Published in final edited form as:
*Behav Sleep Med*. 2011 December 28; 10(1): 6–24. doi:10.1080/15402002.2012.636266.

# Development of Short Forms from the PROMIS Sleep Disturbance and Sleep-Related Impairment Item Banks

**Lan Yu**,
University of Pittsburgh

**Daniel J. Buysse**,
University of Pittsburgh

**Anne Germain**,
University of Pittsburgh

**Douglas E. Moul**,
Cleveland Clinic Foundation Sleep Disorders Center

**Angela Stover**,
University of Pittsburgh

**Nathan E. Dodds**,
University of Pittsburgh

**Kelly L. Johnston**, and
University of Pittsburgh

**Paul A. Pilkonis**
University of Pittsburgh

## Abstract

We report on the development of short forms from the Patient-Reported Outcomes Measurement Information System (PROMIS™) Sleep Disturbance (SD) and Sleep-Related Impairment (SRI) item banks. Results from post-hoc computerized adaptive testing (CAT) simulations, item discrimination parameters, item means, and clinical judgment were used to select the best-performing 8 items for SD and SRI. The final 8-item short forms provided less test information than the corresponding full banks, but correlated strongly with the longer forms. The short forms had greater measurement precision than the Pittsburgh Sleep Quality Index (PSQI) and the Epworth Sleepiness Scale (ESS) as indicated by larger test information values across the continuum of severity, despite having fewer total items, a major advantage for both research and clinical settings.

Sleep and wakefulness are fundamental neurobiological states regulated by homeostatic and circadian processes. Sleep and wake function in humans can be measured along many dimensions, including qualitative and quantitative aspects, as well as signs and symptoms of specific sleep disorders. Likewise, many different measurement tools are available: retrospective self-reports, prospective self-reports (sleep diaries), longitudinal measures of rest-activity patterns using wrist actigraphy, physiological recordings (polysomnography), and even functional imaging measures.

Address correspondence to Lan Yu at 3811 O'Hara Street, Pittsburgh, PA 15213, Telephone (412) 246-5831, Fax (412) 246-5840, yul2@upmc.edu.

Among self-report measures, the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989) is the most widely used scale for sleep disturbance, and the Epworth Sleepiness Scale (ESS; Johns, 1991; Johns, 1992) is the most widely used measure of daytime sleepiness, each with over 500 literature citations. Because of the nature of its items and its component structure, however, individual items in the PSQI are not conducive to validation with modern psychometric techniques such as item response theory (IRT) models. As discussed below, IRT models are useful for selecting items with the greatest information for describing a trait such as sleep disturbance. One consequence in the case of the PSQI is that the instrument has relatively poor ability to discriminate lower levels of severity because of the positive skewness of its distribution of scores. The ESS also has limitations, the major one being that it assesses behaviors (e.g., falling asleep in daily situations) that may not apply to all respondents. Thus, given content and psychometric concerns with the PSQI and ESS, there is a critical need for improved patient-reported outcomes (PROs) of sleep and sleep-related impairment during wakefulness. Measures of this sort can be thought of as general "thermometers" that provide continuous, relative values for every individual in the population, rather than as condition-specific measures that categorize individuals based on a cut-off score.

The Patient-Reported Outcomes Measurement Information System (PROMIS™) is an NIH Roadmap initiative designed to improve PROs using state-of-the-art psychometric methods (e.g., models from IRT; for detailed information, see www.nihpromis.org). The PROMIS Sleep Disturbance (SD) and Sleep-Related Impairment (SRI) item banks were developed using a rigorous and systematic methodology, including literature reviews, qualitative item review, focus groups, cognitive interviewing, and psychometric testing using methods from both classical test theory (CTT) and IRT (Buysse et al., 2010). This work is the most ambitious attempt to date to apply IRT methods to self-report measures of sleep and sleep-related waking impairments. The final SD and SRI item banks have 27 and 16 items each. The SD and SRI item banks assess qualitative aspects of sleep and wake function. They do not include quantitative or time-based items and do not assess symptoms of specific sleep disorders. Thus, they function as generic measures appropriate for gauging the severity of sleep-wake problems on a continuum, applicable across a range of conditions.

As noted, the use of IRT models was critical to the development of all PROMIS scales, but the distinction between CTT and IRT methods deserves emphasis. CTT, also called *true score theory*, assumes that each *observed* score equals the individual's *true* score plus some *error* (Gulliksen, 1950; Lord & Novick, 1968). The relationship among *observed* score, *true* score, and *error* yields *reliability*, which is defined as the ratio of *true* score variance to the *observed* score variance. Different approaches have been developed to estimate reliability, such as alternate-form reliability, examining the particular *form* of a test; test-retest reliability, examining the *occasion* of test administration; and internal consistency, examining the individual items of a test. Under the CTT framework, the standard error of measurement, describing the expected score fluctuations due to error, is constant among scores in the same population.

Unlike CTT, IRT refers to a class of psychometric techniques in which the probability of choosing each item response category is modeled as a function of a latent trait of interest. By convention, the latent trait is scaled along a dimension called theta ($\theta$), which has a mean of 0 and a standard deviation of 1. Item *discrimination* and item *difficulty* are the two major parameters used to define IRT models and describe individual items. The item discrimination parameter ($a$), also called slope parameter, indicates the shape of the category response curves, with higher slope parameters yielding steeper curves. Curves that are narrow and peaked indicate that the response categories differentiate well across $\theta$ values. The item difficulty parameter ($b$), also called threshold parameter, indicates the item's

location on the $\theta$ scale, and represents the $\theta$ level necessary to respond above the corresponding threshold with .50 probability.

The relationship between the probability of choosing a certain response category (e.g., *never, rarely, sometimes, often, always*) for a specific item and the underlying severity level can be described by a monotonically increasing function (i.e., an *S*-shaped function) called the *item characteristic function* (ICF). An ICF can be transformed into an *item information curve*, indicating the amount of information a single item contains at all points along the severity ($\theta$) scale. All of the individual item information curves can be combined to form a *test information curve*, which indicates the amount and accuracy of information the entire test contains at every point of $\theta$ (see Figures 3 and 4 for examples). Thus, the amount of information provided by a test may vary depending on the level of a respondent's severity of sleep disturbance or sleep-related impairment ($\theta$). These standardized curves can be used to compare the measurement precision of two or more scales. In this paper, we compare the test information curves for the PROMIS SD and SRI item banks, the SD and SRI short forms, the PSQI, and the ESS.

IRT models permit investigators to evaluate the performance of a single item or subsets of items as well as the entire test. Different items are better at discriminating people having different levels on the continuum of severity. For instance, the question, “Do you fall asleep while watching TV in a dark room late at night” would identify a milder degree of sleepiness than the question, “Do you fall asleep while talking to other people during the daytime?” One practical application of this feature, which CTT cannot provide, is the ability to construct short forms or tailored assessments, using a subset of items selected to maximize precision along clinically important ranges of severity.

