HHS Public Access

Author manuscript

Behav Sleep Med. Author manuscript; available in PMC 2017 October 26.

Published in final edited form as:

Behav Sleep Med. 2013 ; 11(3): 173–188. doi:10.1080/15402002.2012.654549.

Pittsburgh and Epworth Sleep Scale Items: Accuracy of ratings across different reporting periods

Joan E. Broderick, Ph.D., **Doerte U. Junghaenel, Ph.D.**, **Stefan Schneider, Ph.D.**, **John J. Pilosi, B.A.**, and **Arthur A. Stone, Ph.D.**

Department of Psychiatry & Behavioral Science Stony Brook University

Abstract

This study examined the ecological validity of sleep experience reports across different lengths of reporting periods. The accuracy of item responses on the Pittsburgh Sleep Quality Index (PSQI) and Epworth Sleepiness Scale (ESS) across 3, 7, and 28-day reporting periods was examined in relation to electronic daily item ratings. Primary care clinic patients (*N*=119) were recruited and were not required to have sleep problems to participate. Analyses found few differences in item scores when electronic daily ratings were compared with recall ratings regardless of the length of the reporting period. However, within-subject analyses indicated low levels of accuracy in recall of sleep items for specific days in the last week. Thus, for the purpose of between-subject comparisons, patients generally can provide accurate recall of sleep experiences; studies requiring finer grained analysis across time within subjects will require daily diary methodology.

Keywords

sleep; measurement; validity; diary; patient reported outcomes

Introduction

In 2004 the National Institutes of Health awarded multi-site collaboration grants to improve the measurement of patient reported outcomes (PROs) for clinical trials (see http://www.nihpromis.org/default.aspx). One component of this initiative has been to examine the ecological validity of PROs, that is, the level of association of real-time PROs with recall (A. Stone & Shiffman, 1994). Our first studies focused on pain, fatigue, and interference with daily functioning in patients with chronic illness (Broderick, Schneider, Schwartz, & Stone, 2010; Broderick et al., 2008). These and other studies demonstrated some discrepancies between ratings based on real-time assessment and recall ratings of a week or more (Salovey, Smith, Turk, Jobe, & Willis, 1993). Here, we extend this work by looking at sleep PROs. Approximately 30-50% of the general population report insomnia, daytime sleepiness, or other sleep difficulties, yet the problems are often clinically overlooked (Buysse et al., 2008; Hossain & Shapiro, 2002). The current study was designed to examine

Corresponding author: Joan E. Broderick, Ph.D. Department of Psychiatry and Behavioral Science Putnam Hall, South Campus Stony Brook University Stony Brook, NY 11794-8790 Phone: 631-632-8083 Fax: 631-632-3165 Joan.Broderick@StonyBrook.edu.

the ecological validity of the Pittsburgh Sleep Quality Index (PSQI) and the Epworth Sleepiness Scale (ESS) items in a sample of patients attending a primary care clinic.

This study examined the ecological validity for each item on the PSQI and ESS across three different reporting periods: 3 days, 7 days, and 28 days. Ecological validity reflects the degree to which a measure is a true index of experience in the respondent's daily life (A. Stone & Shiffman, 1994). Daily reports on a hand-held computer provided ratings less subject to memory loss and recall bias (Stone, Shiffman, Schwartz, Broderick, & Hufford, 2003). These electronic daily reports were compared with recall ratings of items referencing the three different reporting periods during a month of daily ratings. Based on results from prior work, we hypothesized (1) that patients on average would over-report nighttime sleep problems and daytime sleepiness problems on the recall ratings relative to aggregated daily reports, (2) that the amount of over-reporting of problems would be greater for longer than for shorter recall-periods, and (3) that the correlation between aggregated daily reports and recall reports would differ across reporting periods, in that the correlation would be greater for shorter recall periods than for longer recall periods. We also examined the accuracy of 7 day-by-day recall ratings on 4 items made during the final study visit. These data will help to determine the optimal reporting period for accuracy in the reporting of these PRO domains.

Methods

Participants

Patients were recruited from the Primary Care Clinic at Stony Brook University Medical Center. The research staff approached patients in the waiting room, and those who were interested could provide contact information, and were screened on the phone for eligibility. Eligibility criteria were: greater than age 18, fluent in English, no visual or hearing impairment, no difficulty holding a pen or writing, typically awake before 10 AM and asleep after 7PM, no night shift job that leads to daytime sleep, no substance abuse or cognitive deficits, and able to travel to the research laboratory two times in a month. Patients were included in the study without regard to sleep disorders or other diseases. This approach was used to ensure a high degree of variability in responses to PSQI and ESS items needed for generalizable results.

Procedure

The Stony Brook University Institutional Review Board approved the study (Approval #6845). During their first laboratory visit, patients gave written informed consent and were trained in the use of the electronic diary (ED) to record daily completion of PSIQ and ESS items. The ED is a Palm Pilot computer that used the open-source Experience Sampling Program (ESP; http://www.experience-sampling.org) to capture symptom ratings. The software records the time and date of each entry. A 24-hour and four weekly follow-up phone calls were conducted to ensure that electronic recording was going well. During the next month, participants completed both morning (PSQI items) and evening (PSQI and ESS items) reports on the ED shortly after waking and shortly before going to sleep. The ED had an alarm feature that alerted participants to complete their daily ratings. After 30 to 36 days,

patients returned to the laboratory with their ED and completed additional questionnaires about their health and sleep quality. During this visit, as a further probe of recall accuracy, participants were presented with 7 recall cards with 4 items for each for the previous 7 days. This would enable assessment of actual memory accuracy across 7 previous days.

