Development and Validation of Patient-Reported Outcome Measures for Sleep Disturbance and Sleep-Related Impairments

Daniel J. Buysse, MD[1,2]; Lan Yu, PhD[1]; Douglas E. Moul, MD, MPH[3]; Anne Germain, PhD[1,2]; Angela Stover, MA[1]; Nathan E. Dodds, BS[1]; Kelly L. Johnston, MPH[1]; Melissa A. Shablesky-Cade[1]; Paul A. Pilkonis, PhD[1]

[1]*Department of Psychiatry and* [2]*Sleep Medicine Institute, University of Pittsburgh School of Medicine;*
[3]*Departments of Psychiatry and Neurology, Louisiana State University in Shreveport*

Study Objectives: To develop an archive of self-report questions assessing sleep disturbance and sleep-related impairments (SRI), to develop item banks from this archive, and to validate and calibrate the item banks using classic validation techniques and item response theory analyses in a sample of clinical and community participants.
Design: Cross-sectional self-report study.
Setting: Academic medical center and participant homes.
Participants: One thousand nine hundred ninety-three adults recruited from an Internet polling sample and 259 adults recruited from medical, psychiatric, and sleep clinics.
Interventions: None.
Measurements and Results: This study was part of PROMIS (Patient-Reported Outcomes Information System), a National Institutes of Health Roadmap initiative. Self-report item banks were developed through an iterative process of literature searches, collecting and sorting items, expert content review, qualitative patient research, and pilot testing. Internal consistency, convergent validity, and exploratory and confirmatory factor analysis were examined in the resulting item banks. Factor analyses identified 2 preliminary item banks, sleep disturbance and SRI. Item response theory analyses and expert content review narrowed the item banks to 27 and 16 items, respectively. Validity of the item banks was supported by moderate to high correlations with existing scales and by significant differences in sleep disturbance and SRI scores between participants with and without sleep disorders.
Conclusions: The PROMIS sleep disturbance and SRI item banks have excellent measurement properties and may prove to be useful for assessing general aspects of sleep and SRI with various groups of patients and interventions.

Keywords: Sleep, wake, measurement, questionnaire, psychometric, item response theory
Citation: 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 impairments. *SLEEP* 2010;33(6):781-792.

SLEEP AND WAKEFULNESS ARE FUNDAMENTAL NEUROBEHAVIORAL STATES THAT CAN BE DISRUPTED BY A VARIETY OF PHYSIOLOGIC, BEHAVIORAL, AND environmental factors. Approximately 30% to 40% of adults complain of insomnia,[1] and 5% to 15% complain of excessive sleepiness.[2] Conversely, sleep-wake disturbances are strongly related to health. Sleep duration, sleep quality, and daytime sleepiness are risk factors for obesity, weight gain, hypertension, metabolic syndrome, vulnerability to the common cold, major depression, and all-cause mortality.[3-10] Sleep deprivation in humans is associated with changes in vigilance and psychomotor performance, mood and affect regulation, memory consolidation, moral reasoning, metabolic and appetite regulation, and immune function.[11-15] Developing research and clinical tools to rapidly and accurately assess sleep disturbances and sleep-related impairments (SRI) is, therefore, a high priority.

Human sleep and SRI can be evaluated in a variety of ways. Functional imaging studies during sleep and wakefulness provide a detailed examination of regional metabolic rate, blood flow, and activation patterns[16,17] but are confined for practical reasons to small numbers of highly selected participants. Electrophysiologic techniques, including electroencephalography, topographic mapping, event-related potentials, and polysomnography, are the current "gold standard" for characterizing sleep-wake patterns and disturbances in clinical and epidemiologic samples but require specialized equipment and facilities. Actigraphy relies on the strong correlation between sleep-wake state and motor activity to provide convenient and inexpensive estimates of sleep and wake times.[18,19]

However, self-report instruments remain the most practical and widely used tools to characterize sleep-wake function. A wide variety of published instruments address different research and clinical goals. For instance, questionnaires have been used in many epidemiologic studies to evaluate general types of sleep disturbance (e.g., difficulty falling asleep, difficulty staying asleep, daytime sleepiness, napping),[20,21] but many of these are not standardized and have not been validated. Other instruments assess temporal and quantitative aspects of sleep and wakefulness, either on a habitual basis or on a daily basis (sleep-wake diaries). Most sleep diaries and habitual-timing questionnaires have not been formally validated (with some exceptions[22,23]). Still other instruments characterize symptoms of specific sleep disorders (e.g., restless legs syndrome,[24,25] sleep apnea,[26,27] narcolepsy,[28] insomnia,[29-33] or multiple sleep disorders[34,35]). Finally, some self-report instruments are designed to quantify latent constructs, such as sleep quality or likelihood of sleep, that are not readily measurable with other tools.[36,37]