Another advantage of IRT is that individuals' $\theta$ estimates are independent of the specific items administered from a larger calibrated item bank. With this feature, IRT serves as the basis for computerized adaptive testing (CAT), a method that provides a unique sequence of items tailored to the individual's personal severity ($\theta$). CAT avoids administering test items that add little information to an individual's assessment. For instance, during CAT administration of the SD item bank, item *S90* (*I had trouble sleeping*) might be administered first. *S90* is a useful initial item because it has a high slope parameter (‘*a*’ in Table 1), indicating high information content. If the individual endorses the most severe category (*always*), the CAT would be unlikely to choose item *S116* (*My sleep was refreshing*) as the next item, because *S116* mainly addresses a lower range of severity (indicated by small values for threshold values *b*1-*b*4). The net result is that CAT can provide an extremely efficient method of PRO administration. For more information regarding technical issues in IRT methodology, see Embretson & Reise (2000). A more detailed description of the specific PROMIS analytic framework is available elsewhere (see Reeve et al., 2007).

Individual items from the IRT-calibrated SD and SRI item banks can be selected to create short forms for assessing SD and SRI. The short forms can be constructed adaptively in real time based on each respondent's answers to previous items, as in computerized adaptive testing (CAT). Alternatively, static short forms (i.e., containing a fixed set of items) can be created so that they could be administered without CAT, e.g. in pencil-and-paper format. In this study, we report on the short form development from the PROMIS SD and SRI item banks. In particular, we report the performance of static 8-item short forms of PROMIS SD and SRI in comparison with their full banks and legacy measures including PSQI and ESS.

# Method

## Sample

Item response data for the SD and SRI item banks were obtained from an internet (YouGov Polimetrix) sample and a clinical sample at the University of Pittsburgh Medical Center. YouGov Polimetrix is a national, web-based polling firm based in Palo Alto, CA. YouGov Polimetrix customized the sample to include individuals with various health conditions (Polimetrix, 2006).

The YouGov Polimetrix sample consisted of 1,993 respondents (41% women, 11% Hispanic, 16% minority, and mean age [S.D.] 52 [15.9]), including 1,259 adults from the general population without self-reported sleep problems, and 734 with self-reported sleep problems. Sleep problems were identified by self report with 4 branching questions: "*Have you ever been told by a doctor or health professional that you have a sleep disorder?*" "*What type of sleep disorder* (with 13 options)*?*" "*Has your sleep disorder been treated?*" and "*Did the treatment help you?*". In order to have adequate observations of each response category for each item, especially for response categories indicating high severity, a separate clinical sample was added to enrich the Polimetrix sample and included 259 patients with sleep problems obtained from sleep medicine clinics in psychiatry and general medicine (61% women, 2% Hispanic, 30% minority, mean age [S.D.] 44 [13.8]). In aggregate, the Polimetrix sample of 1, 993 participants plus the clinical sample of 259 participants, the final pooled sample included 2, 252 participants. For a detailed description of this pooled sample, see Buysse et al. (2010).

## Measures

The "full" SD and SRI item banks consisted of 27 and 16 items each. Respondents rated various aspects of their sleep over the past 7 days on 5-point scales. Most of the items used an intensity scale (*not at all, a little bit, somewhat, quite a bit, very much*), with a smaller number using a frequency scale (*never, rarely, sometimes, often, always*), and one item (*S109*) assessing overall sleep quality using a scale of *very poor, poor, fair, good, very good*. Items assessing sleep disturbance or sleep-related impairment were scored 1 to 5 with 1 for the lowest category (i.e., *not at all*) and 5 for the highest category (i.e., *very much*). In order to be consistent with PROMIS conventions, some items were reverse scored so that, for all items, higher scores corresponded to greater sleep disturbance or sleep-related impairment. Participants also completed two commonly used measures for comparative analyses, the PSQI and the ESS. The PSQI was scored based on standard procedures, with 7 component scores summed together to yield a global score with a range of 0 (good sleep quality) to 21 (poor sleep quality); only the component scores were considered in IRT analyses. The ESS contains 8 items with 4 response categories for each item. ESS items are scored 0 to 3 with 0 for the lowest category and 3 for the highest category. The score for the ESS is obtained by summing the 8 items, and has a range of 0 (no propensity for dozing during daytime activities) to 24 (high propensity for dozing during daytime activities). Demographic and global health information including global health and fatigue items were also collected, as described in Buysse et al (2010).

## Procedures

**Post-hoc CAT simulations—**Post-hoc simulations, also called "real data" simulations, are used to reduce the length of a test that has been administered conventionally. Reise and Henson (2000) showed that a fixed short form, which consists of items most often administered at the start of CAT, performed as well as CATs of the same length. Fixed short forms based on CAT simulations optimize total test information for most individuals. Given that the expected information may vary under different distributions, especially under

distributions with larger standard deviations (Choi, Reise, Pilkonis, Hays & Cella, 2010), a standard normal distribution and a normal distribution with a mean of 0 and standard deviation of 1.5 were both investigated. Items were then rank ordered based on five criteria: Raw score means for each item, discrimination parameters for each item, the percentage of time selected in CAT simulations, and the expected information under the two normal distributions.

The objective of this CAT procedure is to determine how much reduction in test length can be achieved by "re-administering" the items adaptively, without introducing significant changes in the psychometric properties of the test scores. Post-hoc simulations involve the following steps: 1) Use the final item parameter estimates for SD and SRI to estimate each respondent's $\theta$ score using maximum likelihood estimation. 2) Apply CAT with maximum likelihood $\theta$ estimation to adaptively estimate the $\theta$ score for each respondent based on the actual item responses from the calibration sample. 3) Compare the CAT $\theta$ estimates with the conventional test $\theta$ estimates as a function of the numbers of item administered in the CAT. 4) Determine adaptive test lengths that result in greatest similarity between the CAT $\theta$ estimates and those of the conventional test, with a minimum number of CAT items.

Here we use the SD and SRI item banks as an example to illustrate the above simulation procedure. The initial item administered was determined based on maximum information at the mean value of the population distribution of the severity scale ($\theta$). The Maximum Posterior Weighted Information (MPWI) method was used for item selection because MPWI has been demonstrated to perform better than other item selection methods (Choi, 2009). We examined response patterns for every possible length of CAT, from 1-27 items for SD, and 1-16 items for SRI. We used the program Firestar (version 1.2.2; Choi, 2009) to conduct the post-hoc simulations. To be consistent with the literature on the optimal length of short forms (Reise & Henson, 2000), we constrained ourselves to selecting the best 8 items across these criteria. Content experts then reviewed and finalized the short form items from a clinical perspective. See Appendices A and B for the final SD and SRI static short forms with scoring instructions.

**Concurrent calibrations with the PSQI and ESS—**The term calibration has various meanings under different contexts (Angoff, 1971; Linn, 1993; Lord, 1980; Thissen & Wainer, 2001; Kolen & Brennan, 2004). In this paper, concurrent calibration refers to estimating item parameters across multiple measures (i.e., SD/SRI, PSQI, and ESS) on one single computer run. Using the final item parameters for the SD and SRI banks, the PSQI and ESS were calibrated using the Graded Response Model (GRM; Samejima, 1969) with parameters from SD and SRI fixed. This procedure places the PSQI and ESS on the same $\theta$ scales of SD and SRI. The program MULTILOG 7.03 (Thissen, 2003) was used to conduct the concurrent calibrations. Test information curves of full banks, short forms, PSQI, and ESS were then plotted for SD and SRI.

