https://doi.org/10.5664/jcsm.9200

JCSM | Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Evaluation of the Consensus Sleep Diary in a community sample: comparison with single-channel electroencephalography, actigraphy, and retrospective questionnaire

Jessica R. Dietch, PhD[1]; Daniel J. Taylor, PhD[2]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, California; [2]Department of Psychology, University of Arizona, Tucson, Arizona

Study Objectives: The Consensus Sleep Diary (CSD) was developed by experts to promote standardization of sleep diary data across the field, but studies comparing the CSD with other assessments of sleep parameters are scarce. This study compared the CSD with 3 other methods to assess sleep duration, efficiency, and timing.

Methods: Participants (n = 80) were community adults (mean age = 32.65 years, 63% female) who completed the time-stamped CSD and used single-channel electroencephalography (EEG) and actigraphy for 7 days at home, then completed a retrospective sleep questionnaire. Total sleep time (TST), sleep efficiency (SE), and sleep midpoint were compared using correlations, Bland-Altman plots, and limits of agreement (adjusted for repeated measures).

Results: Correlations between the CSD and all methods on TST were large (*r*s = .63–.75). Adjusted CSD average TST was 40 minutes greater than with EEG and 31 minutes greater than with actigraphy. Correlations between CSD, actigraphy, and EEG for SE were small (*r*s = .18), and there was a medium correlation with questionnaire (*r* = .42). Adjusted CSD average SE was 7% greater than EEG and 6% greater than actigraphy; both demonstrated heteroscedasticity. Sleep midpoint correlations between CSD and all methods were large (*r* = .92–.99). Adjusted CSD was, on average, 6 minutes later than EEG and 1 minute later than actigraphy. Questionnaire-derived sleep parameters demonstrated nonconstant bias; lesser values had positive bias and greater values had negative bias.

Conclusions: The time-stamped CSD led to meaningful overestimations of TST and SE as measured by objective/inferred methods. However, sleep timing was rather accurately assessed with the CSD in comparison to objective/inferred measures. Researchers should carefully consider which sleep assessment methods are best aligned with their research question and parameters of interest, as methods do not demonstrate complete agreement.

Keywords: sleep diary, validation, EEG, actigraphy

Citation: Dietch JR, Taylor DJ. Evaluation of the Consensus Sleep Diary in a community sample: comparison with single-channel electroencephalography, actigraphy, and retrospective questionnaire. *J Clin Sleep Med*. 2021;17(7):1389–1399.

BRIEF SUMMARY

Current Knowledge/Study Rationale: The Consensus Sleep Diary (CSD) was originally developed by expert consensus to standardize the sleep diary tool across the field, but few studies have evaluated the CSD in comparison to other commonly used sleep assessment methods. The current study compared the CSD with electroencephalography, actigraphy, and retrospective sleep questionnaire measures of total sleep time, sleep efficiency, and sleep midpoint.

Study Impact: The CSD was most closely aligned with all other measures for sleep midpoint, followed by total sleep time, followed by sleep efficiency, and the CSD was in closer agreement with actigraphy compared with electroencephalography across all sleep parameters. It is difficult to compare questionnaire and CSD due to nonconstant bias and CSD is recommended over retrospective questionnaires for this reason.

INTRODUCTION

Poor sleep measured across many dimensions has been linked to many adverse physical and mental health outcomes.[1,2] However, the definition of "poor sleep" has historically been imprecisely defined and assessed (eg, a single-item retrospective question querying total sleep time [TST] or "sleep disturbance") in the literature.[2] Three primary domains of sleep thought to impact health and safety are duration, continuity (ie, efficiency), and timing.[1] Current research in health disciplines often relies upon brief, self-reported, inadequately validated methods to assess limited dimensions of sleep, which results in inaccurate measurements that impede and cloud scientific progress and discovery.[3] In order to better understand the relationship between sleep and health, precise and well-validated measurement strategies must be employed.

Prospective monitoring of sleep with daily sleep diaries is an essential component of self-reported sleep parameter assessment.[4] Sleep diaries offer several inherent advantages over single-time-point retrospective assessments, including ability to capture night-to-night variability in sleep and reduction in potential retrospective bias. Additionally, sleep diaries are standard tools in the treatment of sleep disorders with behavioral sleep medicine interventions.[5] However, until recently, many different versions of sleep diaries were used across published studies, which reduced comparability across studies and complicated examination of sleep diary validity. In an attempt to standardize sleep diary assessment, the Consensus

Sleep Diary (CSD)[6] was developed based on expert consensus and tested with patient focus groups. The CSD allows for extraction of the typical sleep parameters most often examined in the health literature, including those that describe sleep duration, efficiency, and timing.

Despite the rigorous development process for the CSD, it was not compared with other sleep assessment methods at the time of its development and few studies have done so since its inception. One prior study[7] compared CSD sleep parameters with those derived from actigraphy in a sample of individuals with (n = 37) and without (n = 37) insomnia and found CSD-derived TST and sleep efficiency (SE) were greater compared with actigraphy. Another prior study[8] compared CSD parameters with 2 single-time-point retrospective measures (ie, Pittsburgh Sleep Quality Index [PSQI], Self-Assessment of Sleep Survey [SASS] and SASS-split [SASS-Y]). This study found CSD-derived TST was greater compared with the PSQI and SASS, and lesser compared with the SASS-Y. SE was greater on the CSD compared with all 3 questionnaire measures. Another study[9] compared the CSD with actigraphy and a variety of wearable sleep trackers and found that CSD-derived TST was generally greater compared with actigraphy and most activity trackers. CSD-derived SE was fairly similar to actigraphy and mixed in regard to other activity trackers. Gaps in the literature still remain, namely (1) no study has compared the CSD with other assessment methods of sleep timing and (2) no study has compared the CSD with electroencephalography (EEG)- or polysomnography (PSG)-derived sleep parameters.

It is crucial for researchers and clinicians to understand how sleep parameters derived from different sleep assessment methods relate to one another in order to facilitate comparison and contextualization of findings across studies. In the current study, we sought to compare daily-collected CSD parameters of sleep duration (ie, TST), continuity (ie, efficiency, or the percentage of intended sleep time that is filled with sleep), and timing (ie, sleep midpoint) with single-channel EEG, actigraphy, and retrospective questionnaire across 7 nights in a community sample.