Submitted for publication July, 2009
Submitted in final revised form February, 2010
Accepted for publication March, 2010

Address correspondence to: Dr. Buysse, 3811 O'Hara ST, Pittsburgh, PA 15213; Tel: (412) 246-6413; Fax: (412) 246-5300; E-mail: buyssedj@upmc.edu

Despite their widespread use, self-report measures of sleep-wake function face several methodologic challenges. First, awareness is reduced during sleep, and sleep onset is associated with brief retrograde amnesia.[38] This limits the validity of sleep self-reports, relative to other methods, for assessing quantitative variables and phenomena such as snoring, apnea, or leg jerks. A second challenge is that sleep usually occurs in a single, relatively consolidated, block every 24 hours, and sleep-wake function can vary considerably from day to day. Thus, sleep-wake self-reports require longer time frames than do self-reports for other common symptoms and must survey an adequate number of days to derive stable estimates. Third, self-reported sleep can be characterized along multiple dimensions, including sleep quality, quantitative aspects such as durations and numbers of events, timing within the 24-hour day, and specific sleep-related symptoms. Several aspects of waking function, such as sleepiness, fatigue, cognitive efficiency, and emotional control, are also related to nocturnal sleep characteristics and disturbances.

No self-report instrument can address all of these challenges. However, there are currently no validated, flexible, brief instruments that measure general aspects of sleep and SRI across the entire range of the adult populations. Of the available self-report instruments, the Pittsburgh Sleep Quality Index (PSQI)[39] and the Epworth Sleepiness Scale (ESS)[40,41] are the most widely used, with several thousand citations in aggregate. However, each scale has limitations. For instance, the 7 component scores of the PSQI are not consistent with factor analytic solutions,[42] and the individual items and components have not been carefully evaluated psychometrically. The ESS is simple to complete and score but has been criticized on the basis of content validity (e.g., inclusion of some situations an individual may not routinely encounter and omitting other important situations such as work) and limited correlation with other measures of sleepiness, such as the Multiple Sleep Latency Test.[43-45] Previous reviews have documented the properties, strengths, and weaknesses of other self-report instruments for sleep-wake function, particularly those related to insomnia.[46-48]

The PROMIS (Patient-Reported Outcomes Information System) sleep disturbance and SRI item banks were developed to improve self-report instrumentation of sleep-wake function. PROMIS (www.nihpromis.org) is a National Institutes of Health-funded consortium that aims to build item pools and develop core questionnaires that measure key health-outcome domains manifested in a variety of chronic diseases. PROMIS also aims to build an electronic Web-based resource for administering computerized adaptive tests, collecting self-report data, and providing instant reports of the health assessments (see www.assessmentcenter.net). PROMIS item banks have been developed through a systematic process of literature reviews, expert consensus, qualitative research methods, classic test theory (CTT) methods, and item response theory (IRT) analyses. These methods have been designed to calibrate individual items for high precision and minimal respondent biases across major symptom domains affecting health status and quality of life.

In the setting of the broader PROMIS objectives, the specific aims of the sleep-wake project were (1) To develop an archive of self-report measures that assess sleep and SRI, (2) to develop item banks from these measures that assess sleep disturbance and SRI, (3) to test the item banks in broad samples of patients and community participants to determine the dimensionality of sleep-wake symptoms and to identify the psychometric properties of individual items using IRT models, and (4) to examine the validity of the new item banks against widely used existing measures (PSQI, ESS).

METHODS

Overview

Development of the sleep-wake item banks was a single-site project within the broader multisite PROMIS initiative. The methods for this process were similar to those used for the other PROMIS item banks[49-52] and included the articulation of a conceptual framework, development of the item banks, testing of the initial item bank, and psychometric analyses using both CTT and IRT.

Conceptual framework

The PROMIS domain framework (see http://www.nihpromis.org) is drawn from the World Health Organization's framework of physical, mental, and social health domains.[53] Within this framework, PROMIS places sleep-wake function as a physical health measure that is influenced by mental health. Investigators on the sleep-wake project generated a list of 17 potentially distinct conceptual categories across the broad spectrum of sleep-wake function (see Table S1 in supplemental material available online only at www.journalsleep.org). Categories included qualitative, quantitative, behavioral, and symptom-based dimensions of sleep, as well as domains assessing sleepiness and the perceived daytime correlates of nocturnal sleep.