**Preliminary Validity Evidence—**In order to evaluate the convergent and discriminant validity of the final SD and SRI 8-item short forms, short form $\theta$ scores were correlated with their corresponding full banks, PSQI, and ESS. In order to evaluate the face validity of the final SD and SRI 8-item short forms, $\theta$ scores were compared between individuals who did and who did not report a previously-diagnosed sleep disorder. Given the nature of the sample collected from YouGov Polimetrix, we were not able to verify the presence or absence of self-reported clinical diagnoses in that cohort.

# Results

## Sample Characteristics

The combined YouGov Polimetrix and clinical sample (n = 2,252) included 43.8% women and had a mean age (S.D.) of 51 (15.9) years, a median age of 52 years, and 20.7% aged 65 or older. Eighty-two percent were White, 12.6% Black, 2.7% Native American or Alaskan, 0.7% Asian, 0.4% Native Hawaiian or Pacific Islander, and 1.6% unknown. Ten percent of the sample was Hispanic or Latino. Educational attainment ranged from high school or less (13.6%), some college (38.6%), college degree (27.9%), to advanced degree (19.9%). The combined sample had a PSQI mean (S.D.) score of 6.93 (4.57). Fifty five percent of the combined sample were over the PSQI cut-off for poor sleep quality (>5). The combined sample had an ESS mean (S.D.) score of 6.98 (4.30). Twenty percent of the combined sample were over the ESS cut-off for clinically significant sleepiness (>10).

To characterize the health status of the sample, participants were presented with 25 chronic health conditions and asked to identify if a health care professional had ever told them that they had any of these conditions. Twenty-two percent of the total sample reported having none of the 25 conditions, whereas 25% reported one condition, 21% reported two conditions, and 33% reported having three or more. The most frequently reported conditions were hypertension (41%), sleep disorder (40%), depression (33%), arthritis (24%), anxiety (23%), and migraines (20%).

## Post-hoc Simulations

Based on CAT simulations for all items of the SD and SRI item banks, we estimated $\theta$ scores for each respondent from single-item administration to the full-bank administration (i.e., from 1-item administration to 27-item administration for SD and 1-item administration to 16-item administration for SRI). We then correlated each of these $\theta$ scores from CATs with the $\theta$ scores based on the final calibrations of the full-scale SD and SRI item banks. These correlations were plotted as a function of number of items administered, which are represented as lines with diamond symbols in Figures 1 and 2 for SD and SRI. These correlations were very high, indicating that CAT can yield equivalent $\theta$ score estimates with far fewer items. For example, the 2-item CAT for SD provided a $\theta$ score correlation of .95 with the full SD bank, and the 4-item CAT for SRI provided a $\theta$ score correlation of .95 with the full SRI bank.

## Short Form Development

We rank ordered all SD and SRI items based on the following evaluation criteria (Choi et al., 2010): discrimination parameters ($a$), raw score mean, percentage of times selected in CAT across all possible number of items administered, expected information under the standard normal distribution with a mean of 0 and a standard deviation of 1, and the expected information under a normal distribution with a larger standard deviation (i.e., a mean of 0 and a standard deviation of 1.5). Tables 1 and 2 display the rank order results for items in the SD and SRI item banks.

For the SD item bank, the best 8 performing items based on the simulation results (i.e., the last three columns of Tables 1) and the discrimination parameters (i.e., the second column of Table 1) were: *S20: I had a problem with my sleep; S44: I had difficulty falling asleep; S72: I tried hard to get to sleep; S90: I had trouble sleeping; S105: My sleep was restful; S109: My sleep quality was…; S115: I was satisfied with my sleep; and S116: My sleep was refreshing*. The raw score mean criteria (i.e., the third column of Table 1) provided a rationale for the additional selection of *S42: It was easy for me to fall asleep; S87: I had trouble staying asleep; S107: My sleep was deep; and S110: I got enough sleep*. These 12

items were further reviewed by content experts (DJB, DEM, and AG) for clinical importance. *S20* was removed because it was similar to *S90*. *S42* was removed because it contained essentially the opposite wording of *S44*. *S72* was removed because it was similar to *S44*. *S105* and *S107* were removed because of conceptual redundancy. After removing the 5 items, another item, *S108, My sleep was restless*, was added back because the concept of restless sleep is seen as salient by individuals describing their sleep, and was not covered by other short form items. The items in bold in the first column of Table 1 were the final selected items for the SD 8-item short form.

For the SRI item bank, the best 8 performing items based on the expected information under the two normal distributions and the discrimination parameters (i.e., the second and the last two columns of Table 2) were: *S10: I had a hard time getting things done because I was sleepy; S11: I had a hard time concentrating because I was sleepy; S18: I felt tired; S25: I had problems during the day because of poor sleep; S29: My daytime activities were disturbed by poor sleep; S30: I felt irritable because of poor sleep*; and *S33: I had a hard time controlling my emotions because of poor sleep*. The criteria of percentage of time being selected in CAT (i.e., the fourth column of Table 2) added one more item: *S6: I was sleepy during the daytime*. The criteria of raw score mean (i.e., the third column of Table 2), however, provided a rationale for the additional selection of 6 items: *S4: I had enough energy; S19: I tried to sleep whenever I could; S119: I felt alert when I woke up; S120: When I woke up I felt ready to start the day; S123: I had difficulty waking up*; and *S124: I still felt sleepy when I woke up*. These 15 items were further reviewed by content experts (DJB, DEM, and AG) for clinical importance. *S11* was removed because it was similar to *S27* but it had a lower ranking. *S29* was removed because of the desire to achieve content balance between items assessing consequences of poor sleep and items assessing sleepiness: *S29* focused on poor sleep and *S10* focused on sleepiness. *S33* was removed because it was similar to *S30* but it had a lower ranking. *S4, S19, S120, S123*, and *S124* were removed because they had very low rankings under CAT simulations, although they indicated a separate perspective of staying awake. Therefore, one other item, *S7: I had trouble staying awake during the day*, was added back to cover this important clinical perspective after removing the 8 items. The items in bold in the first column of Table 2 were the final selected items for the SRI 8-item short form.

The product-moment correlations between $\theta$ values for the short forms and their corresponding full item banks were very high (0.96 for SD and 0.98 for SRI). We also correlated each of the $\theta$ scores from CATs with the $\theta$ scores based on the 8-item short forms of SD and SRI. We plotted these correlations as a function of the number of items administered for SD and SRI, which are displayed as lines with square symbols in Figures 1 and 2. The lines with square symbols showed the 8-item static short form correlated highly with CAT theta scores. Correlations between CAT and 8-item static short forms (lines with squares) were equivalent or larger than correlations between CAT and full banks (lines with diamond symbols) for CAT simulations with 8 or fewer items.

## Concurrent Calibrations with PSQI and ESS

In order to examine the final SD and SRI item banks and the two legacy measures (ESS and PSQI) on the same scale, items from SD and SRI item banks were calibrated concurrently with ESS and PSQI items, by fixing SD and SRI item parameters to their final bank calibration values. Figures 3 and 4 display the test information curves for the full SD and SRI item bank, SD and SRI short forms, ESS, and PSQI. Overall the full PROMIS SD and SRI item banks provided the greatest test information, followed by the SD and SRI short forms, PSQI, and ESS.

