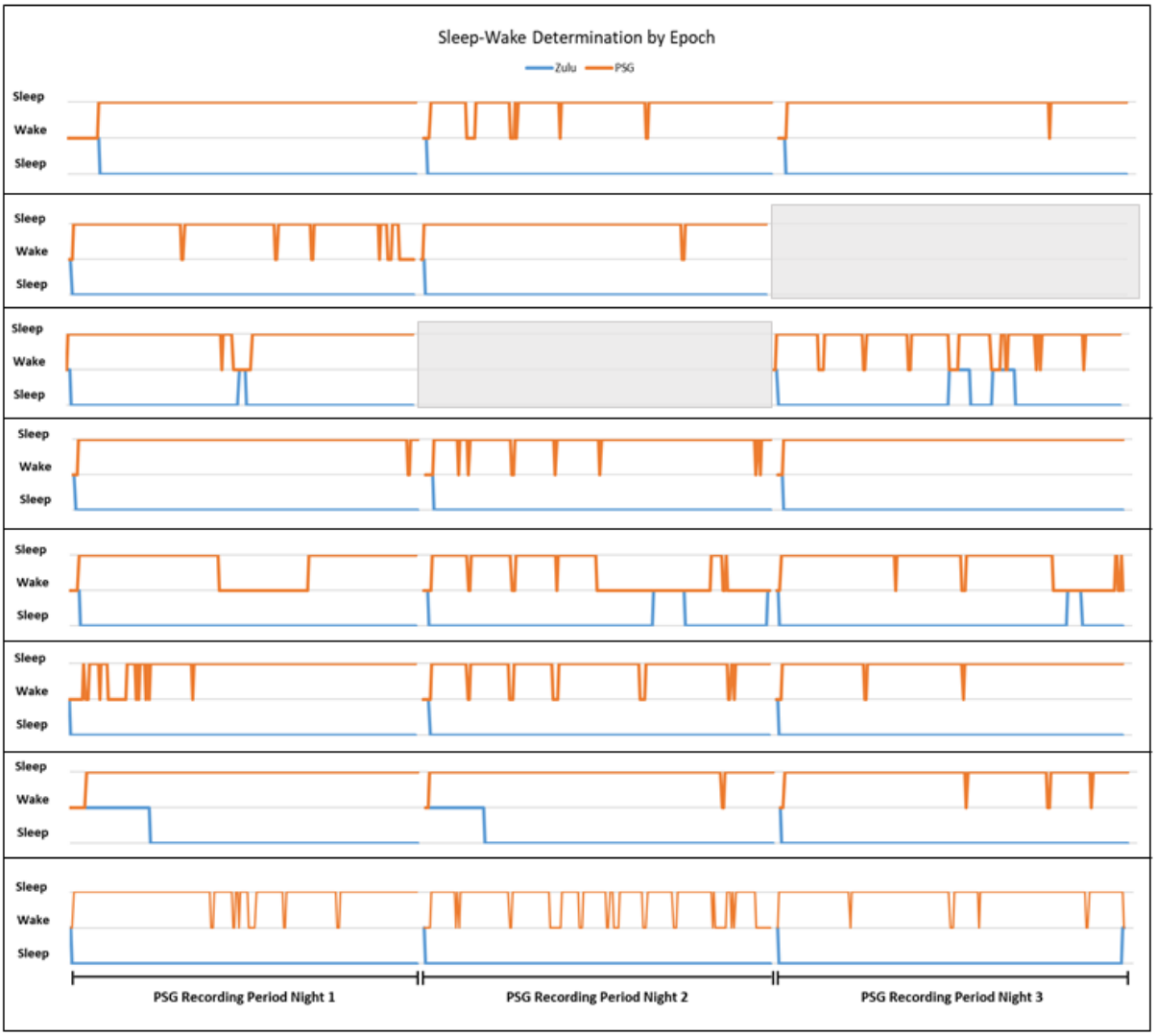
Sleep summary statistics (TIB, TST, SE, time of sleep onset, and time of final awakening) were calculated from the PSG recording period and the Zulu sleep interval summary data. It should be noted that Zulu watches detect periods of wakefulness during sleep episodes but do not report wake after sleep onset (WASO) or number of awakenings. Paired samples t-tests were then conducted to determine if a statistically significant mean difference existed between PSG and Zulu measures of TST, TIB, SE, time of sleep onset, and final awakening.