In addition to the daily ED ratings, recall ratings for the three different reporting periods were made with an Interactive Voice Recording (IVR) system. At the laboratory visit, participants were provided with five numerically labeled envelopes holding the recall questionnaires. Participants were informed that they would be telephoned by the IVR system on five randomly selected nights during the month of the study. On those nights, they would be informed which recall questionnaire they should complete (3-, 7-, or 28-day). In order to avoid anticipatory monitoring of sleep quality and daytime fatigue, participants were not informed of the dates they would be contacted to complete the recall questionnaires via IVR. When called, they were instructed to record their responses on the paper questionnaire, and then to call back and record their responses via IVR. This procedure allowed time and date stamping of the recall ratings. To enhance data quality, patients were instructed to place each completed paper questionnaire into the mail the next day to permit comparison with IVR entries in the event of missing entries or outliers. If a participant missed their scheduled call on a particular evening, a research assistant called the participant the next morning to inquire about reasons for and/or difficulties in completing the ratings. Another attempt to reach the participant was made by the IVR system the subsequent evening without prior knowledge of the participant.

Each participant was randomly assigned to 1 of 6 different recall schedules that specified when, within the 30-36 day study period, the two 3-day and the two 7-day recall assessments would take place. The schedules were designed with no overlap of days for the two 3-day and two 7-day recall periods, and one of each of the 3-day and 7-day recalls took place during the week and the others on the weekend. The one 28-day recall was fixed at the end of the study for all participants (and overlapped with the other recall periods). Participants' length of participation varied between 30 and 36 days based on the specific IVR calling schedule to which they were randomized. Participants were compensated $125 for full completion of this research study.

Measures

The PSQI and the ESS are two very widely used instruments for measurement of sleep dysfunction. The PSQI is a self-report global measure of sleep (Buysse, Reynolds, Monk, Berman, & Kupfer, 1989). The reporting period is one month, and scores >5 on the global scale indicate clinically meaningful sleep disturbance. Test-retest reliability (46 days) for the global score is good (r= .86); although for some of the subscales, it can be substantially less (e.g., r= .23 for sleep quality) (Backhaus, Junghanns, Broocks, Riemann, & Hohagen, 2002; Buysse, et al., 1989). The PSQI has demonstrated the ability to differentiate among a number of patient populations with varying sleep quality, convergent and discriminant construct validity (Carpenter & Andrykowski, 1998), and responsivity to treatment to improve sleep (Krakow et al., 2004). Correlations between paper sleep diaries and PSQI sleep duration reports (r=.81) and sleep onset latency reports (r=.71) are good, however, the

diaries indicate longer sleep duration than PSQI and shorter sleep onset latency (Backhaus, et al., 2002; Sharkey et al., 2010).

The ESS measures self-reported excessive daytime sleepiness (Johns, 1991). It has been used to probe for indicators of a variety of sleep disorders including obstructive sleep apnea, insomnia, and narcolepsy. Higher scores indicate greater levels of daytime sleepiness, and scores >10 indicate excessive daytime sleepiness. No reporting period is specified. Correspondence between ESS scores and polysomnography, respiratory disturbance index, and the apnea-hypopnea index are often low suggesting that the measure is not a sensitive measure of objectively measured sleep disturbance (Kaminska et al., 2010; Sauter et al., 2000) and correspondence with the Multiple Sleep Latency Index is also generally modest (Johns, 2000). Nevertheless, it has been found to be stable across repeated measures (7 months; ICC = .87) and sensitive to treatment for apnea (Kaminska, et al., 2010; Massie & Hart, 2003).

Patients completed the standard PSQI and ESS scales during their first laboratory visit. For daily ED ratings and IVR recall ratings, the PSQI items were slightly modified to accommodate the recall period of the ratings. For example, the PSQI item, “During the past month, how long (in minutes) has it usually taken you to fall asleep each night? # of minutes ___” was presented on the ED as “Last night, it took me about __ minutes to fall asleep.” Likewise, the IVR item was “During the past (3, 7, or 28 days), how long (in minutes) has it usually taken you to fall asleep each night?” The same strategy was applied to ESS items. The ESS instructs respondents to rate each activity on a 0-3 scale for the likelihood of dozing off while doing the activity. Thus, we wanted to look at how often the patient reported sleepiness when in the ESS activities. For our ED and IVR ratings, we asked two questions. For the ED we asked, “At any time today, were you sitting and reading?” (yes or no). For patients who responded “yes,” it was followed by “While you were sitting and reading, did you doze off or fall asleep?” (yes or no). Similarly, for the IVR ratings, we asked “During the past (3, 7, or 28) days, on how many days were you sitting and reading? __# of days.” This was followed by “On how many of those days did you doze off or fall asleep while sitting and reading? __ # days” (details available from authors).

At the last laboratory visit, patients made recall ratings for 4 items for each of the last 7 days: (1) what time the participant went to bed, (2) if the participant had trouble sleeping because he/she could not get to sleep within 30 minutes, (3) a sleep quality rating (1-4 scale). The fourth item, overall, how sleepy he/she was during the day (1-7 scale), was used to capture daytime sleepiness. Participants could write “CR” (can’t remember) next to any items that could not be recalled.