Development of Item Banks

Development of the item banks included 4 major steps (Figure 1).

Literature review

Comprehensive literature searches ensured broad content validity. An earlier literature review on instruments related to insomnia[48] helped to identify more than 100 sleep questionnaires and almost 3000 items. Health science librarians at the University of Pittsburgh conducted a more systematic literature search in Medline, Psych Info, and Health and Psychosocial Instruments databases using a list of 291 sleep-wake search terms developed by the research team. Sleep-wake search terms were crossed with measurement terms (e.g., *validity*, *psychometric*) to focus the search field. The search identified 535 candidate citations, of which 126 were further examined by the content experts and 71 were found to have adequate psychometric documentation. The 2 literature search and review processes yielded a final pool of 82 questionnaires, 2529 sleep items, and a refined conceptual model of the sleep-wake domain.

Item banking

An Access database was created for the initial pool of items, which were coded into the 17 sleep-wake content conceptual categories and further subcoded into 76 "bins." These codings were based on independent content expert (DJB, DEM, AG) ratings, with resolution of disagreements by discussion and consensus. Qualitative item review based on the larger PROMIS Network protocol[51] then substantially reduced the number of items by deleting those with redundant content. Items were rewritten to be consistent with PROMIS Network standards of verb tense, time frame, and response set. For most items, a 7-day time frame, first-person subject, past tense, and either frequency scaling (*never, rarely, sometimes, often*, and *always*) or intensity scaling (*not at all, a little bit, somewhat, quite a bit*, and *very much*) were used. For some items (e.g., those referring to infrequent behaviors) a 1-month (28-day) time frame was used, and, for other items, different response options were deemed more appropriate to their content (e.g., sleep quality responses range from *very poor* to *very good*). After qualitative item and expert item reviews, 310 items in 53 bins were retained for further testing (see supplemental material).

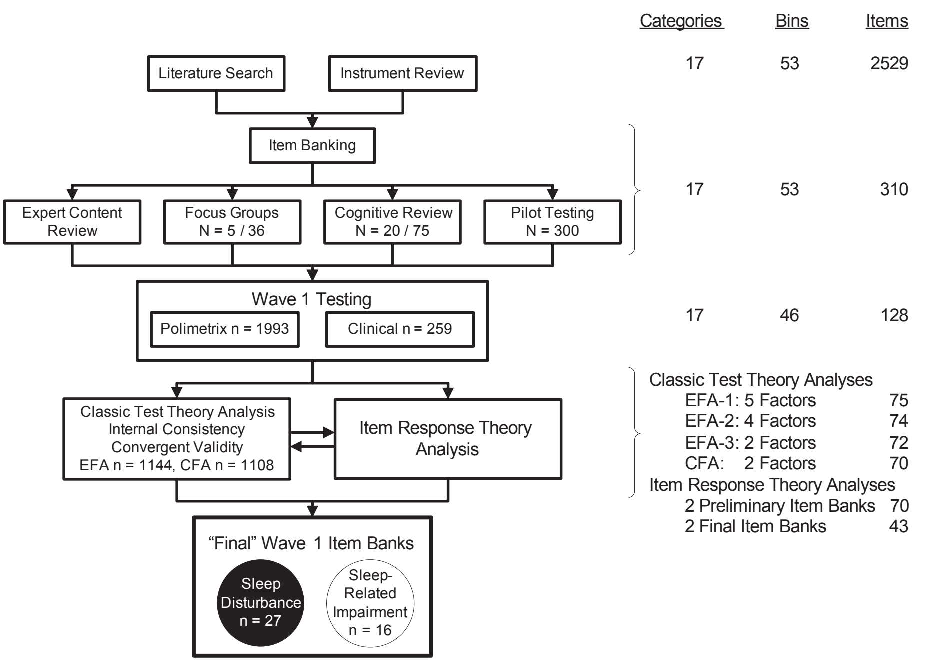

Figure 1— Development of PROMIS sleep-wake function item banks. The flow chart on the left illustrates the major steps in the development of the Patient-Reported Outcomes Information System (PROMIS) sleep-wake function instruments. The table on the right side indicates the number of categories, subcategories, and items at each stage of the item bank development. EFA refers to exploratory factor analysis; CFA, confirmatory factor analysis.