The reliability, or measurement precision, under the IRT framework may vary as a function of $\theta$, whereas the conventional reliability of the test (rho) is fixed. In order to make a direct comparison with conventional reliability, two lines were drawn at the test information values of 10 and 20 in Figures 3 and 4. Test information of 20 corresponds approximately with conventional reliability of .95, and test information of 10 corresponds approximately with conventional reliability of .90. That is, the SD full item bank provided a reliability of .95 or above for respondents with $\theta$ scores from -1.5 to 2.8, and a reliability of .90 or above for respondents with $\theta$ scores from -2 to 3. The short form provided a reliability of .90 or above for respondents with $\theta$ scores from -1.5 to 2.5. The SRI full item bank provided a reliability of .95 or above for respondents with $\theta$ scores from -0.5 to 3.1, and a reliability of .90 for respondents with $\theta$ scores from -1.0 to 3.2. The short form provided a reliability of .90 for respondents with $\theta$ scores from -0.5 to 3.0. Taken together, these results demonstrate that the most precise severity estimates of Sleep Disturbance and Sleep-Related Impairment are provided by the full PROMIS SD and SRI item banks, followed by SD and SRI 8-item static short forms, PSQI, and ESS.

### Preliminary Validity Evidence

We also examined convergent validity between $\theta$ scores for the SD and SRI, both 8-item short forms and full banks, and commonly used measures of sleep-wake functioning, PSQI and ESS. The results, summarized in Table 3, demonstrate the SD and SRI full banks and 8-item short forms yielded similar results. Specifically, we found larger product-moment correlations between SD and PSQI (hypothesized to measure similar attributes) than between SRI and the ESS (hypothesized to measure a related but slightly different construct, the propensity to doze during activities). This expected pattern of results supports the validity of SD and SRI full banks and 8-item short forms. Contrary to expectations, SRI $\theta$ values, both 8-item short form and full bank, correlated more strongly with the PSQI than the ESS. However correlations with the ESS were larger for SRI than for SD, again supporting the validity of SD and SRI full banks and short forms.

In order to evaluate the construct validity of the final SD and SRI 8-item short forms, $\theta$ scores were compared between self-reported sleep disorder and no sleep disorder groups. As hypothesized, subjects reporting each sleep disorder had higher $\theta$ values for both SD and SRI, in both full banks and short forms, compared to those with no sleep disorder (Table 4). These findings suggest that the SD and SRI 8-item short form do, in fact, differ in expected ways among known groups, supporting their construct validity.

## Discussion

IRT analyses of the PROMIS SD and SRI item banks permitted the development of CAT and 8-item static short forms of SD and SRI. Both CAT versions and the two static short forms adequately represent the Sleep Disturbance and Sleep-Related Impairment domains in general population and clinical samples. The static 8-item short forms for SD and SRI were developed based on CAT simulations and clinical judgment from content experts. These short forms correlated strongly with the full SD and SRI item banks and had high total test information and low standard error across a broad range of $\theta$ values. Taken together, these findings provide support for the reliability and validity of the PROMIS SD and SRI item banks and the short forms derived from them.

Scales developed with IRT have several desirable attributes, including the ability to characterize measurement properties of individual items as well as those of an entire scale. By understanding the measurement properties of individual items, investigators can customize item selection to specific applications. The development of short forms reported here was conducted using a sample that included individuals with and without sleep

disorders in order to reflect a wide range of symptom severity. Calibration in a sample of sleep disorder patients alone may have led to the selection of different items for the short forms, reflecting a higher level of severity.

An additional benefit of IRT is the possibility of administering a PRO using CAT methods. CAT uses individual item measurement properties to develop a progressively more precise estimate of an individual's severity, as described in the Introduction. Depending on the desired level of precision, CAT administration of PROs typically require responses to only 5-8 items as an alternative to “fixed” forms containing 2-4 times as many items (e.g., Gardner et al., 2004). The CAT method resembles the information-seeking practices of skilled diagnosticians, who zero in on precise diagnostic questions after reviewing answers to screening questions.

Scoring of IRT-calibrated item banks differs from CTT-derived scales. The limitation of IRT-derived scoring is that it requires access to programs such as MULTILOG (Thissen, 2003). For this reason, SD and SRI short form conversion tables were created by the PROMIS Cooperative Group and can be found in the PROMIS User Manual (Version 1.1; PROMIS Cooperative Group, 2008). With these tables, shown in Appendix C, the static SD and SRI short forms can be administered in pencil-and-paper format. In this case, the individual items for a respondent can be summed, and the corresponding $\theta$ scores or T-scores estimated from a nonlinear transformation contained in a conversion table. Although the conversion tables for short form scores are simple and convenient, they do not offer the same measurement precision as MULTILOG scoring. The conversion table assumes that any combination of item scores yielding the same total score are equivalent, whereas MULTILOG uses the unique calibrated values for each response to each item, with different responses and different items measuring different levels of severity. Thus, scoring with MULTILOG is encouraged in order to take advantage of IRT calibration and provide the most precise estimates.

The five criteria for selecting short form items reflected a combination of psychometric and clinical input. From the rankings of Tables 1 and 2, we found the rankings based on item discrimination parameters ($a$), percentage of times selected in CAT across all possible number of items administered, and expected information under the two distributions were quite consistent. There was less consistency with the ranking based on raw score mean. This finding is expected because all criteria are based on IRT analysis except the raw score mean, which is a typical CTT indicator. Although raw score mean was not an important criterion for whether an item would be retained or dropped, we report this information to facilitate a comparison of IRT and CTT results.

The PROMIS SD and SRI item banks have multiple potential uses, including the characterization of clinical and research samples with or without sleep disorders. For instance, it would be possible to select even fewer items than the 8-item short forms for an epidemiological study with a normal population that measures lower levels of severity by choosing items with low threshold parameters. Similarly, a clinical trial of a new medication for sleep-related impairment might employ high threshold items if sensitivity to change over time in a severely affected sample. The PROMIS SD and SRI, whether in full-scale, CAT, or short-form versions, will be most appropriately used for clinical and research applications that require unidimensional severity scales. Since the PROMIS scales did not include actual clock time items and were intended to provide generic “thermometers,” they are not appropriate for deriving “quantitative” estimates of sleep such as total sleep time or sleep onset latency, nor for measuring the symptoms of specific sleep disorders. The one exception may be for insomnia. Sleep quality and sleep dissatisfaction appear to exist on a continuum of severity, with good sleep represented at one end and insomnia at the other

(Buysse et al., 1989; Ohayon & Partinen, 2002; Ohayon & Smirne, 2002; Ohayon & Paiva, 2005). Thus, the PROMIS scales may prove useful for grading the global severity of insomnia.

Although SD and SRI short forms have many advantages, we cannot recommend the SD and SRI scales “instead of” PSQI or ESS until further validation work is done in more of the settings where PSQI and ESS have been used. Given the correlations among the SD, SRI, ESS, and PSQI, users of the instruments also need to recognize that there are other important differences, e.g., PSQI includes actual time-based quantitative items, which SD and SRI does not.