Analytic Strategy

Compliance criteria for ED reports and IVR recall reports—ED reports and IVR recall reports were examined for compliance with the protocol assessment schedule. An insufficient number of ED reports during a reporting period could result in inaccurate estimates of the reporting period, and comparisons with IVR recall reports could be biased if IVR reports were not completed on the last day of the reporting period.

An IVR recall rating was considered compliant if it was completed on the evening that the participant was contacted to make a rating. If an IVR report was missed, the participant was contacted the following evening; completing the report on that evening was considered compliant for the 28-day report, as well as for the 3-day and 7-day reports if it took place for the first time during the study protocol.

Compliance criteria for the ED ratings were: a morning-report had to be completed between 5 AM and 2 PM; an evening-report had to be completed between 6 PM and 3 AM for any given day. We required that a patient complete all 3 ED assessments for the 3-day period, at least 6 for the 7-day period, and at least 21 for the 28-day period. These criteria were assessed separately for morning- and evening-reports; that is, patients could meet the criteria for neither report, only either for the morning- or evening report, or for both reports.

To be included in the analyses, we required that a patient met the IVR and ED compliance criteria for at least one out of the two 3-day and 7-day reporting periods, as well as for the 28-day period. If a participant met the criteria for both of the 3-day or both of the 7-day periods, the data from the period that occurred first in the study protocol was analyzed.

Aggregation of ED data across days of a reporting period—For each participant, ED responses were averaged across the days of a reporting period. ED responses for items addressing the occurrence of discrete events (e.g., whether sleep medication was taken that day) were averaged across days and multiplied by 100, to represent the *percentage* of days an event was endorsed. IVR recall responses for these items (e.g., on how many days sleep medication was taken) were transformed correspondingly to represent the *percentage* of days a problem was indicated.

Analysis of level differences between ED and IVR recall—Repeated measures analysis of variance was used to examine level differences between ED and IVR recall reports. To address the hypothesis that symptom severity and frequency would be higher in IVR recall than in aggregated ED reports, we examined the main effect of reporting method (ED/IVR, within-subjects) across all three reporting periods (3-, 7-, 28-day, within subjects). To address the hypothesis that the difference between IVR recall and aggregated ED reports would be greater for longer than for shorter reporting periods, we examined the reporting method (ED/recall) by reporting period (3-, 7-, 28-day) interaction term in 2×3 repeated-measures analysis of variance.

Analysis of correspondence between ED and IVR recall—Correspondence between ED reports and IVR recall reports was examined with correlation analyses. To address the hypothesis that the correlation between IVR recall and aggregated ED reports would differ across the 3-, 7-, and 28-day reporting periods, we had to take into account that these correlations were non-independent (i.e., they came from the same sample). Thus, we estimated the 3 correlations simultaneously as part of the full correlation matrix of all reports (a 6×6 correlation matrix of 3-, 7-, 28-day ED, and 3-, 7-, 28-day IVR). A Wald Chi-square test was used to test the null-hypothesis of no differences across reporting periods. Analyses were conducted separately for each item.

Missing data—The compliance criteria minimized missing data, but allowed for missing ED reports on some days (see compliance criteria); multiple imputation was used to account for the missing data. For each missing ED data point, we randomly selected a value from all available ED ratings for that person and item, generating a set of five imputed databases. Multiple imputations allow for variation across the five data sets to reflect the uncertainty of imputed data, and research indicates that five is a sufficient number of datasets when the rate of missing responses is low (Schafer, 1997). Analyses were performed using M*plus* (Version 4).

RESULTS

Compliance and analysis sample

Figure 1 shows the flow of patients into the study. Telephone screening of 195 patients found 22 (11%) were ineligible. Of the 173 eligible patients, 48 (28%) declined participation. Six patients dropped out, and 119 completed the study.

Overall, compliance criteria were met for 89% of IVR recall reports (87% for 3-day, 87% for 7-day, and 95% for 28 day recall, respectively), for 86% of ED morning reports, and for 88% of ED evening reports. Data were excluded due to IVR or ED noncompliance or ED malfunction. The analysis samples were $n = 83$ for morning report items, and $n = 87$ for evening report items (with a combined sample size of $n = 94$). In the analysis samples, there were no missing IVR ratings. There were no missing days for the 3-day reporting period. For the 7-day reporting period, 15 out of 581 (2.6%) days were missing for morning reports, and 20 out of 609 (3.3%) for evening reports. For the 28-day reporting period, 128 out of 2,324 (5.5%) days were missing for morning reports, and 121 out of 2,436 (5.0%) days were missing for evening reports.

Participants had a mean age of 57 years (range 21 to 83, $SD = 13.9$) and tended to be female (68%), married (61%), and White (87%) (see Table 1). As would be expected in a Primary Care Clinic, the age distribution was skewed with the majority (68%) of the participants > 50 years of age. Self-reported general health was described as “fair” or “poor” by 8% of the sample. Half (51%) of the patients met the cutoff for “poor sleepers” (score >5) on the PSQI (Buysse, et al., 1989). Twenty-five percent of the patients had ESS scores >10, the most common cut-off for excessive daytime sleepiness (Sanford et al., 2006).