METHODS

Participants

Participants were recruited broadly in the Denton County, Texas area. Recruitment efforts aimed to increase generalizability by including a wide age range and a diverse racial/ethnic breakdown similar to that in the community. Recruitment materials directed interested individuals to complete an online informed consent and a brief screening questionnaire that collected contact information and assessed the following inclusion criteria: (1) English-language fluency, (2) over the age of 18, (3) had a phone number at which they could be regularly reached, and (4) had regular (daily) internet and personal email access. Exclusion criteria included a pacemaker, cardiac defibrillator, or other medical electronic device due to interference with using the single-channel EEG device. Data were collected from February 2017 through August 2017.

Procedures

All procedures were approved by the University of North Texas Institutional Review Board prior to the start of data collection. After completing the brief screening measure, eligible participants were contacted and asked to complete baseline measures online at home via a secure online data-collection tool (REDCap).[10] Participants were scheduled for an initial in-person appointment in the sleep laboratory during which they were trained in the use of the EEG device via videos provided by the equipment manufacturer and live demonstration. Participants were also trained in the use of actigraphy via verbal instruction from the research assistants and live demonstration. Participants were trained in the use of daily CSD via a sample questionnaire sent to their internet-enabled device and hands-on demonstration with a research assistant. Participants were then given an EEG device, actigraph, and written instructions for all items. Finally, participants' understanding of study procedures was verified via a brief written quiz and their second in-laboratory appointment was scheduled.

Participants used the time-stamped CSD, EEG device, and actigraph in their typical sleep environment for 7 days. After the first study night, research assistants securely messaged participants to ensure that equipment worked properly and check if they had any questions. At the first laboratory visit, participants and research assistants mutually chose a time to receive the first CSD reminder each day of the study (shortly after estimated awakening time each day). Participants received up to 2 additional reminders, at 3-hour intervals, if they did not complete the CSD. Additionally, if they had not completed the CSD by noon, research assistants messaged the participants to remind them to complete it. After 7 days of data collection, participants returned to the laboratory to return the equipment and complete final measures including retrospective sleep questionnaires (including the PSQI) and a structured interview for sleep disorders. The compensation offered for participation in the study was $20, a comprehensive report of the participant's sleep over the study duration and sleep disorders resources, and a decorative refrigerator magnet.

Measures

Demographics and baseline sleep characteristics

In order to describe the sample, participants were asked to report demographic and sleep characteristics using well-validated questionnaires. Insomnia status was assessed with the Structured Clinical Interview for Sleep Disorders–Revised (SCISD-R).[11,12]

Sleep diary

The CSD–Core version was used in the current study.[6] The CSD is a self-reported measure typically administered as soon as possible after the end of the major sleep period (eg, in the morning upon awakening), which asks participants to give an estimate of their sleep on the previous night (eg, bedtime, sleep onset latency, wake time). These variables allow for the calculation of additional sleep parameters (eg, TST, SE, sleep midpoint). The CSD included in the current study was an online version that was collected via electronic data capture software

(REDCap) and was therefore time-stamped.[10] The CSD has been validated for online use.[13] In general, sleep diaries generally correlate moderately with both PSG and actigraphy (eg, .33–.71).[6]

Single-channel EEG

The Zmachine (General Sleep Corporation, Cleveland, OH) Insight Plus is an ambulatory device that processes a single channel of EEG data using information from 2 mastoid-placed electrodes and 1 neck-placed ground electrode. The Zmachine is capable of differentiating between wake, light sleep (stages N1 and N2), deep sleep (stage N3), and rapid eye movement sleep.[14] The Zmachine electrodes are single-use and participants were instructed to self-apply them 30 or more minutes prior to bedtime per the manufacturer's instructions. In the original validation study, compared with full PSG rated by 2–4 scorers, the Zmachine demonstrated 96% sensitivity and 93% specificity for sleep-wake detection.[14] Correlations between PSG and Zmachine for sleep parameters were high: TST, $r = .95$; SE, $r = .93$; sleep onset latency, $r = .96$; and wake after sleep onset, $r = .89$.[14] In the current study, data were visually examined using the Zmachine Data Viewer software v3.5.0 and poor-quality data (eg, missing over 1 hour of recording time, illogical values) was removed. Firmware on the Zmachine Insight Plus devices was version 5.0.

Actigraphy

Actigraphs are wrist-worn, wristwatch-like devices that use an accelerometer to capture motion as a proxy for activity. Computer software uses an algorithm to analyze activity and estimate sleep parameters such as TST, sleep onset latency, number of awakenings, wake after sleep onset, and terminal wakefulness.[15] In the current study, the actigraphs used were Philips Respironics (Philips Respironics, Bend, OR) Actiwatch Spectrum devices, and data were analyzed with Respironics Actiware version 6.0. Data were scored by 2 trained scorers using an internally developed, publicly available scoring hierarchy,[16] and discrepancies were resolved by a third scorer. In brief, this scoring hierarchy prioritizes event markers (assuming congruence with sleep diary and activity/light patterns), then sleep diaries, then activity/light patterns, for making decisions about setting intervals. Settings used for data export in Actiware were the following: low threshold (activity count: 10), 20 epochs inactivity for sleep onset/offset.

Retrospective sleep questionnaire

The PSQI[17] is a 19-item self-rated questionnaire designed to measure 7 domains of sleep. Domain scores range from 0 (no difficulty) to 3 (severe difficulty). In the current study, only questions that queried average TST, SE, and variables used in the assessment of sleep midpoint over the past week were included for analyses.

Data analysis

The following sleep parameters were computed across all 4 assessment methods when possible: sleep midpoint (bedtime – wake time/2), TST (time in bed – total wake time [sleep onset latency + wake after sleep onset + terminal wakefulness]), and SE (TST/time in bed × 100). For the CSD, time in bed is calculated at the time between lights out ("What time did you try to go to sleep?") and rise time ("What time did you get out of bed for the day?") For comparison to questionnaire method, data were averaged across the week for CSD. Averages for CSD were only calculated if ≥ 5 days of data existed for an individual on that measure.

In order to compare agreement between CSD and actigraphy, retrospective questionnaire, and EEG assessments of sleep timing, duration, and efficiency, parameters were compared across measurement method using Pearson correlations and the Bland and Altman/limits of agreement technique to examine systematic bias and agreement.[18] One deficit in the sleep measure validation literature is the erroneous use of product-moment correlation coefficients (r) and other global indices to demonstrate agreement between 2 measures; it is inappropriate to assess agreement between measures using solely correlation, regression, comparison of means, structural equations, or intraclass correlation methods, which signify association rather than agreement.[19] Instead, Bland and Altman[19] recommend examining plots of 2 methods' means against mean differences (Bland-Altman plots) and estimates of where 95% of differences between measures are expected to fall (limits of agreement). These items give information as to potential systematic bias and variability of estimates in addition to mean differences.