Taken together, the findings of this paper demonstrate the precision and efficiency that the 8-item SD and SRI short forms provide compared with their corresponding full scales and the two commonly used scales (PSQI and ESS). SD and SRI 8-item short forms may prove useful in both research and clinical settings.

## Acknowledgments

The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project have included Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Laura Lee Johnson, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Susana Serrate-Sztein, MD, and James Witter, MD, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee, and their comments incorporated, prior to external peer review. See the web site at www.nihpromis.org for additional information on the PROMIS cooperative group.

The authors gratefully acknowledge the support and guidance of Seung Choi, PhD at the statistical coordinating center for the PROMIS network.

## Appendix A: PROMIS Sleep Disturbance short form

Please respond to each item by marking one box per row.

| In the past 7 days… | | Not at all | A little bit | Somewhat | Quite a bit | Very much |
|---|---|---|---|---|---|---|
| Sleep108 | My sleep was restless…….... | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep115 | I was satisfied with my sleep… | ☐ 5 | ☐ 4 | ☐ 3 | ☐ 2 | ☐ 1 |
| Sleep116 | My sleep was refreshing…….. | ☐ 5 | ☐ 4 | ☐ 3 | ☐ 2 | ☐ 1 |
| Sleep44 | I had difficulty falling asleep... | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| **In the past 7 days…** | | **Never** | **Rarely** | **Sometimes** | **Often** | **Always** |
| Sleep87 | I had trouble staying asleep…. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep90 | I had trouble sleeping……... | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep110 | I got enough sleep…….. | ☐ 5 | ☐ 4 | ☐ 3 | ☐ 2 | ☐ 1 |
| **In the past 7 days…** | | **Very poor** | **Poor** | **Fair** | **Good** | **Very good** |
| Sleep109 | My sleep quality was…….. | ☐ 5 | ☐ 4 | ☐ 3 | ☐ 2 | ☐ 1 |

## Appendix B: PROMIS Sleep-Related Impairment Short Form

Please respond to each item by marking one box per row.

| In the past 7 days… | | Not at all | A little bit | Somewhat | Quite a bit | Very much |
|---|---|---|---|---|---|---|
| Sleep10 | I had a hard time getting things done because I was sleepy…… | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep119 | I felt alert when I woke up…… | ☐ 5 | ☐ 4 | ☐ 3 | ☐ 2 | ☐ 1 |
| Sleep18 | I felt tired…… | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep25 | I had problems during the day because of poor sleep…….. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep27 | I had a hard time concentrating because of poor sleep…….. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep30 | I felt irritable because of poor sleep…… | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep6 | I was sleepy during the daytime…… | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| Sleep7 | I had trouble staying awake during the day…… | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |

## Appendix C: PROMIS Sleep Disturbance and Sleep-Related Impairment conversion table

| Raw Score | Sleep Disturbance | | Sleep-Related Impairment | |
|---|---|---|---|---|
| | T-score | SE | T-score | SE |
| 8 | 28.9 | 4.8 | 30.0 | 5.4 |
| 9 | 33.1 | 3.7 | 35.1 | 4.6 |
| 10 | 35.9 | 3.3 | 38.7 | 4.2 |
| 11 | 38.0 | 3.0 | 41.4 | 3.8 |
| 12 | 39.8 | 2.9 | 43.6 | 3.6 |
| 13 | 41.4 | 2.8 | 45.5 | 3.4 |
| 14 | 42.9 | 2.7 | 47.3 | 3.1 |
| 15 | 44.2 | 2.7 | 48.9 | 2.9 |
| 16 | 45.5 | 2.6 | 50.3 | 2.7 |
| 17 | 46.7 | 2.6 | 51.6 | 2.6 |
| 18 | 47.9 | 2.6 | 52.9 | 2.6 |
| 19 | 49.0 | 2.6 | 54.0 | 2.5 |
| 20 | 50.1 | 2.5 | 55.1 | 2.5 |
| 21 | 51.2 | 2.5 | 56.1 | 2.5 |
| 22 | 52.2 | 2.5 | 57.2 | 2.5 |
| 23 | 53.3 | 2.5 | 58.2 | 2.4 |
| 24 | 54.3 | 2.5 | 59.3 | 2.4 |
| 25 | 55.3 | 2.5 | 60.3 | 2.4 |
| 26 | 56.3 | 2.5 | 61.3 | 2.4 |
| 27 | 57.3 | 2.5 | 62.3 | 2.3 |
| 28 | 58.3 | 2.5 | 63.3 | 2.3 |
| 29 | 59.4 | 2.5 | 64.3 | 2.3 |
| 30 | 60.4 | 2.5 | 65.3 | 2.3 |
| 31 | 61.5 | 2.5 | 66.3 | 2.3 |
| 32 | 62.6 | 2.5 | 67.3 | 2.3 |

| Raw Score | Sleep Disturbance | | Sleep-Related Impairment | |
|---|---|---|---|---|
| | T-score | SE | T-score | SE |
| 33 | 63.7 | 2.6 | 68.4 | 2.3 |
| 34 | 64.9 | 2.6 | 69.5 | 2.4 |
| 35 | 66.1 | 2.7 | 70.7 | 2.4 |
| 36 | 67.5 | 2.8 | 71.9 | 2.5 |
| 37 | 69.0 | 3.0 | 73.3 | 2.6 |
| 38 | 70.8 | 3.2 | 75.0 | 2.8 |
| 39 | 73.0 | 3.5 | 76.9 | 3.1 |
| 40 | 76.5 | 4.4 | 80.0 | 3.9 |

Note:

Conversion table applies only when ALL items on the short form have been answered.

T-score metric is a linear transformation from the IRT theta scale: T-score=10*theta+50 SE in the table is the standard error on T-score metric.

## References


Angoff, WH. Scales, norms and equivalent scores. Princeton, NJ: Educational Testing Service; 1971.

Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. Psychiatry Research. 1989; 28:193–213. [PubMed: 2748771]

Buysse DJ, Yu L, Moul DE, Germain A, Stover A, Dodds NE, Johnston KL, Shablesky-Cade MA, Pilkonis PA. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairment. Sleep. 2010; 33(6):781–792. [PubMed: 20550019]

Choi SW. Firestar: Computerized Adaptive Testing (CAT) simulation program for polytomous IRT models. Applied Psychological Measurement. 2009; 33:644–645. [PubMed: 20011609]

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Quality of Life Research. 2010; 19(1):125–136. [PubMed: 19941077]

Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, et al. Computerized adaptive measurement of depression: A simulation study. BMC Psychiatry. 2004; 4:13. [PubMed: 15132755]

Gulliksen, H. Theory of mental tests. New York: Wiley; 1950.

Linn RL. Linking results of distinct assessments. Applied Measurement in Education. 1993; 6(1):83–102.

Lord, F. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum; 1980.

Lord, FN.; Novick, MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.

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

Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. Sleep. 1992; 15:376–381. [PubMed: 1519015]

Kolen, MJ.; Brennan, RL. Test equating, scaling, and linking Methods and practices. 2nd. New York: Springer; 2004.