There were no statistically significant differences between the analyzed and excluded participants on the demographic variables. Analyzed and excluded participants also did not differ on baseline ESS scores ($p = .60$); however, baseline PSQI indicated greater sleep problems in excluded than in analyzed participants (mean PSQI scores of 9.8 versus 6.8, respectively; $p = .006$).

Comparison of rating level differences between aggregated daily and recall ratings across reporting periods

Our first hypothesis was that on average patients would over-report nighttime sleep problems and daytime sleepiness in recall reports relative to aggregated ED reports. This hypothesis was not supported (see Tables 2 and 3). For only 1 (4%) out of 25 items was the mean IVR

recall response significantly ($p < .05$) higher than the mean ED response: patients reported "dozing off while watching TV" on 23.0% (IVR) versus 19.2% (ED) of the days. In contrast to the hypothesis, patients reported fewer sleep problems on recall on several items. Specifically, "problems with keeping up enthusiasm to get things done" was noted significantly less often on recall, as were sleeping difficulties due to breathing problems and due to coughing/snoring. However, these differences were small in magnitude, in that on all items the self-reported occurrence differed by less than 4% of the days between IVR recall and ED methods. In addition, patients reported going to bed 24 minutes earlier ($p < .001$) and getting up 11 minutes earlier ($p < .01$) on the recall ratings compared with the daily ratings.

The hypothesis that the degree of over-reporting of symptoms would be greater for longer than for shorter reporting periods was also not supported. Statistically significant ($p < .05$) interactions between reporting method (ED / IVR) and reporting period (3-, 7-, 28 day) were found for 8 (32%) of the items (see Tables 2 and 3). The direction of these interactions suggested that the IVR recall ratings indicated more sleep problems over the shorter (3 day) reporting period than ED reports. Whereas, recall ratings for the longer reporting period (28-day) indicated fewer sleep problems than ED reports. However, on average across all items, the percentage of days for which problems were reported was 1.3% greater, 0.2% greater, and 1.1% smaller in IVR recall than ED reports for the 3-, 7- and 28-day reporting periods, respectively. This suggests that the overall evidence for an effect of the length of recall on the degree of under- or over-reporting of symptoms is very weak.

Correspondence between ED and recall ratings across reporting periods

Correlations between aggregated ED reports and IVR recall reports are shown in Tables 4 (for morning reports) and 5 (for evening reports). For time ratings (time gone to bed and time woke up) and durations (sleep latency and duration of sleep), the correlations between recall and aggregated ED ratings were high for all reporting periods, with an average correlation of $r = .85$ (range .69 to .95). Similarly, moderate to high correspondence between recall and aggregated ED reports was evident for items reporting frequencies of sleeping problems with an average correlation of $r = .84$ (range .49 to .97). However, correlations were generally lower for ESS items pertaining to ratings of the number of days of daytime sleepiness during various activities with an average correlation of $r = .62$ (range .06 to .95). This latter finding may in part be explained by the low prevalence of daytime sleepiness problems reported for several items (below 5% of the days for some items, see Table 3).

The hypothesis that longer reporting periods would result in lower correspondence was not supported. Even though the magnitude of the correlations between ED and IVR recall reports differed significantly ($p < .05$) between reporting periods for 11 (41%) out of the 25 items, the highest correlations were evident inconsistently for the 3-, 7-, and 28-day period (see Tables 4 and 5). The average correlation between ED and IVR recall reports was $r = .74$ for the 3-day, $r = .75$ for the 7-day, and $r = .77$ for the 28-day reporting period, respectively.

Recall ratings of each of the last 7 days

At the last laboratory visit, patients made recall ratings on 4 items for each of the last 7 days. They were given the option of responding that they could not remember. Ten percent of patients reported inability to remember 3 days previously for the 3 sleep quality items, and 10-20% could not remember these experiences by the 5th through 7th previous days. Patients found it even more difficult to remember bedtime: 24% could not remember 3 days ago, and 44% could not remember 7 days ago.

We examined level differences and correlations between recall ratings for *each* of the 7 days and the ED ratings obtained on corresponding days (during the patients' last week in the study). The analyses were conducted on the 94 patients that were included in the analysis sample for the main hypotheses. Across 7 days and 4 items, only 2 of 28 level differences between recall ratings and ED ratings were significant ($p \leq .05$). Across the 4 items, the average correlation for "yesterday" was .80. Starting with two days ago through 7 days ago, the correlations range from .51 to .66. Thus, the between-subject correlations between recall and ED ratings decline quickly, starting with two days ago. Indeed, 71% of the correlations were less than .70, and 25% were ≤ .50 indicating sub-optimal correspondence.

An even more relevant analysis to determine recall-rating accuracy is to examine the *within-subject* correlations between recall and ED ratings across the 7 days for each item. Whereas the between-subject analyses examined differences between people for a given day, within-subject analyses examine the variation of responses from day to day within a person. Thus, the within-subject correlations more directly address the question of whether patients can recall and differentiate a given day from other days in the past week. The average within-subject correlation pooled across patients for the bedtime item was .58, while the correlations for the 3 qualitative items relating to sleep ranged from .20 to .26, indicating poor correspondence.