Bland-Altman analyses were conducted in R software version 3.1.3 (R Foundation for Statistical Computing, Vienna, Austria)[20] using the using the MethComp package version 1.22.2.[21] Separate analyses were used for each sleep parameter and each measure comparison. Data were first examined via 2 plots: (1) prediction plots, gold standard on the x-axis against comparison method on the y-axis with a line of equality (ie, perfect agreement between the methods), and (2) Bland-Altman plots, the mean of both methods ([method 1 + method 2]/2) on the x-axis against difference between methods (method 1 – method 2) on the y-axis. These plots allow for visual examination of agreement between methods and detection of systematic or unsystematic bias. For comparison between CSD with EEG and actigraphy (repeated measures) a mixed model was used to estimate the 95% limits of agreement while controlling for nesting of repetitions within participants (data were considered "linked" or paired replicates).[22] This method includes measure (ie, EEG, actigraphy, CSD) and participant as fixed effects and the measure × participant interaction as a random effect.

For comparison between the CSD and retrospective sleep questionnaire, traditional calculations for Bland-Altman plots were created and visually examined.[19] Bland-Altman plots were examined and found to demonstrate nonconstant bias (ie, significant slope; all $P < .05$). Given nonconstant bias, traditional limits of agreement could not be calculated,[23] nor can an average bias be meaningfully interpreted. Therefore, plots were recomputed allowing differences to depend on averages in a linear rather than constant fashion per recommendations by Carstensen.[23] Using the DA.reg function of the MethComp R package,[21] coefficients were calculated that can be used to convert 1 method to another and prediction intervals were used rather than traditional limits of agreement. Conversion from

questionnaire to the other method can be achieved using the following formula[23]:

$$y_{2|1} = \alpha_{2|1} + \beta_{2|1} y_1 \pm 2 \times \mathrm{SD}\left(y_{2|1}\right).$$

Power analysis

Carstensen[23] argues that power analysis calculations for measure comparison studies are irrelevant and instead recommends a sample of at least 50, with at least 3 days of measurement per person. Data collected far exceed that recommendation, even given missing data.

RESULTS

Data were cleaned by examining and applying necessary corrections for outliers, variable normality, and missing data in accordance with recommendations from Tabachnick and Fidell.[24] Initially, 120 people expressed interest in the study, 101 completed the screening questionnaire, and 87 completed the baseline questionnaire. A total of 81 participants attended the first laboratory appointment and completed some measures and a final n = 80 were included in any analyses, resulting in a total of 560 measurement opportunities. Participants were considered "completers" and therefore included for analysis if they had ≥ 5 days of usable data on at least 2 measures; 1 participant was excluded for noncompletion. For CSD, 6 days were missing across the entire study, but no participants were removed from analyses (n = 80; all had ≥ 5 days of sleep diary data). For the EEG, 65 days were excluded for bad data and 19 days were missing, for a total of 84 excluded days (due to the nature of the data errors, sleep midpoint was retained for 19 days with bad data). A total of 11 participants had data removed for EEG due to < 5 days of usable data (n = 69 with complete data). For actigraphy, 21 days were excluded for bad data and 15 days were missing, for a total of 36 days missing (sleep midpoint retained for 15 days). A total of 5 participants had data removed for actigraphy due to < 5 days of usable data (n = 75 with complete data). For the questionnaire, 1 participant's data were completely missing and 2 additional participants' data were removed for SE due to impossible values given (n = 77 for all questionnaire variables). Participant characteristics are presented in **Table 1**. The majority of participants were female, non-Hispanic White, married or in a relationship, well-educated, and employed.

Unadjusted means and correlations

Unadjusted means and standard deviations of sleep variables of interest are presented in **Table 2**. The CSD demonstrated the greatest TST and SE and latest sleep midpoint compared with all other measures. Correlations (Pearson *r*) were conducted to examine relationships between each variable of interest (TST, SE, sleep midpoint) as measured by the CSD, EEG, actigraphy, and questionnaire (**Table 3**). For TST, correlations between the CSD and all other measures were large. For SE, correlations between the CSD, actigraphy, and EEG were small and correlations between the CSD and questionnaire were medium. For sleep midpoint, correlations between the CSD and all other measures were very large.

Table 1—Participant demographic and psychosocial characteristics.

	Values
Age, mean (SD); range, y	32.7 (10.1); 19–69
Sex	
Male	30 (37.5)
Female	50 (62.5)
Race/ethnicity	
NH Black	3 (3.8)
NH White	68 (85.0)
Asian	4 (5.0)
Biracial/other	5 (6.2)
Married/committed relationship	
Yes	58 (72.5)
No	22 (27.5)
Educational attainment	
High school or less	2 (2.5)
≤ 4 years post–high school education	36 (45.0)
> 4 years post–high school education	42 (52.5)
Employment status	
Full time	49 (61.3)
Part time	19 (23.8)
Retired/unemployed	12 (15.1)
Insomnia diagnosis by clinical interview	
Yes	24 (30)
No	56 (70)

Values are presented as n (%) except for age. NH = non-Hispanic, SD = standard deviation.

Bland-Altman plots and limits of agreement: EEG and actigraphy

Bland-Altman plots of EEG, actigraphy, and CSD for TST, SE, and sleep midpoint are presented in **Figure 1**. These plots were visually examined and found to demonstrate constant bias (ie, homoscedasticity of differences) at all levels, so slope was fixed to 1 and bias and limits of agreement were calculated with mixed-model adjustments for repeated measures (see **Table 4**).[22] The 95% limits of agreement are displayed as the thin blue outer horizontal lines on these plots, signifying that 95% of differences between measures are expected to fall between these lines. Adjusted mean difference, or bias, is represented as a single, thicker blue horizontal line on these plots, with positive values indicating greater values for the first measure compared with the second and negative values indicating lower values for the first measure compared with the second. The values reported in the following sections reflected means adjusted for repeated measures.

CSD TST was, on average, 40.2 minutes greater than EEG and demonstrated wide limits of agreement (4.1 hours). CSD TST was, on average, 31.2 minutes greater than actigraphy and demonstrated wide limits of agreement (3.7 hours)

CSD SE was, on average, 7.1% greater than EEG, with wide limits of agreement (46.7%). Notably, the comparison for SE

Table 2—Unadjusted means (SD) for Consensus Sleep Diary, single-channel electroencephalography (EEG), actigraphy, and questionnaire total sleep time, sleep efficiency, and sleep midpoint.