Qualitative focus group research

Focus groups provide essential patient input in the development of patient-reported outcomes.[54] Five groups were conducted: 2 sleep disorder groups, 2 sleep disorder and psychiatric patient groups, and 1 group of normal sleepers. Thirty-six participants (64% women, 39% minority, 31% married, 50% with a college or graduate degree, mean age 45.3 years, sleep disturbance 13.8 years, range 23-80 years) were recruited from sleep medicine centers, outpatient clinics, and advertisements. Focus group facilitators elicited participants' perceptions of sleep symptoms and difficulties, sleep patterns, bedtime and wake time routines, mood symptoms and their interactions with sleep difficulties, daytime alertness, sleepiness, fatigue, and functioning in relationship to sleep. A preliminary qualitative review of participant comments suggested themes of a lack of understanding of sleep problems by family and health-care workers, the unpredictability of sleep, the substantial effects of sleep problems on waking function, and the effort required to cope with sleep problems. These themes led to the inclusion of 10 additional items for initial testing.

Cognitive interviews and Lexile analyses

Cognitive interviews and Lexile analyses were used to evaluate whether proposed items were readily understandable. Seventy-five items were selected for cognitive interviewing in 20 participants (55% women, 30% minority, 30% married, 45% with a college or graduate degree, mean age 51.9 years, sleep disturbance 11.0 years, range 30-72 years). Participants completed the Wide Range Achievement Test[55] to estimate their reading levels, which ranged from third grade to post-high school. Twenty participants of different races, both sexes, and a range of reading levels reviewed each item. Item stems and response options were reviewed for clarity, meaning, and vocabulary. Of the 75 items reviewed, 10 items (13%) were subsequently rewritten to clarify item stems or response choices based on participant feedback. The mean score from Lexile analysis was 406.1 (260.8), indicating an average third-grade reading level, with a maximum at the seventh-grade reading level.

Pilot Testing

Pilot testing of the item bank was conducted in a national sample of 300 participants (150 with self-reported sleep disorders and 150 control subjects; 51% women, 6% Hispanic, 13% minority) empanelled by YouGov Polimetrix, an internet polling firm (http://www.polimetrix.com/). Participants completed the preliminary item bank of 310 items via the Internet. Frequency distributions across each item's response categories were examined to identify items with floor or ceiling effects. "High threshold" items were identified when a majority of the control sample endorsed only the bottom 2 response-option categories. Pilot-item responses were also compared with the prior cognitive interview results from patients and subjected to a second round of expert item review. Data from pilot-testing were used to trim the item bank from 310 to 128 items for subsequent testing and evaluation.

Psychometric Testing

Psychometric testing was conducted using 128 PROMIS sleep-wake items, the PSQI and ESS, and patient-reported ratings of global health. Global health items included ratings of 5 primary PROMIS domains (2 items each for physical function, emotional distress, and satisfaction with social roles and discretionary activities, and 1 item each for fatigue and pain), as well as general ratings of perceived health (1 item each for general health and quality of life [56]).

Two samples were used for this psychometric testing: a second sample collected by YouGov Polimetrix and a clinical sample. Both samples completed a computerized questionnaire containing the studied items. The Polimetrix sample included 1993 individuals from the community (41% women, 11% Hispanic, 16% minority); of these, 1259 reported no sleep problems and 734 reported sleep problems in response to 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*? *Did the treatment help you*? Given the nature of the YouGov Polimetrix sampling frame, we were not able to verify the presence or absence of participants' self-reported clinical diagnoses. The clinical sample included 259 patients recruited from sleep medicine, general medicine, and psychiatric clinics at the University of Pittsburgh Medical Center (61% women, 2% Hispanic, 30% minority). All endorsed sleep and wake symptoms during a scripted telephone screening interview for common sleep disorders (sleep apnea, insomnia, restless legs syndrome). In aggregate, the YouGov Polimetrix and clinical sample (n = 2252) included 43.8% women and had a median age of 52 years, with 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 10% Hispanic or Latino. Education attainment was high school or less (13.6%), some college (38.6%), college degree (27.9%), and advanced degree (19.9%).

CTT Analyses

CTT analyses included descriptive statistics, internal consistency reliability (Cronbach α), convergent validity with PSQI and ESS, and factor analysis. We assumed no preconceptions regarding the most appropriate factor structure for the PROMIS sleep-wake item bank. Accordingly, the entire sample was randomly split into 2 subsamples, one for exploratory factor analysis (EFA) (n = 1144) and the other for subsequent confirmatory factor analysis (CFA) (n = 1108). Both EFA and CFA were conducted using Mplus 4.21 with Promax rotation.[57] Following the guidance of previous PROMIS data-analysis plans,[50] we evaluated indices such as the Tucker-Lewis Index (TLI > 0.95 for good fit), comparative fit index (CFI > 0.95 for good fit), root mean square error of approximation (RMSEA < 0.06 for good fit), and standardized root mean residuals (SRMR < 0.08 for good fit) for EFA models. Scree plots, eigenvalues, and factor loadings were also examined. A ratio in excess of 4 for the first 2 eigenvalues, significant factor loadings on the primary factor, and small residual correlations represented evidence in support of unidimensionality.[50] Following factor analysis, additional items with low factor loadings were dropped.