Discussion

This study was designed to examine the ecological validity of Pittsburgh Sleep Quality Index and Epworth Sleepiness Scale item responses across 3 reporting periods: 3-day, 7-day, and 28-day. Specifically, the data were analyzed for comparability of reports of items such as sleep quality, latency, duration, and instances of daytime sleepiness when aggregated daily reports were compared with recall reports. Good comparability would indicate acceptable ecological validity. Research on other patient reported outcomes, such as pain, fatigue and physical functioning, suggested that, as the reporting period gets longer, patients report higher levels of symptoms, and the between-subject correlations of daily and recall ratings can vary by length of reporting period (Broderick, et al., 2010; Broderick, et al., 2008). We hypothesized that the ecological validity of sleep-related items might show some of the problems observed in other populations and item domains, given that the specific nature of problems assessed in many of the PSQI and ESS items might be difficult to remember across a month reporting period.

Our results did not support the first and second hypotheses that patients would over report sleep difficulties on recall and that the length of the recall period would systematically

impact patients' accuracy of PSQI and ESS item responses. Patients' accuracy on recall was comparable, regardless of the length of the reporting period from 3 days through a month. Some rating differences were observed when specific items were compared on recall versus daily reports. There were statistically significant differences, but they are probably not meaningful. Generally, across all of the items, the levels on the recall ratings were very close to those generated by the aggregated daily ratings. This suggests that patients are able to accurately report both sleep and daytime sleepiness experiences for recall periods of up to a month. Likewise, between-subject correlations of aggregated daily and recall reports demonstrated good correspondence for all items except those with very low frequencies observed, particularly the daytime sleepiness items. Thus, the third hypothesis was generally not supported.

A final set of analyses testing the accuracy of day-by-day recall for the last 7 days of the protocol (data collected on the last laboratory visit) provides some insight into the nature of recall of PROs. As we observed in a previous study of pain ratings where 19-34% of patients could not remember their pain 5 to 7 days ago (Broderick, et al., 2008), 10-40% of this study's patients acknowledged that they had difficulty remembering their sleep experiences beyond more than several days. When we examined the within-subject correlations of the 7 day-by-day ratings made at the end of the study with the corresponding daily ratings, we found exceedingly low correlations (.20's). This means that patients cannot accurately identify days from the past week that had relatively high and low levels of symptoms. Nevertheless, the between-subject level and correlation analyses on the 7 day-by-day ratings show good correspondence, suggesting that patients are able to generate ratings that accurately reflect their level relative to other patients. We interpret these results as evidence that patients "know" their typical/average sleep and daytime sleepiness experiences, and they can report them accurately for between-subject comparisons for short and long reporting periods. However, our data suggest that this accuracy is not due to patients being able to always remember their sleep-related experiences on the specific days included in the reporting period. For example, if patients "know" when they generally go to bed, when they wake up, and how often they fall asleep watching television – this would be enough to generate high between-subject correlations. However, they cannot remember whether last Monday was better or worse than last Tuesday. For these reasons, we conclude that sleep-related PROs, with items like the PSQI and ESS, can be administered with confidence for week or month-long reporting periods for between-subject analyses. In the case of studies requiring a finer grain within-subject analysis across days, daily diary measurement is recommended.

There are strengths and limitations in this study. One study limitation is that the response options that patients were given on the ED and IVR recall questions in this study were modified from the original PSQI and ESS scales to allow for direct comparison between the daily and recall assessment methods. The response options on the PSQI and ESS standard instruments are often ranges, whereas the options for the IVR recall items required that the patient report a specific number of days for a sleep problem. Thus, the recall task in this study could be more challenging than completion of the standard instruments. Therefore, this study does not directly address the accuracy of patient reports using those response options. However, the fact that more specific responses were required in this study lends

strength to the conclusions. Second, this study was observational and did not collect data during an intervention. The accuracy data reported in this study may not generalize to recall ratings during a reporting period with change, as in a clinical trial. Third, it is possible that because patients were focusing on their sleep experiences each day, their ability to recall them might have been improved relative to recall ratings made in the absence of daily reports. However, we tested this possibility in a previous study of pain and found no evidence (Stone et al., 2003). Fourth, we noted that patients who were removed from the analysis sample due to non-compliance with assessments had scores on the PSQI that reflected greater sleep disturbance than the analyzed sample. This could be due to many factors including greater sleep disturbance interfering with concentrating or completing tasks. It is possible that their recall data would have revealed more recall bias. In fact, as suggested by reviewers, we analyzed subgroup differences (n=48 poor sleepers vs. n=46 good sleepers) and found that --relative to good sleepers-- poor sleepers reported significantly more sleep problems on recall than reflected in their daily ratings on 4 of the 24 PSQI and ESS items (uncorrected for multiple comparisons, $p<.05$). However, the magnitude of these differences was small.

The strengths include recruitment of a Primary Care sample of adult men and women with a range of sleep difficulties. At baseline, half of the sample was above the threshold (score >5) on the PSQI for clinically relevant sleep difficulties, and 25% of the sample met the threshold (score >10) for daytime sleepiness on the ESS. These rates are in line with other studies of the US general population, thus suggesting a sample that is broadly speaking relevant to both general and clinical populations (Buysse, et al., 2008; Hossain & Shapiro, 2002). Second, the size of the sample and overall protocol compliance yielded an analysis sample with very little missing data. Thus, the comparisons between recall and daily ratings have a solid basis. Third, the study examined three different reporting period lengths including a month used by the PSQI. Fourth, the design of the study (patient did not know when the recall ratings would be requested) allows generalization to survey and clinical situations where the patient may not anticipate being asked to provide recall ratings of sleep-related experiences.