	Mean (SD)	n
Total sleep time, h		
CSD	7.08 (0.86)	80
EEG	6.14 (0.75)	69
Actigraphy	6.35 (0.84)	75
Questionnaire	6.89 (1.20)	79
Sleep efficiency, %		
CSD	90.37 (5.59)	80
EEG	81.22 (6.59)	69
Actigraphy	81.59 (5.71)	75
Questionnaire	84.69 (11.63)	77
Sleep midpoint, t (min)		
CSD	3:46 (73.20)	80
EEG	3:37 (75.0)	72
Actigraphy	3:43 (71.40)	78
Questionnaire	3:11 (82.20)	79

Values are unadjusted means (SD) for CSD, single-channel EEG, actigraphy, and questionnaire total sleep time, sleep efficiency, and sleep midpoint. CSD = Consensus Sleep Diary, EEG = single-channel electroencephalography, SD = standard deviation, t = clock time.

Table 3—Correlations among CSD, single-channel EEG, actigraphy, and questionnaire total sleep time, sleep efficiency, and sleep midpoint across 7 nights at home.

	CSD	EEG	Actigraphy
Total sleep time			
EEG	.71***	—	
Actigraphy	.75***	.75***	—
Questionnaire	.63***	.57***	.68***
Sleep efficiency			
EEG	.18	—	
Actigraphy	.18	.25*	—
Questionnaire	.42***	.15	.38**
Sleep midpoint			
EEG	.98***	—	
Actigraphy	.99***	.99***	—
Questionnaire	.92***	.94***	.93***

*P < .05, **P < .01, ***P < .001. CSD = Consensus Sleep Diary, EEG = single-channel electroencephalography.

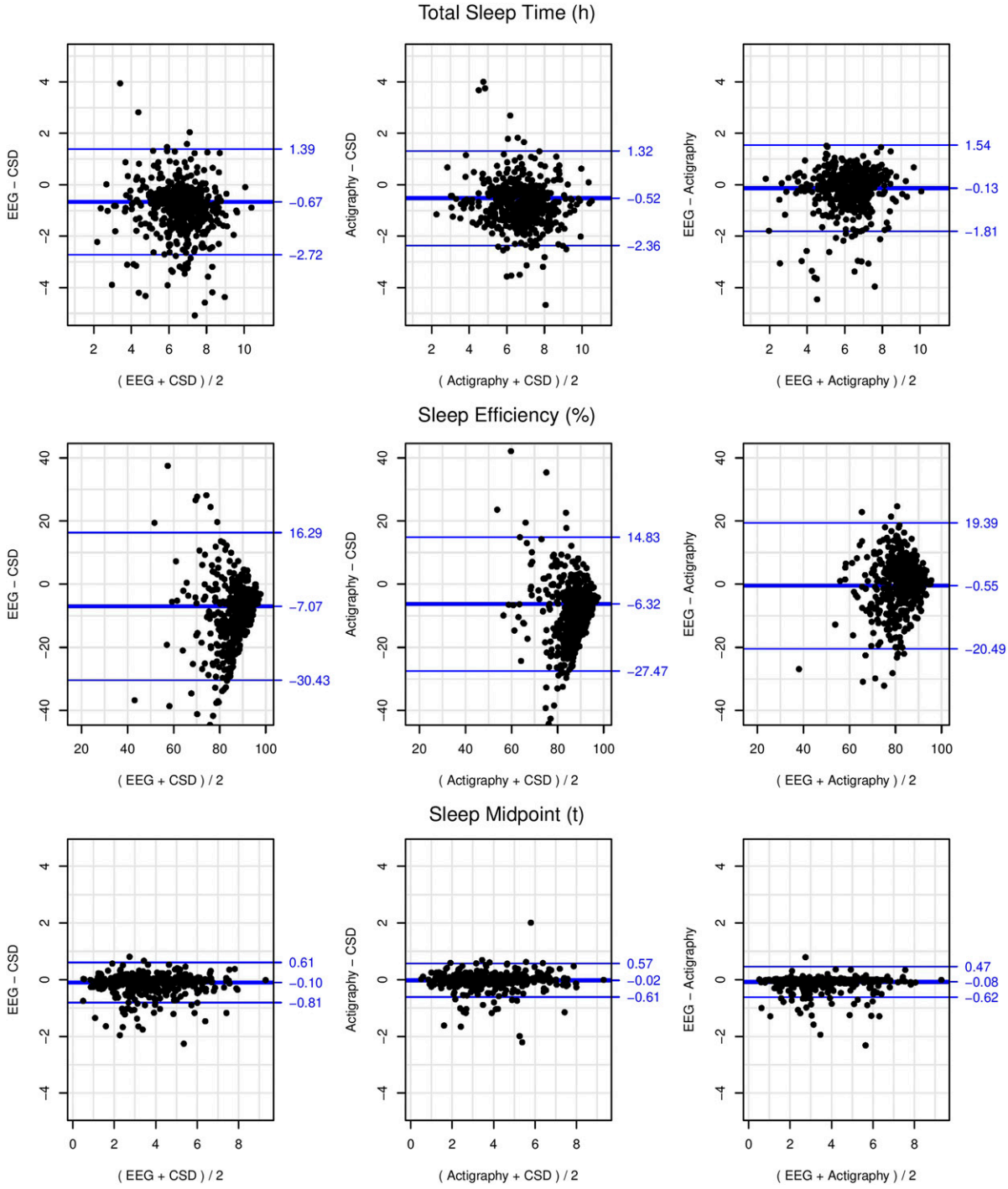
demonstrated substantial heteroscedasticity, with greater variability in differences at lesser average values of SE (see Figure 1). In other words, days with greater SE (average of CSD and EEG) were more precisely measured by CSD with respect to EEG. CSD SE was, on average, 6.3% greater than actigraphy, with wide limits of agreement (42.3%). A similar pattern of heteroscedasticity was seen with actigraphy compared with CSD.

CSD sleep midpoint was, on average, 6.0 minutes later than EEG and limits of agreement were moderate (1.4 hours). CSD sleep midpoint was, on average, 1.2 minutes later than actigraphy and limits of agreement were moderate (1.2 hours).