Ohayon MM, Paiva T. Global sleep dissatisfaction for the assessment of insomnia severity in the general population of Portugal. Sleep Medicine. 2005; 6:435–441. [PubMed: 16085459]

Ohayon MM, Partinen M. Insomnia and global sleep dissatisfaction in Finland. Journal of Sleep Research. 2002; 11:339–346. [PubMed: 12464102]

Ohayon MM, Smirne S. Prevalence and consequences of insomnia disorders in the general population of Italy. Sleep Medicine. 2002; 3:115–120. [PubMed: 14592229]

Polimetrix Inc.. Scientific sampling for online research. Palo Alto, CA: Author; 2006.

PROMIS Cooperative Group. User Manual: Patient-reported outcomes measurement information system (PROMIS), Version 1.1. Unpublished Manual for the Patient Reported Outcomes Measurement Information System (PROMIS). 2008. www.nihpromis.org

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai J, Cella D. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). Medical Care. 2007; 45(Suppl 1):S22–S31. [PubMed: 17443115]

Reise SP, Henson JM. Computerization and adaptive administration of the NEO PI-R. Assessment. 2000; 7:347–364. [PubMed: 11151961]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, No 17. 1969

Takegami M, Suzukamo Y, Wakita T, Noguchi H, Chin K, Kadotani H, Inoue Y, Oka Y, Nakamura T, Green J, Johns MW, Fukuhara S. Development of a Japanese version of the Epworth Sleepiness Scale (JESS) based on item response theory. Sleep Medicine. 2009; 10(5):556–565. [PubMed: 18824408]

Thissen, D. MULTILOG 7: Multiple categorical item analysis and test scoring using item response theory [computer program]. Chicago: IL: Scientific Software; 2003.

Thissen, D.; Wainer, H. Test scoring. Mahwah, NJ: Erlbaum; 2001.