"Zulu watch achieves comparability against PSG"

Zulu Watch Sleep-Wake Determination Compared to PSG

Figure 2 depicts a comparison of epoch-by-epoch paired PSG and Zulu sleep-wake determination data for each participant in this data set. Two participants supplied only two nights of viable comparison data; missing data are reflected in Figure 1 by the gray blocks. Two-minute epochs are graphed along the x-axis and shown in study days. Sleep versus wake determination is plotted in orange on the top line for PSG and in blue directly below for the corresponding Zulu watch data. PSG sleep was coded as 1 for scored sleep epochs and 0 for wake epochs, while Zulu-determined sleep was coded as -1 for scored sleep epochs and 0 for wake epochs. This coding was done specifically to create this plot.

Figure 2: Comparison of Epoch Level Sleep-Wake Determination by PSG and Zulu Watch by Participant



Accuracy over all epochs was high (90.53%±8.23%) as was sensitivity for the detection of sleep epochs (97.89%±4.17%). Zulu watches did not show good specificity (35.23%±24.75%) for identifying epochs of wake during a sleep interval. Low specificity is an issue across research actigraphy devices and scoring algorithms (Marino et al., 2013; Paquet, Kawinska, & Carrier, 2007), with longer epoch lengths relating to lower specificity (Ancoli-Israel et al., 2015).

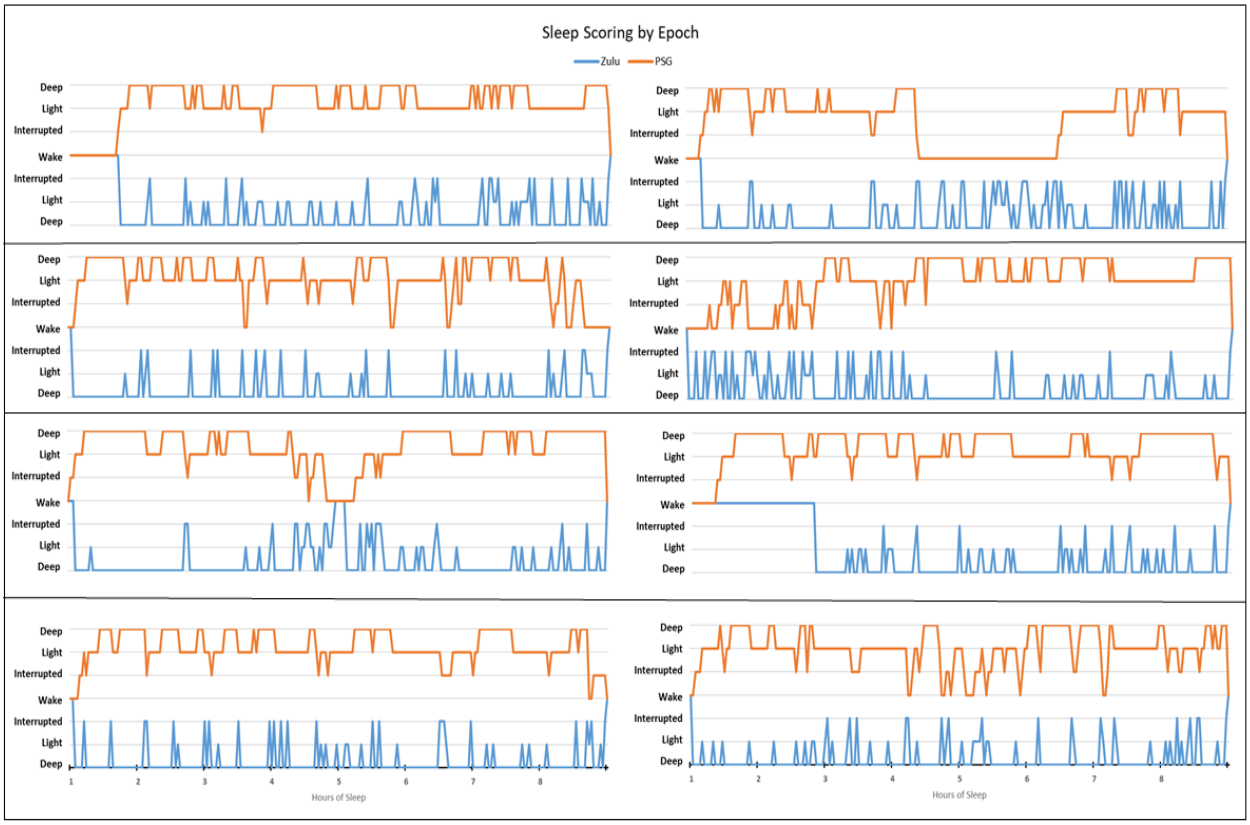
Zulu Watch Sleep Scoring Compared to PSG