Acknowledgements

This research was supported by grants from the National Institutes of Health (1 U01-AR052170-01, 1 U01-AR05794801) and from the National Center for Research Resources (M01-RR10710). We would like to thank David Goodrich, MD, Director of Stony Brook Primary Care, for facilitating recruitment of patients; Leighann Litcher-Kelly for her assistance with the electronic diaries; and Justin Teague and Kristen Derkevics for their assistance in collecting data. Prosodie Interactive, Inc. provided Programming and data management services for the IVR assessments.

Abbreviations

ED	electronic diary
ESS	Epworth Sleepiness Scale
IVR	Interactive Voice Recording
PRO	Patient reported outcome

PSQI Pittsburgh Sleep Quality Index

References

Backhaus J, Junghanns K, Broocks A, Riemann D, Hohagen F. Test-retest reliability and validity of the Pittsburgh Sleep Quality Index in primary insomnia. Journal of Psychosomatic Research. 2002; 53(3):737–740. [PubMed: 12217446]

Broderick JE, Schneider S, Schwartz JE, Stone AA. Interference with activities due to pain and fatigue: accuracy of ratings across different reporting periods. Quality of Life Research. 2010; 19(8):1163–1170. doi: 10.1007/s11136-010-9681-x. [PubMed: 20535565]

Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. Pain. 2008; 139:146–157. [PubMed: 18455312]

Buysse DJ, Hall ML, Strollo PJ, Kamarck TW, Owens J, Lee L, Matthews KA. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and clinical/polysomnographic measures in a community sample. Journal of Clinical Sleep Medicine. 2008; 4(6):563–571. [PubMed: 19110886]

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 Research. 1989; 28(2):193–213. [PubMed: 2748771]

Carpenter JS, Andrykowski MA. Psychometric evaluation of the Pittsburgh Sleep Quality Index. Journal of Psychosomatic Research. 1998; 45(1 Spec No):5–13. [PubMed: 9720850]

Hossain JL, Shapiro CM. The prevalence, cost implications, and management of sleep disorders: an overview. Sleep and Breathing. 2002; 6(2):85–102. [PubMed: 12075483]

Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep. 1991; 14(6):540–545. [PubMed: 1798888]

Johns MW. Sensitivity and specificity of the multiple sleep latency test (MSLT), the maintenance of wakefulness test and the epworth sleepiness scale: failure of the MSLT as a gold standard. Journal of Sleep Research. 2000; 9(1):5–11. [PubMed: 10733683]

Kaminska M, Jobin V, Mayer P, Amyot R, Perraton-Brillon M, Bellemare F. The Epworth Sleepiness Scale: Self-administration versus administration by the physician, and validation of a French version. Canadian Respiratory Journal. 2010; 17(2):27–34.

Krakow B, Melendrez D, Lee SA, Warner TD, Clark JO, Sklar D. Refractory insomnia and sleep-disordered breathing: a pilot study. Sleep and Breathing. 2004; 8(1):15–29. doi: 10.1007/s11325-004-0015-5. [PubMed: 15026935]

Massie CA, Hart RW. Clinical outcomes related to interface type in patients with obstructive sleep apnea/hypopnea syndrome who are using continuous positive airway pressure. Chest. 2003; 123(4):1112–1128. [PubMed: 12684301]

Salovey P, Smith A, Turk DC, Jobe JB, Willis GB. The accuracy of memories for pain: not so bad most of the time. American Pain Society Journal. 1993; 2:184–191.

Sanford S, Lichstein K, Durrence H, Riedel B, DJ T, Bush A. The influence of age, gender, ethnicity, and insomnia on Epworth sleepiness scores: A normative US population. Sleep Medicine. 2006; 7:319–326. [PubMed: 16713340]

Sauter C, Asenbaum S, Popovic R, Bauer H, Lamm C, Klosch G, Zeitlhofer J. Excessive daytime sleepiness in patients suffering from different levels of obstructive sleep apnoea syndrome. Journal of Sleep Research. 2000; 9(3):293–301. [PubMed: 11012870]

Schafer, JL. Analysis of Incomplete Multivariate Data. Chapman & Hall; London: 1997.

Sharkey KM, Kurth ME, Anderson BJ, Corso RP, Millman RP, Stein MD. Assessing sleep in opioid dependence: A comparison of subjective ratings, sleep diaries, and home polysomnography in methadone maintenance patients. Drug and Alcohol Dependence. 2010; 113(2-3):245–248. doi: 10.1016/j.drugalcdep.2010.08.007. [PubMed: 20850231]

Stone A, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Annals of Behavioral Medicine. 1994; 16:199–202.

Stone A, Shiffman S, Schwartz J, Broderick J, Hufford M. Patient compliance with paper and electronic diaries. Controlled Clinical Trials. 2003; 24(2):182–199. [PubMed: 12689739]

Stone AA, Broderick JE, Schwartz JE, Shiffman S, Litcher-Kelly L, Calvanese P. Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. Pain. 2003; 104(1-2):343–351. [PubMed: 12855344]