Bland-Altman plots and prediction interval: questionnaire

Coefficients that can be used to convert 1 method to another, standard deviation of the prediction (precision), and 95% limits of prediction are presented in Table 5. Plots of differences between questionnaire and other measures are presented in Figure 2. For TST, the prediction interval for the comparison between the CSD and questionnaire was 2.8 hours. Additionally, the slope indicated that differences were not the same for individuals with high and low TST. For individuals with longer average TST (average of both measure), the questionnaire was greater compared with the CSD. For individuals with shorter average TST, the questionnaire was lesser compared with the CSD.

Figure 1—Bland-Altman plots adjusted for repeated measures (7 days) of single-channel EEG-, actigraphy-, and CSD-assessed total sleep time, sleep efficiency, and sleep midpoint with diagonal blue lines indicating mean bias (thicker middle line) and 95% limits of agreement (thinner outer lines).



CSD = Consensus Sleep Diary, EEG = electroencephalography.

For SE, the prediction interval for the comparison between the CSD and questionnaire was 21.6%. The slope indicated, similar to TST, that differences were not the same for individuals with high and low SE. For individuals with greater average SE (average of both measures), the questionnaire was greater compared with the CSD. For individuals with lesser average SE, the questionnaire was lesser compared with the CSD.

For sleep midpoint, the prediction interval for the comparison between the CSD and questionnaire was 1.9 hours. The slope indicated, similar to TST and SE, that differences were not the same for individuals with early and late sleep midpoint.

Table 4—Bias, precision, and 95% limits of agreement (adjusted for repeated measures) for 7 days of total sleep time, sleep efficiency, and sleep midpoint as compared between single-channel EEG, actigraphy, and diary.

	α	σ	95% LoA Lower	95% LoA Upper
Total sleep time, h				
EEG vs CSD	−0.67	1.03	−2.72	1.39
Actigraphy vs CSD	−0.52	0.92	−2.36	1.32
EEG vs actigraphy	−0.13	0.84	−1.81	1.54
Sleep efficiency, %				
EEG vs CSD	−7.07	11.68	−30.43	16.29
Actigraphy vs CSD	−6.32	10.58	−27.47	14.83
EEG vs actigraphy	−0.55	9.97	−20.49	19.39
Sleep midpoint, t				
EEG vs CSD	−0.10	0.35	−0.81	0.61
Actigraphy vs CSD	−0.02	0.30	−0.61	0.57
EEG vs actigraphy	−0.08	0.27	−0.62	0.47

CSD = Consensus Sleep Diary, EEG = single-channel electroencephalography, LoA = limit of agreement, t = time, α = bias, σ = precision.

Table 5—Intercepts, slopes, standard deviation of prediction, and 95% prediction intervals for averaged-over-week total sleep time, sleep efficiency, and sleep midpoint as measured by single-channel EEG, actigraphy, and diary compared with questionnaire.

	α	β	σ	95% p.i.
Total sleep time, h				
CSD vs questionnaire	2.42	0.68	0.73	±1.42
EEG vs questionnaire	2.04	0.60	0.67	±1.31
Actigraphy vs questionnaire	1.68	0.67	0.66	±1.30
Sleep efficiency, %				
CSD vs questionnaire	59.41	0.36	5.50	±10.79
EEG vs questionnaire	43.92	0.43	7.52	±14.74
Actigraphy vs questionnaire	48.69	0.38	5.69	±11.15
Sleep midpoint, t				
CSD vs questionnaire	0.92	0.89	0.49	±0.95
EEG vs questionnaire	0.79	0.88	0.44	±0.85
Actigraphy vs questionnaire	0.95	0.87	0.45	±0.89

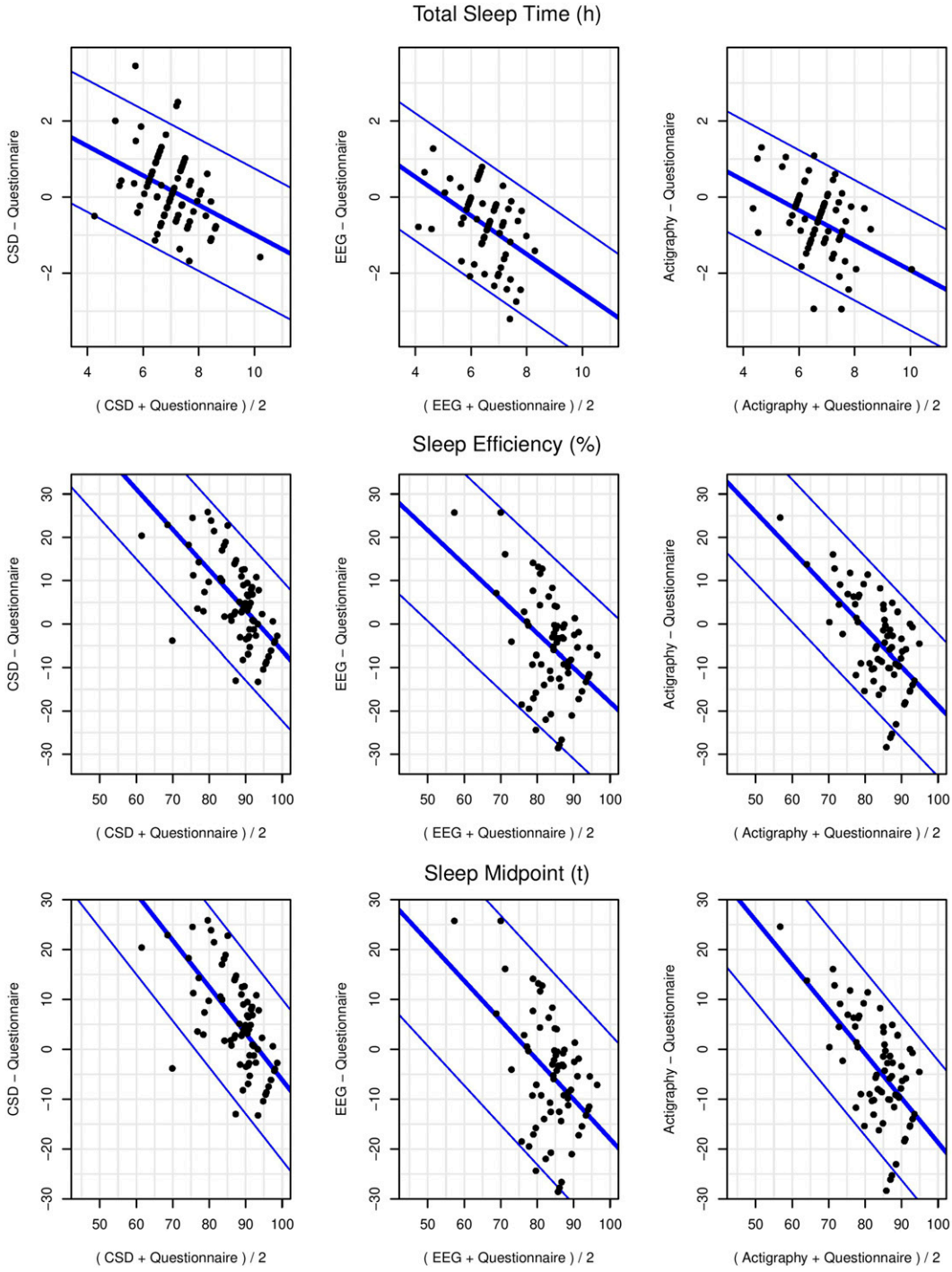
CSD = Consensus Sleep Diary, EEG = single-channel electroencephalography device, p.i. = prediction interval, t = time, α = intercept, β = slope, σ = standard deviation.