Figure 3 depicts a comparison of epoch-by-epoch paired PSG and Zulu sleep scoring data from the first study night for each participant in this data set. Only one night of data is depicted in Figure 2 in order to improve readability. Two-minute epochs are graphed along the x-axis and shown in hours of sleep over the study night. Sleep stages are plotted in orange on the top line for PSG and in blue directly below for the corresponding Zulu watch data. For the purpose of creating this plot, PSG sleep epochs scored as N3 or REM sleep were coded as 3 (indicating deep sleep) while Zulu-determined deep sleep was coded as -3; PSG sleep epochs scored as N2 sleep were coded as 2 (indicating light sleep) while Zulu-determined light sleep was coded as -2;

PSG sleep epochs scored as N1 sleep was coded as 1 (indicating interrupted sleep) while Zulu-determined interrupted sleep was coded as -1 and all wake epochs were coded as 0. This coding was done specifically to create this plot.

Zulu sensitivity for detecting deep sleep (PSG-scored N3 and REM sleep) was high (84.16%±7.20%) but sensitivity was poor for the determination of light sleep/N2 (10.73%±3.48%) or interrupted sleep/N1 (27.07%±19.38%). Poor sensitivity for Zulu estimates of interrupted and light sleep may be due to the discrepancy between sleep scoring terminology by Zulu versus PSG. We have attempted to achieve equivalence by recoding PSG sleep stages into their logical Zulu sleep scores equivalents (see Figure 1). However, what constitutes "interrupted sleep" is open to interpretation and the distinction between wake, interrupted sleep and light sleep could differ on an epoch-to-epoch basis. Zulu sleep scoring may provide greater insight into neurobiological sleep architecture than research-grade actigraphy. Zulu sleep scores are intended as an estimation of sleep depth, not as a direct equivalent to PSG sleep staging. Researchers must use their best judgement when interpreting Zulu sleep scores in the context of their study design.

Figure 3: Comparison of Epoch Level Sleep Scoring by PSG and Zulu Watch by Participant



Zulu Watch Determination of Sleep Summary Statistics Compared to PSG

Sleep summary statistics as determined by PSG and Zulu watches and comparisons are summarized in Table 1. Significance is indicated in Table 1 by an asterisk (*) at $p<0.05$. Paired samples t-test analyses indicated agreement between PSG and Zulu measures of TST, time of sleep onset and time of final awakening, but a lack of agreement between PSG and Zulu measures of TIB and SE. The underestimation of TIB by the Zulu watch, as well as the device's poor specificity for wake detection, most likely accounts for the overestimation of SE. Another factor which may contribute to the Zulu's underestimation of TIB is the lack of sleep onset latency (SOL) time spent in bed after final awakening (snooze time) measurement by the Zulu watch. In the current study, the amount of time that participants spent in bed was closely controlled by study staff. Therefore, TIB was a constant measure (480 minutes) across all nights of PSG data collection. The Zulu watch-estimated TIB was based on wrist activity, and was, in fact, very accurate for measuring sleep onset, i.e., the time at which the participant first fell asleep, as well as the time that the participant awoke in the morning. However, the watch does not provide an indication of how long it takes an individual to first fall asleep after getting into bed, which could account for the underestimation of TIB. To explore this possibility, SOL and snooze time were estimated by subtracting the Zulu-determined time of sleep onset from the time that participants were first put to bed by study staff (bedtime) and when they were told to get out of bed in the morning (wake time). Results were not statistically different between Zulu and PSG (all $p>0.48$). This finding indicates that shorter TIB as measured by Zulu is due to the fact that the watch does not measure SOL or snooze time.