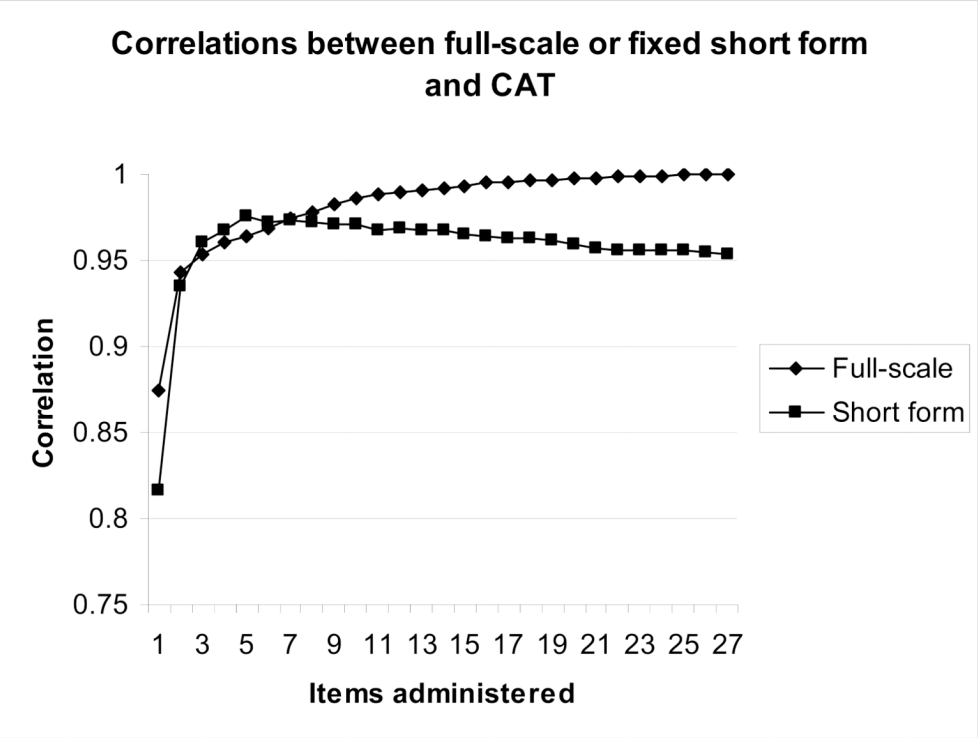


**Figure 1. Correlations between full-scale or fixed short form and CAT for Sleep Disturbances**

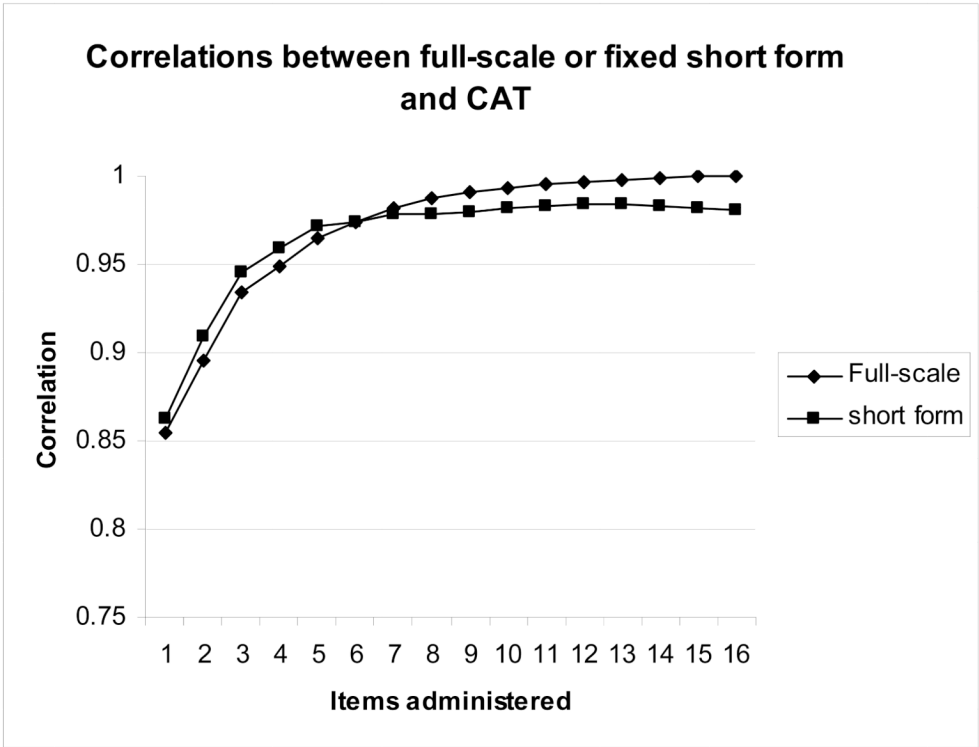


**Figure 2. Correlations between full-scale or fixed short form and CAT for Sleep-related Impairment item bank**

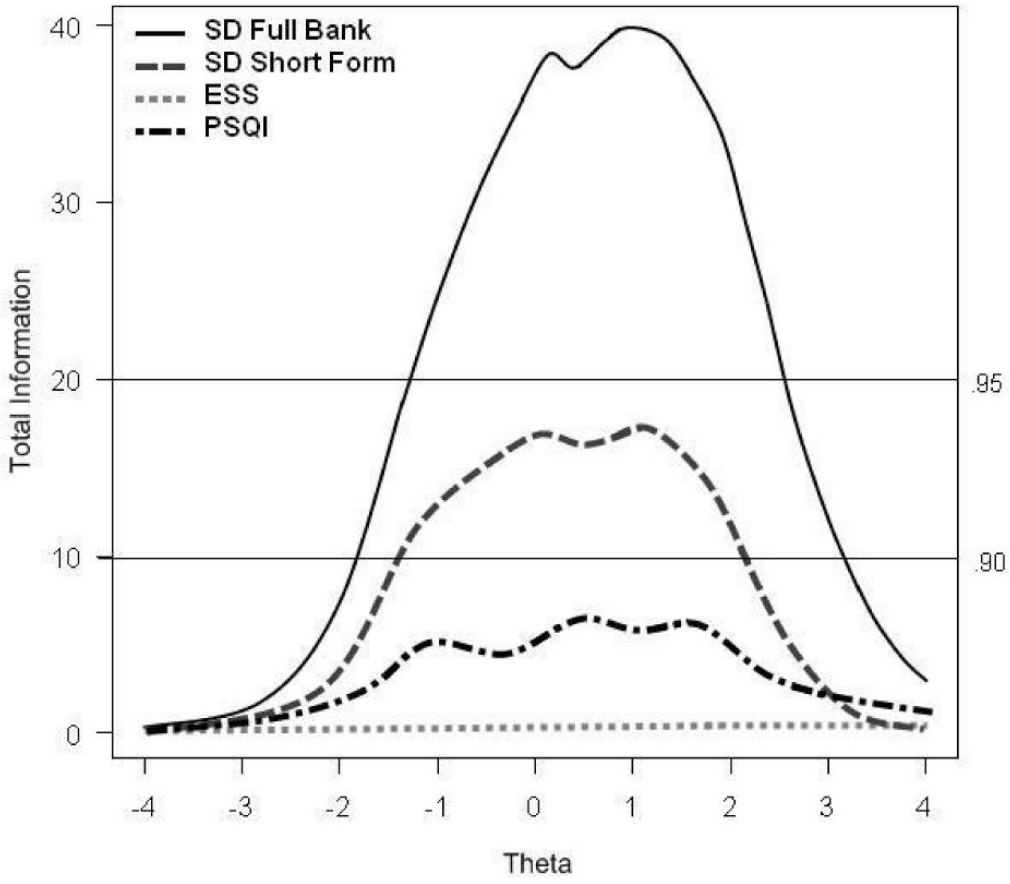


**Figure 3. Test Information Curves for the PROMIS Sleep Disturbance Full Item Bank, Short Form, Epworth Sleepiness Scale (ESS), and Pittsburgh Sleep Quality Index (PSQI)**

Note: The test information of 10 derived from IRT on the left-side y-axis is roughly equivalent to the reliability of .90 derived from CTT on the right-side y-axis. Therefore, the curves above the horizontal line (test information of 10 to reliability of .90) indicate the section on the theta scale has reliability of .90 or above.

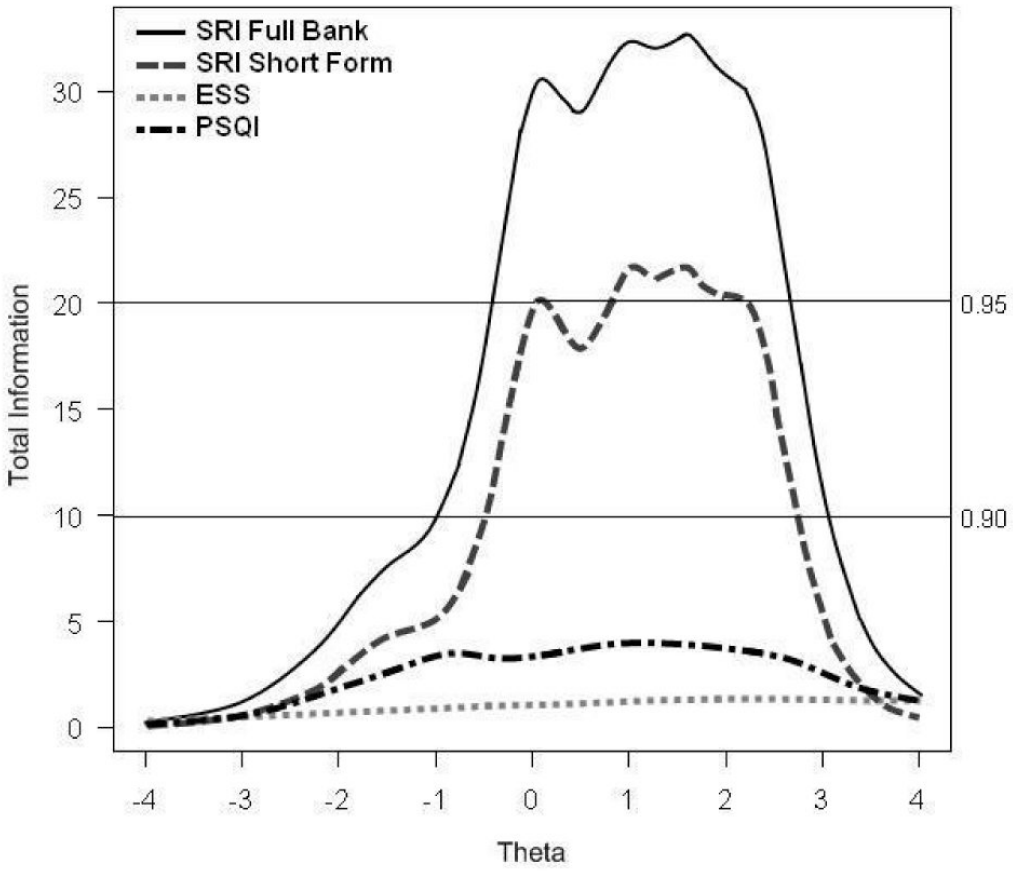


**Figure 4. Test Information Curves for the PROMIS Sleep-Related Impairment Full Item Bank, Short Form, Epworth Sleepiness Scale (ESS), and Pittsburgh Sleep Quality Index (PSQI)**

Note: The test information of 10 derived from IRT on the left-side y-axis is roughly equivalent to the reliability of .90 derived from CTT on the right-side y-axis. Therefore, the curves above the horizontal line (test information of 10 to reliability of .90) indicate the section on the theta scale has reliability of .90 or above.