For individuals with later average sleep midpoint (average of both measures), the questionnaire was later compared with the CSD. For individuals with earlier average sleep midpoint, the questionnaire was earlier compared with the CSD.

DISCUSSION

The primary objective of the current study was to compare measures of TST, SE, and sleep midpoint as assessed by CSD with assessments by single-channel EEG, actigraphy, and retrospective questionnaire in a community sample in the naturalistic sleep environment. We do not consider this to be a "validation" study per se, as each of these methods measure distinct, yet overlapping constructs of sleep. However, the findings may be interpreted as evidence for or against the validity of the CSD to assess a given sleep parameter in comparison to a given assessment method. In general, the CSD was in closer agreement with actigraphy compared with EEG across sleep parameters. It is difficult to make a comparison for questionnaire due to the nonconstant bias compared with all 3 other measures. Based on the overall picture of the correlations, Bland-Altman plots, and limits of agreement, the CSD was most closely aligned with other measures for sleep midpoint, followed by TST, followed by SE. In sum, the time-stamped CSD leads to meaningful overestimations of objective/inferred measurements of TST and SE. However,

Figure 2—Bland-Altman plots for averaged-across-week single-channel EEG-, actigraphy-, CSD-, and questionnaire-assessed total sleep time, sleep efficiency, and sleep midpoint with slopes allowed to vary linearly and with diagonal blue lines indicating mean bias (thicker middle line) and 95% limits of agreement (thinner outer lines).



CSD = Consensus Sleep Diary, EEG = electroencephalography.

sleep timing is rather accurately assessed with the CSD in comparison to objective/inferred measures.

Total sleep time

CSD-assessed TST was 40 minutes longer than EEG, on average, after adjusting for repeated measures. This was similar to findings of a study that found that sleep diary (not CSD) overestimated TST compared with PSG by 50 minutes (night 1) and 28 minutes (night 2) in a group of n = 27 adults with bipolar disorder.[25] Matthews et al[26] found that sleep diary (not CSD) overestimated TST compared with PSG by 15 minutes, on average, across 2 nights in n = 223 middle-aged and older adults.

In contrast, McCall and McCall[27] found that sleep diary (not CSD) was 55 minutes shorter, on average, than PSG in a sample of n = 54 young, middle-aged, and older adults with insomnia and depression. These differences may be attributable to differences in sleep diary format, as none of the aforementioned studies used the time-stamped CSD or single-channel EEG, and population differences. In the current study, limits of agreement were broad (ie, 95% of differences fell within a range of > 4 hours). McCall and McCall[27] found even broader 95% limits of agreement (6.8 hours). Given current and previous studies demonstrating that sleep diary substantially overestimated TST compared with PSG/EEG, and produced broad limits of agreement, it is reasonable to suggest that sleep diaries (including CSD) and PSG/EEG are not interchangeable methods for measuring TST. Instead, self-reported TST may reflect an overlapping, yet distinct construct from objectively assessed TST and should be treated as such.

Consistent with prior research,[7,9] actigraphy-derived TST was shorter than TST derived from time-stamped CSD. However, unlike those prior studies with discrepancies ranging from 70 to 124 minutes, actigraphy and CSD TST only differed by 31 minutes, on average (adjusted). One prior study of individuals with insomnia collected on 1 night found minimal differences between TST derived from PSG, actigraphy, and sleep diary (not CSD), which is unsurprising given the highly controlled nature of the study (ie, a single night in the sleep laboratory).[28] Given the lack of consistency in reporting and high degree of variability in rest-interval setting procedures, devices, and algorithm choices/export settings, direct comparison is difficult. It is possible that the smaller discrepancy in measures in the current study is attributable to the rest-interval setting procedure used by our group, which uses sleep diary information to inform this process,[16] and the time-stamped nature of the CSD used in the current study, which encouraged close adherence. Regardless, the finding in the current study that limits of agreement were broad suggests a high degree of variability in the differences between these measures in individual cases. Thus, for a given individual, the accuracy of actigraphy may be highly variable.

On average, similar to 1 prior study,[8] PSQI questionnaire–derived TST was approximately 11 minutes lower than CSD (unadjusted; no adjustment for repeated measures was used since the PSQI is only measured at a single time point). This is unsurprising given that both measures assess self-reported perception of sleep parameters. However, the questionnaire TST demonstrated nonconstant bias, such that lesser values of TST had positive bias (questionnaire was lesser than CSD), and the reverse was true for greater values of TST (questionnaire was greater than CSD). Additionally, precision was poor for CSD vs questionnaire comparisons of TST. The current findings shed light on the potential for bias in a questionnaire depending on level of the parameter. Given this bias, the current study confirms that a daily sleep diary is the preferred self-reported measure compared with questionnaire, given that a sleep diary does not result in substantial nonconstant bias compared with actigraphy and EEG, which is difficult to adjust for either clinically or statistically.