Table 1: Comparison of Sleep Summary Statistics by PSG and Zulu Watch

	PSG	*Zulu*	*Mean Difference*	*Statistics*
Time In Bed (TIB)	480'±0'	456'±34'	-24'	t=3.43, p=0.001*
Total Sleep Time (TST)	408'±57'	414'±34'	+6'	t=0.45, p=0.65
Sleep Efficiency (SE)	85%±12%	91%±3%	+6%	t=2.29, p=0.03*
Time of sleep onset	22:02±0:31'	22:05±0:44'	+3'	t=0.22, p=0.83
Time of final awakening	05:47±0:33'	05:47±0:28'	0'	t=0.06, p=0.95

Conclusions

In conclusion, the Zulu watch showed mixed results but performed well in determining TST, sleep onset, final awakening, and deep/REM sleep in comparison to PSG in a sample of healthy young adults. Moreover, the watch achieves this accuracy in light of factors which could contribute to inaccuracy, such as long epoch lengths (2 minutes), on-wrist automatic scoring of sleep and single sensor (accelerometer) input. The Zulu also has positive off-wrist detection so that periods when the watch is not being worn are not mistakenly scored as sleep and can measure sleep episodes as short as 20 minutes occurring at any time of the day. It is noteworthy that the Zulu watch achieves comparability against PSG and can score sleep automatically on-wrist without requiring any intervention from the wearer or additional processing by a researcher or technologist.

References

Ancoli-Israel, S., Martin, J. L., Blackwell, T., Buenaver, L., Liu, L., Meltzer, L. J., . . . Taylor, D. J. D. J. (2015). The SBSM Guide to Actigraphy Monitoring: Clinical and Research Applications. *Behav Sleep Med, 13 Suppl 1*, S4-S38. doi:10.1080/15402002.2015.1046356

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., & Vaughn, B. V. (2012). The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine, 176*, 2012.

Bhat, S., Ferraris, A., Gupta, D., Mozafarian, M., DeBari, V. A., Gushway-Henry, N., . . . Chokroverty, S. (2015). Is There a Clinical Role For Smartphone Sleep Apps? Comparison of Sleep Cycle Detection by a Smartphone Application to Polysomnography. *J Clin Sleep Med, 11*(7), 709-715. doi:10.5664/jcsm.4840

de Zambotti, M., Cellini, N., Goldstone, A., Colrain, I. M., & Baker, F. C. (2019). Wearable Sleep Technology in Clinical and Research Settings. *Med Sci Sports Exerc, 51*(7), 1538-1557. doi:10.1249/MSS.0000000000001947

Liang, Z., & Chapa-Martell, M. A. (2019). Accuracy of Fitbit Wristbands in Measuring Sleep Stage Transitions and the Effect of User-Specific Factors. *JMIR Mhealth Uhealth, 7*(6), e13384. doi:10.2196/13384

Liang, Z., & Martell, M. A. C. (2018). Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *Journal of Healthcare Informatics Research, 2*(1-2), 152-178.

Marino, M., Li, Y., Rueschman, M. N., Winkelman, J. W., Ellenbogen, J. M., Solet, J. M., . . . Buxton, O. M. (2013). Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep, 36*(11), 1747-1755. doi:10.5665/sleep.3142

Paquet, J., Kawinska, A., & Carrier, J. (2007). Wake detection capacity of actigraphy during sleep. *Sleep, 30*(10), 1362-1369. doi:10.1093/sleep/30.10.1362