**Table 1**

**Short form item selection order of PROMIS Sleep Disturbance item bank**

| Items[1] | *a*[2] | Raw score mean[3] | % of times selected in CAT simulation[4] | Expected Informatio n for distribution (0, 1)[5] | Expected Information for distribution (0, 1.5)[6] |
|---|---|---|---|---|---|
| **S90: I had trouble sleeping.** | 1 | 9 | 1 | 2 | 1 |
| **S109: My sleep quality was…** | 2 | 7 | 2 | 1 | 2 |
| S20: I had a problem with my sleep. | 3 | 11 | 6 | 4 | 3 |
| **S115: I was satisfied with my sleep.** | 4 | 1 | 3 | 3 | 4 |
| **S116: My sleep was refreshing.** | 5 | 2 | 4 | 5 | 6 |
| **S44: I had difficulty falling asleep.** | 6 | 12 | 7 | 7 | 5 |
| S72: I tried hard to get to sleep. | 7 | 19 | 8 | 10 | 8 |
| S105: My sleep was restful. | 8 | 5 | 5 | 6 | 7 |
| S67: I worried about not being able to fall asleep. | 9 | 25 | 11 | 13 | 10 |
| **S108: My sleep was restless.** | 10 | 18 | 13 | 9 | 9 |
| **S87: I had trouble staying asleep.** | 11 | 8 | 9 | 12 | 12 |
| S45: I laid in bed for hours waiting to fall asleep. | 12 | 22 | 15 | 15 | 13 |
| **S110: I got enough sleep.** | 13 | 4 | 10 | 8 | 14 |
| S92: I woke up and had trouble falling back to sleep. | 14 | 15 | 16 | 11 | 11 |
| S42: It was easy for me to fall asleep. | 15 | 6 | 12 | 14 | 15 |
| S78: Stress disturbed my sleep. | 16 | 21 | 17 | 16 | 16 |
| S93: I was afraid I would not get back to sleep after waking up. | 17 | 23 | 17 | 18 | 17 |
| S125: I felt lousy when I woke up. | 18 | 17 | 16 | 19 | 18 |
| S86: I tossed and turned at night. | 19 | 16 | 17 | 17 | 19 |
| S68: I felt worried at bedtime. | 20 | 26 | 18 | 21 | 20 |
| S69: I had trouble stopping my thoughts at bedtime. | 21 | 10 | 18 | 22 | 21 |
| S65: I felt physically tense at bedtime. | 22 | 24 | 16 | 23 | 22 |
| S107: My sleep was deep. | 23 | 3 | 14 | 20 | 23 |
| S71: I had trouble getting into a comfortable position to sleep. | 24 | 20 | 18 | 25 | 25 |
| S106: My sleep was light. | 25 | 14 | 18 | 24 | 24 |
| S70: I felt sad at bedtime. | 26 | 27 | 18 | 26 | 26 |

| **Items**[1] | ***a***[2] | **Raw score mean**[3] | **% of times selected in CAT simulation**[4] | **Expected Informatio n for distribution (0, 1)**[5] | **Expected Information for distribution (0, 1.5)**[6] |
|---|---|---|---|---|---|
| S50: I woke up too early and could not fall back asleep. | 27 | 13 | 18 | 27 | 27 |

[1] Items in bold are contained in the static short form, and the “S” in front of item number stands for Sleep.

[2] a = ranks based on discrimination parameter (how well the item discriminates between respondents' with low or high symptom levels).

[3] raw score mean=ranks based on arithmetic mean from the original scoring

[4] % of times selected in CAT simulations=ranks based on number of times that each item being selected in CAT simulations.

[5] expected information for distribution of (0, 1) =ranks based on expected information that each item has under the normal distribution with a mean of 0 and standard deviation of 1.

[6] expected information for distribution of (0, 1.5) =ranks based on expected information that each item has under the distribution with a mean of 0 and standard deviation of 1.5.

**Table 2**

**Short form item selection order of PROMIS Sleep-related Impairment item bank**

| Items[1] | *a*[2] | Raw score mean[3] | % of times selected in CAT simulation[4] | Expected Information for distribution (0, 1)[5] | Expected Information for distribution (0, 1.5)[6] |
|---|---|---|---|---|---|
| **S27: I had a hard time concentrating because of poor sleep.** | 1 | 15 | 2 | 1 | 1 |
| **S25: I had problems during the day because of poor sleep.** | 2 | 10 | 1 | 2 | 2 |
| S29: My daytime activities were disturbed by poor sleep. | 3 | 11 | 3 | 3 | 3 |
| **S10: I had a hard time getting things done because I was sleepy.** | 4 | 14 | 6 | 4 | 4 |
| S11: I had a hard time concentrating because I was sleepy. | 5 | 12 | 5 | 5 | 5 |
| **S30: I felt irritable because of poor sleep.** | 6 | 13 | 7 | 6 | 7 |
| **S18: I felt tired.** | 7 | 5 | 4 | 7 | 6 |
| S33: I had a hard time controlling my emotions because of poor sleep. | 8 | 16 | 10 | 8 | 8 |
| **S6: I was sleepy during the daytime.** | 9 | 6 | 8 | 9 | 9 |
| **S7: I had trouble staying awake during the day.** | 10 | 9 | 13 | 10 | 10 |
| S120: When I woke up I felt ready to start the day. | 11 | 1 | 9 | 12 | 12 |
| S4: I had enough energy. | 12 | 4 | 11 | 11 | 11 |
| S124: I still felt sleepy when I woke up. | 13 | 3 | 16 | 13 | 13 |
| **S119: I felt alert when I woke up.** | 14 | 2 | 12 | 14 | 14 |
| S19: I tried to sleep whenever I could. | 15 | 7 | 14 | 15 | 15 |
| S123: I had difficulty waking up. | 16 | 8 | 15 | 16 | 16 |

[1] Items in bold are contained in the static short form, and the "S" in front of item number stands for Sleep.

[2] a =ranks based on slope parameter (how well the item discriminates between respondents' with low or high symptom levels).

[3] raw score mean=ranks based on arithmetic mean from the original scoring

[4] % of times selected in CAT simulations=ranks based on number of times that each item being selected in CAT simulations.

[5] expected information for distribution of (0, 1) =ranks based on expected information that each item has under the normal distribution with a mean of 0 and standard deviation of 1.

[6] expected information for distribution of (0, 1.5) =ranks based on expected information that each item has under the distribution with a mean of 0 and standard deviation of 1.5.

**Table 3**

**Convergent and Discriminant Validity of the PROMIS Sleep Disturbance and Sleep-Related Impairment Item Banks (*correlations*)**

| | PSQI | ESS |
|---|---|---|
| SD full item bank (27 items) | .85 | .25 |
| SD short form (8 items) | .83 | .30 |
| SRI full item bank (16 items) | .70 | .45 |
| SRI short form (8 items) | .68 | .46 |

**Table 4**

**Known-Groups Validity of the PROMIS Sleep Disturbance and Sleep-Related Impairment Item Banks**

| | Mean $\theta$ score (standard deviation) | | | | *p* value for pairwise comparisons I, A, R vs. N |
|---|---|---|---|---|---|
| | **No Sleep Disorder (N) *n* = 1342** | **Insomnia (I) *n* = 358** | **Apnea (A) *n* = 504** | **Restless Legs Syndrome (R) *n* = 132** | |
| SD Full Bank $\theta$ | -.27 (.97) | 1.00 (.76) | -.06 (.96) | .73 (.89) | <.001 (I, A, R) |
| SD Short Form $\theta$ | -.25 (.89) | .85 (.74) | -.01 (.90) | .70 (.83) | <.001 (I, A, R) |
| SRI Full Bank $\theta$ | -.25 (.91) | .75 (.85) | .14 (.91) | .71 (.91) | <.001 (I, A, R) |
| SRI Short Form $\theta$ | -.23 (.84) | .68 (.84) | .14 (.92) | .64 (.92) | <.001 (I, A, R) |