Sleep efficiency

CSD overestimated SE compared with EEG by 7%, on average (adjusted). Two existing studies demonstrated minimal differences between sleep diary–derived (not CSD) and PSG-derived SE.[25,28] Consistent with 1 prior study, CSD SE was also fairly similar to actigraphy, overestimating by approximately 6% (adjusted).[9] Notably, for both EEG and actigraphy, comparisons demonstrated substantial heteroscedasticity, indicating that, for lower values of SE, the differences between measures were more variable. In other words, CSD estimation of SE was more precise compared with EEG and actigraphy for nights with greater values of SE. This is consistent with prior research demonstrating that individuals with insomnia, a disorder of reduced SE, report greater discrepancies between diary/self-report sleep parameters compared with good sleepers or those who have been treated for insomnia.[29–32]

PSQI questionnaire–derived SE demonstrated nonconstant bias, such that lesser values of SE had positive bias (questionnaire was lesser than CSD), and the reverse was true for greater values of SE (questionnaire was greater than CSD). Additionally, precision was poor for CSD vs questionnaire comparisons of SE. On average, SE was quite similar between questionnaire and CSD; however, the nonconstant bias and broad prediction interval suggest caution should be used in interpreting questionnaire-derived SE.

Sleep midpoint

CSD-assessed sleep midpoint was quite similar, on average, to both actigraphy and EEG. Questionnaire sleep midpoint demonstrated slight nonconstant bias, such that earlier sleep midpoints had positive bias (questionnaire was earlier than CSD) and the reverse was true for later sleep midpoints (questionnaire was later than CSD). Precision was acceptable for all methods. To our knowledge, this is the first study to compare accuracy across measures of sleep midpoint. The close agreement between measures of sleep midpoint is likely attributable, in part, to the attention that individuals pay to bedtime and wake time and the proximity of these events to periods of sustained wakefulness. Because sleep is an inherently amnesiac state, memory for nighttime periods of wakefulness (eg, sleep onset latency, wake after sleep onset) that factor into the calculation of TST and SE are likely to be recalled less accurately. Further, setting alarms for morning wake-up provides a time stamp for bedtime and an anchor point for wake time, which can improve recall ability. The stability of sleep midpoint for CSD suggests that it is a good proxy for actigraphy- or EEG-assessed sleep timing, whereas a questionnaire is slightly biased and should be used with caution.

Strengths and limitations

To our knowledge, this study was the first in the field of sleep measurement to adjust for repeated measures using a mixed-model approach. Few studies examined Bland-Altman plots or limits of agreement to assess accuracy of measures, and among those that did, none adjusted for repeated measures. Adjusting for repeated measures improves confidence in the accuracy of the limits of agreement calculated (without

adjustments for repeated measures, limits of agreement in the current study were calculated to be much wider). Further, this study extends the scarce literature comparing CSD with other sleep-assessment tools. This is a crucial area of study, as many studies present self-reported measures of sleep parameters as substitutes for objective measures when they appear to reflect different constructs. Finally, to our knowledge, this was the first study to examine accuracy of sleep measures for sleep midpoint.

The current study was not without limitations. First, although this sample of participants was drawn from the community, the sample demographics do not reflect the larger community and substantially limit the generalizability of results. In particular, this was a highly educated and largely non-Hispanic White convenience sample, so results may not apply to individuals with less education or with different racial/ethnic identities. Second, use of a single-channel EEG device as a gold standard sacrifices some accuracy in assessing the true value of the measured construct in trade for greater ecological validity. The exact impact of this trade-off cannot be assessed within the current study but use of this device limits comparison between CSD and full PSG. The results here cannot be generalized to studies conducted in a laboratory environment. Third, it is crucial to consider the way in which the CSD was delivered in the current study when generalizing the results. The method used in the current study, namely digital time-stamped administration of the CSD that was closely monitored by research assistants, likely yielded close to the maximum potential accuracy of the CSD. Using methods without this close attention to measure completion and participants' awareness of time-stamping may attenuate the accuracy of the CSD. Finally, retrospective questionnaires were collected after 1 week of careful attention to sleep via a diary, which likely increased the accuracy of questionnaire measures compared with typical use. Similarly, the version of the PSQI that was used as a retrospective questionnaire in the current study queried participants on their sleep in the past week, rather than the typical use of the PSQI to assess sleep across the past month. This may have also artificially increased the accuracy of the questionnaire measure.

Implications and future directions

The results of this study highlight the importance of careful attention to measurement method, particularly when assessing TST and SE as accuracy varies across method and outcome. Emerging research suggests that self-reported aspects of sleep like "depth" or "restfulness" are not closely related to objective sleep assessments.[33] In other words, self-assessment of sleep may not always closely reflect its underlying biological function.[33] However, both objective and self-reported assessments of sleep are important variables to consider depending on the research question, sleep parameters of interest, and outcome variables.

These results demonstrate the importance of using statistical techniques that describe agreement (eg, Bland-Altman plots) in addition to those that describe relationship (eg, correlation). Although, on average, the CSD demonstrated acceptable agreement with TST and SE compared with EEG and actigraphy, poor precision suggests individual days or participants may not demonstrate acceptable accuracy. Use of these methods also revealed retrospective questionnaire-derived sleep parameters were universally biased compared with other measures. On average, days with shorter TST, lesser SE, and earlier sleep midpoint were fewer than determined by other methods, and the reverse was true for greater values of each parameter. We recommend that brief retrospective questionnaires of sleep parameters should be replaced with other methods (eg, SASS/SASS-Y,[8] CSD), when possible, to facilitate accurate interpretation.

Future studies should examine potential predictors of discrepancies between sleep measures. Future studies should continue to explore measurement validity in the naturalistic sleep environment, as much current validation research focuses on a laboratory environment, whereas a substantial portion of sleep and health research occurs in the home environment. In particular, individuals from low-socioeconomic-status backgrounds may be more likely to have discrepancies between their typical sleep environment and the laboratory environment. Thus, it is important to continue this work in individuals with diverse characteristics, particularly on known sources of sleep disparities such as socioeconomic status, race/ethnicity, and sex/gender.