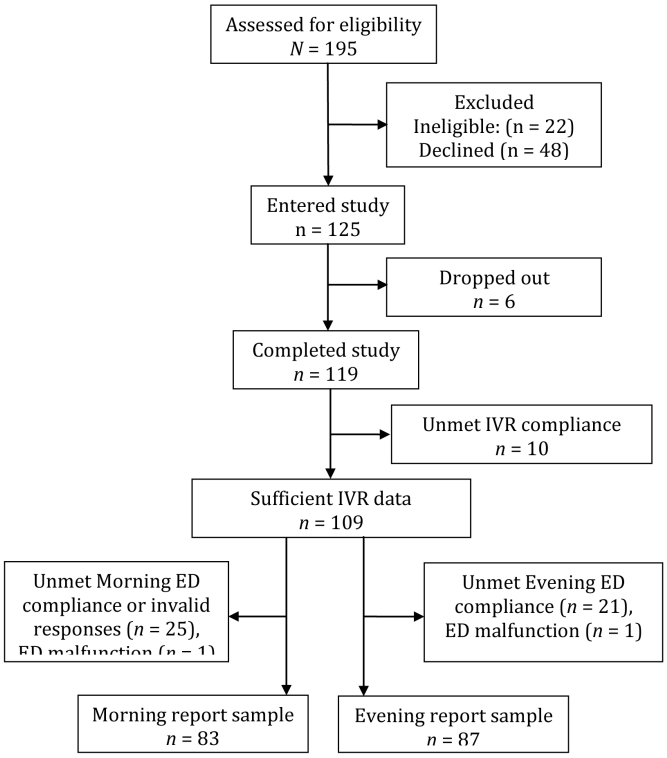

FIGURE 1.

The flow of patients into the study. Note. IVR D Interactive Voice Recording; ED D electronic diary.

Table 1

Demographic and health characteristics of the analysis sample (N = 94)

Age (Mean, *SD*)	57.3 (13.9) (range= 21-83)
Female	68%
Race (White)	87%
Married	61%
Education[a]	
Did not finish high school	2%
High school graduate	18%
Some college	29%
College graduate or more	49%
Employed	47%
Disability benefits	19%
General health perception	
Excellent	10%
Very good	50%
Good	32%
Fair	7%
Poor	1%
PSQI global score (Mean, *SD*)	6.8 (4.6)
Poor sleepers (Score >5)	51%
ESS score (Mean, *SD*)	7.0 (4.7)
Excessive sleepiness (Score >10)	25%

[a] *Note*: not indicated by two participants. PSQI = Pittsburgh Sleep Quality Index. ESS = Epworth Sleepiness Scale

Table 2

Means of IVR recall ratings and aggregated ED ratings for morning reports (n = 83)

Item [metric of scores]	ED mean (*SD*)			IVR recall mean (*SD*)			IVR - ED mean difference	IVR-ED by period interaction
	3-day	7-day	28-day	3-day	7-day	28-day	χ^2 (*df* = 1)	χ^2 (*df* = 2)
Time gone to bed at night [hours and decimal hrs]	23.45 (1.3)	23.41 (1.2)	23.40 (1.1)	23.16 (1.3)	23.08 (1.3)	23.04 (1.3)	28.34*** ED > IVR	1.65
Time got up in the morning [hours and decimal hrs]	7.34 (1.3)	7.39 (1.3)	7.37 (1.1)	7.21 (1.4)	7.17 (1.5)	7.15 (1.3)	7.16** ED > IVR	1.81
Sleep latency [minutes and decimal mins]	19.98 (17.0)	22.89 (21.3)	21.50 (19.1)	20.74 (27.3)	19.84 (20.5)	17.49 (18.7)	3.57	6.55*
Duration of sleep [hours and decimal hrs]	6.91 (1.1)	6.87 (0.9)	6.92 (0.9)	6.97 (1.1)	6.88 (1.0)	6.92 (1.1)	0.18	0.32
Trouble sleeping because …								
… couldn't fall asleep [percent of days]	21.3 (30.4)	21.3 (25.5)	19.6 (24.4)	23.7 (34.1)	25.1 (32.6)	18.3 (29.1)	0.73	6.22*
… woke up at night [percent of days]	45.4 (37.9)	51.1 (32.9)	50.1 (30.4)	52.2 (40.8)	54.2 (39.9)	46.1 (38.7)	0.72	12.30***
… had to use bathroom [percent of days]	39.8 (40.9)	40.5 (37.9)	39.6 (35.4)	43.4 (42.9)	44.8 (41.6)	39.5 (40.3)	2.25	6.31*
… trouble breathing [percent of days]	6.0 (19.5)	10.3 (24.4)	8.8 (20.4)	6.0 (20.8)	6.7 (19.8)	6.3 (17.6)	4.45* ED > IVR	4.89
… coughed/snored [percent of days]	5.6 (19.9)	7.7 (20.8)	6.1 (17.6)	4.8 (18.0)	6.0 (18.4)	3.7 (13.3)	4.04* ED > IVR	1.55
… felt too hot [percent of days]	11.2 (22.1)	10.3 (19.0)	10.7 (16.2)	11.6 (22.2)	10.5 (20.6)	8.5 (15.9)	0.34	4.16
… felt too cold [percent of days]	5.2 (17.5)	5.2 (12.1)	4.8 (10.8)	5.2 (17.5)	6.0 (14.2)	4.3 (9.8)	0.02	1.80
… of pain [percent of days]	17.3 (31.2)	18.4 (27.3)	17.1 (23.6)	17.7 (31.2)	17.9 (27.6)	14.8 (25.6)	0.38	2.06
Sleep medication taken [percent of days]	18.1 (35.6)	20.5 (35.2)	20.7 (33.9)	19.3 (37.0)	20.3 (36.6)	19.2 (34.2)	0.02	1.68
Sleep quality rating [mean rating, scale 1-4]	2.0 (0.6)	2.1 (0.5)	2.0 (0.5)	2.1 (0.6)	2.1 (0.5)	2.1 (0.6)	3.85	0.89

Note: $p < .05$,

** $p < .01$,

*** $p < .001$, ED = electronic diary. All items from PSQI. χ^2 values are from Wald tests.

Table 3

Means of IVR recall ratings and aggregated ED ratings for evening reports (n = 87)

Item [metric of scores]	Daily rating (*SD*)			Recall rating (*SD*)			IVR - ED mean difference	IVR-ED by period interaction
	3-day	7-day	28-day	3-day	7-day	28-day	χ^2 (*df* = 1)	χ^2 (*df* = 2)
Trouble staying awake[#] [percent of days]	5.0 (12.9)	6.6 (11.5)	5.9 (9.8)	3.4 (12.4)	4.8 (14.5)	3.7 (9.8)	2.69	0.26
Problem to keep enthusiasm[#] [percent of days]	20.7 (30.4)	23.3 (25.0)	21.9 (22.7)	19.5 (31.8)	20.0 (27.0)	16.2 (26.3)	4.45* ED > IVR	5.87
Doze off while …[§]								
… sitting & reading [percent of days]	8.4 (20.4)	11.0 (16.2)	10.2 (14.4)	11.1 (25.1)	12.3 (19.4)	11.0 (21.3)	1.30	0.62
… watching TV [percent of days]	17.2 (27.2)	20.3 (24.2)	20.1 (22.2)	26.1 (33.3)	21.0 (28.4)	22.0 (29.0)	4.40* IVR > ED	6.17*
… sitting in public [percent of days]	1.1 (7.9)	2.5 (7.3)	1.9 (4.1)	1.5 (7.0)	2.6 (9.3)	2.0 (8.0)	0.25	0.08
… passenger in car [percent of days]	3.4 (12.4)	3.7 (10.8)	2.5 (6.0)	3.1 (13.0)	2.6 (9.6)	2.8 (7.6)	0.32	7.36*
… lying down to rest [percent of days]	11.9 (24.2)	19.1 (21.8)	15.7 (18.2)	16.5 (28.5)	17.6 (25.4)	13.2 (22.6)	0.02	8.00*
… sitting and talking [percent of days]	0.4 (3.6)	1.2 (4.4)	1.0 (2.9)	0.0 (0.0)	1.8 (10.3)	2.2 (11.8)	0.70	0.10
… sitting after lunch [percent of days]	3.4 (11.3)	4.1 (9.2)	2.8 (5.3)	3.1 (14.0)	4.8 (15.3)	4.4 (14.1)	0.45	1.02
… in car and stopped in traffic [percent of days]	1.1 (6.1)	1.2 (5.9)	1.0 (3.6)	0.4 (3.6)	3.0 (13.4)	2.1 (11.7)	1.43	3.78

*Note: $p < .05$,

[#]Items from PSQI,

[§]Items from ESS. χ^2 values are from Wald tests

Table 4

Correlations between IVR recall ratings and aggregated ED ratings for morning reports (n = 83)

	Correlation			Wald test for difference between correlations
PSQI items	**3-day**	**7-day**	**28-day**	**χ^2 (df = 2)**
Time gone to bed at night	.83	.91	.95	20.02***
Time gotten up in the morning	.88	.88	.89	0.19
Sleep latency	.81	.88	.89	5.75
Duration of sleep	.74	.69	.82	8.02*
Trouble sleeping because …	.79	.80	.79	0.13
… couldn't fall asleep	.68	.81	.84	7.32*
… woke up at night	.85	.90	.91	5.23
… had to use bathroom	.90	.86	.84	4.20
… trouble breathing	.89	.94	.86	9.54**
… coughed/snored	.82	.82	.83	0.08
… felt too hot	.91	.49	.68	29.69***
… felt too cold	.88	.78	.94	20.95***
… of bad dreams	.83	.87	.90	5.35
Sleep medication taken	.90	.97	.95	11.98**
Sleep quality rating	.75	.69	.79	3.77

Note: $p < .05$,

** $p < .01$,

*** $p < .001$.

Table 5

Correlations between IVR recall ratings and aggregated ED ratings for evening reports (n = 87)

	Correlation			Wald test for difference between correlations
ESS and PSQI items	**3-day**	**7-day**	**28-day**	**χ^2 (df = 2)**
Trouble staying awake, percent days ‡	.37	.35	.39	0.21
Problem keeping up enthusiasm ‡	.76	.73	.85	11.30**
Doze off while sitting & reading	.62	.71	.74	2.72
Doze off while watching TV	.63	.80	.77	5.80
Doze off while sitting in public	.66	.70	.74	2.14
Doze off while passenger in car	.73	.80	.44	28.07***
Doze off while lying down to rest	.64	.69	.77	4.26
Doze off while sitting and talking	-- [a]	.31	.06	2.85 [b]
Doze off while sitting after lunch	.26	.58	.56	8.16*
Doze off while in car and stopped in traffic	.57	.67	.95	52.78***

Note: $p < .05$,

** $p < .01$,

*** $p < .001$.

‡ PSQI items.

[a] Coefficient could not be estimated due to zero variance in 3-day recall report.

[b] The test for difference in correlations excluded the 3-day reporting period ($df = 1$).