ABBREVIATIONS

CSD, Consensus Sleep Diary
EEG, electroencephalography
PSG, polysomnography
PSQI, Pittsburgh Sleep Quality Index
SASS, Self-Assessment of Sleep Survey
SE, sleep efficiency
TST, total sleep time

REFERENCES

1. Czeisler CA. Duration, timing and quality of sleep are each vital for health, performance and safety. *Sleep Health*. 2015;1(1):5–8.
2. Buysse DJ. Sleep health: can we define it? Does it matter? *Sleep*. 2014;37(1):9–17.
3. Grandner MA, Hale L, Moore M, Patel NP. Mortality associated with short sleep duration: the evidence, the possible mechanisms, and the future. *Sleep Med Rev*. 2010;14(3):191–203.
4. Buysse DJ, Ancoli-Israel S, Edinger JD, Lichstein KL, Morin CM. Recommendations for a standard research assessment of insomnia. *Sleep*. 2006; 29(9):1155–1173.
5. Hartmann JA, Carney CE, Lachowski A, Edinger JD. Exploring the construct of subjective sleep quality in patients with insomnia. *J Clin Psychiatry*. 2015;76(6): e768–e773.
6. Carney CE, Buysse DJ, Ancoli-Israel S, et al.. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep*. 2012;35(2):287–302.
7. Maich KHG, Lachowski AM, Carney CE. Psychometric properties of the Consensus Sleep Diary in those with insomnia disorder. *Behav Sleep Med*. 2018; 16(2):117–134.
8. Dietch JR, Sethi K, Slavish DC, Taylor DJ. Validity of two retrospective questionnaire versions of the Consensus Sleep Diary: the whole week and split week Self-Assessment of Sleep Surveys. *Sleep Med*. 2019;63:127–136.
9. Lee J-M, Byun W, Keill A, Dinkel D, Seo Y. Comparison of wearable trackers' ability to estimate sleep. *Int J Environ Res Public Health*. 2018;15(6):1265.

10. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377–381.
11. Taylor DJ, Wilkerson AK, Pruiksma KE, et al; STRONG STAR Consortium. Reliability of the structured clinical interview for DSM-5 Sleep Disorders Module. *J Clin Sleep Med*. 2018;14(3):459–464.
12. Taylor DJ, Wilkerson AK, Pruiksma KE, Dietch JR, Wardle-Pinkston S. Structured Clinical Interview for Sleep Disorders–Revised (SCISD-R). 2019. https://insomnia.arizona.edu/SCISD. Accessed May 1, 2020.
13. Tonetti L, Mingozzi R, Natale V. Comparison between paper and electronic sleep diary. *Biol Rhythm Res*. 2016;47(5):743–753.
14. Kaplan RF, Wang Y, Loparo KA, Kelly MR, Bootzin RR. Performance evaluation of an automated single-channel sleep-wake detection algorithm. *Nat Sci Sleep*. 2014;6:113–122.
15. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*. 2003;26(3):342–392.
16. Rijsketic J, Dietch J, Wardle-Pinkston S, Taylor D. Actigraphy (Actiware) Scoring Hierarchy Manual. 2020. https://insomnia.arizona.edu/actigraphy. Accessed May 1, 2020.
17. Buysse DJ, Reynolds CF 3rd, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Res*. 1989;28(2):193–213.
18. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Int J Nurs Stud*. 2010;47(8):931–936.
19. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–160.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2015.
21. MethComp: Functions for Analysis of Agreement in Method Comparison Studies [computer program]. Version 1.22. 22016, https://cran.r-project.org/web/packages/MethComp/index.html
22. Carstensen B, Simpson J, Gurrin LC. Statistical models for assessing agreement in method comparison studies with replicate measurements. *Int J Biostat*. 2008;4(1):16.
23. Carstensen B. Comparing methods of measurement: extending the LoA by regression. *Stat Med*. 2010;29(3):401–410.
24. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 6th ed. Boston, MA: Pearson; 2013.
25. Kaplan KA, Talbot LS, Gruber J, Harvey AG. Evaluating sleep in bipolar disorder: comparison between actigraphy, polysomnography, and sleep diary. *Bipolar Disord*. 2012;14(8):870–879.
26. Matthews KA, Patel SR, Pantesco EJ, et al. Similarities and differences in estimates of sleep duration by polysomnography, actigraphy, diary, and self-reported habitual sleep in a community sample. *Sleep Health*. 2018;4(1):96–103.
27. McCall C, McCall WV. Comparison of actigraphy with polysomnography and sleep logs in depressed insomniacs. *J Sleep Res*. 2012;21(1):122–127.
28. Lichstein KL, Stone KC, Donaldson J, et al. Actigraphy validation with insomnia. *Sleep*. 2006;29(2):232–239.
29. Kay DB, Buysse DJ, Germain A, Hall M, Monk TH. Subjective-objective sleep discrepancy among older adults: associations with insomnia diagnosis and insomnia treatment. *J Sleep Res*. 2015;24(1):32–39.
30. Rezaie L, Fobian AD, McCall WV, Khazaie H. Paradoxical insomnia and subjective-objective sleep discrepancy: a review. *Sleep Med Rev*. 2018;40:196–202.
31. Crönlein T, Lehner A, Schüssler P, Geisler P, Rupprecht R, Wetter TC. Changes in subjective-objective sleep discrepancy following inpatient cognitive behavior therapy for insomnia. *Behav Ther*. 2019;50(5):994–1001.
32. Janků K, Šmotek M, Fárková E, Kopřivová J. Subjective-objective sleep discrepancy in patients with insomnia during and after cognitive behavioural therapy: an actigraphy study. *J Sleep Res*. 2020;29(4):e13064.
33. Kaplan KA, Hirshman J, Hernandez B, et al; Osteoporotic Fractures in Men (MrOS), Study of Osteoporotic Fractures SOF Research Groups. When a gold standard isn't so golden: lack of prediction of subjective sleep quality from sleep polysomnography. *Biol Psychol*. 2017;123:37–46.

ACKNOWLEDGMENTS

The authors thank and acknowledge the study participants and the research assistants who made this project possible: Kirti Veeramachaneni, Brett Messman, Aurora Brown, Ryan Moore, Jenny Quinn, Hanan Rafiuddin, Stormie Garza, Michelle Liu, Ian Dadeboe, Cynthia Yan, Cristabel Abi-Hanna, Sara Koh, and Emily Bready. Data for the current study were collected and presented as part of a dissertation project (Dietch, 2018).

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication November 16, 2020
Submitted in final revised form February 5, 2021
Accepted for publication February 9, 2021
Address correspondence to: Jessica R. Dietch, PhD, Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine, 401 Quarry Road, Palo Alto, CA 94304; Email: jessee.dietch@oregonstate.edu

DISCLOSURE STATEMENT

All authors have seen and approved the final manuscript. Work for this study was performed at the Department of Psychology, University of North Texas, Denton, Texas. Both authors were affiliated with this institution at the time of the study. This work was supported by the Foundation for Rehabilitation Psychology (dissertation grant) and the General Sleep Corporation (in-kind grant). Drs. Dietch and Taylor were loaned study devices and equipment from General Sleep Corporation. Receipt of devices did not influence the reporting of any study results and the authors have no ongoing financial relationship with General Sleep Corporation. The authors report no conflicts of interest.