IRT Analysis, Model Selection, and Item Calibration

IRT refers to a class of psychometric techniques in which the probability of choosing each item-response category for each item is modeled as a function of a latent trait of interest (for further detail, refer to supplemental material). By convention, the latent trait is scaled along a dimension called theta (θ). IRT differs from CTT in 3 important ways. First, IRT models 1 or more parameters that describe each item, such as item difficulty (i.e., at what value of θ an individual has a 0.50 probability of choosing that item-response category) and discrimination (i.e., how well an item distinguishes among individuals along the θ scale).[58] In this study, the latent traits of interest were sleep disturbance and SRI. Thus, IRT provides psychometric information regarding each questionnaire item separately, as well as psychometric information for the overall test. Second, IRT provides not only item-parameter estimates invariant to forms of measurement and applicable across samples and populations, but also θ estimates for *individual participants*. In this way, an individual's responses can be used to precisely estimate his or her severity of sleep disturbance or SRI relative to the population. Third, IRT item-parameter estimates occur on the same θ scale as those of individuals completing the questionnaire, i.e., sleep disturbance or SRI items can be represented along the same severity spectrum as those of individual respondents.

The most commonly used IRT model for polytomous items (e.g., items with 5-point response scales) is the graded response model (GRM[59]). GRM has 1 slope parameter and *n*-1 threshold parameters for each item, where *n* is the number of response categories. The slope parameter measures item discrimination, i.e., how well the response categories differentiate among different θ levels; useful items have large slope parameters. Threshold parameters measure item difficulty, i.e., the location of the item on the θ scale; items with lower threshold parameters identify lower levels of θ, and items with higher threshold parameters identify higher levels of θ. Items were calibrated using MULTILOG 7[60] based on a general GRM, in which individual items are allowed to have different discrimination parameters.

Item Selection

To further refine the sleep disturbance and SRI item banks, additional criteria were considered from IRT analyses and descriptive statistics.

Item information function

An item's total information function is determined by discrimination parameter estimates and the range of threshold parameter estimates along θ. Items with discrimination parameter estimates less than 1.0 (i.e., low item information) were considered for exclusion.

Response distributions

Items with sparse cell distributions can be particularly problematic in IRT because of skewing. In principle, it is not possible to obtain reliable estimates of parameters for response categories with very few observations, i.e., strongly skewed distributions. Because many sleep disturbance and SRI items were right skewed (i.e., more precision was obtained at higher severity levels), items with sparse cells in the 2 response categories indicating greatest severity were considered for exclusion.

Construct validity

For concurrent validity, items demonstrating high correlation with the PSQI but low correlation with ESS were desired for the sleep-disturbance item bank, and the opposite pattern of correlations was desired for the SRI item bank. For discriminant validity, sleep disturbance and SRI items showing high Pearson correlations with the PROMIS network global health items (in particular, the fatigue item) were considered for exclusion.

Monotonicity/scalability

The monotonicity assumption specifies that the probability of selecting an item response category is a nondecreasing or *S*-shaped function of the underlying θ level of the construct being measured. In other words, the probability of selecting an item response indicating greater severity should increase as the overall level of sleep disturbance or SRI (estimated by θ) increases. Two nonparametric methods were used to evaluate the monotonicity assumption. First, the TestGraf program[61] was used to visualize empiric probability-curve estimates using a Gaussian kernel-smoothed model for nonparametric analysis. Second, Mokken Scale *H* coefficients for polytomous items (MSP[62]) were checked for values less than 0.30, indicating poor monotonicity.[63]

Local independence

Local independence assumes that the probability of providing a specific response to one item is independent of the probability of providing a specific response to any other item, after controlling for overall severity and item-parameter estimates. The existence of locally dependent item pairs may inappropriately overestimate or underestimate the probabilities for specific response patterns.[64] A computer program calculated local dependence indices for polytomous items (LDIP)[65] based on item-parameter estimations from MULTILOG. The Q3 statistic was used to evaluate local dependence. Although absolute Q3 values larger than 0.30 are often used to identify local dependence,[66] the nature of items banks, such as sleep disturbance and SRI, is expected to produce more locally dependent pairs and clusters[67] of items than are other types of item banks, such as cognitive tests. Therefore, a less restrictive Q3 of 0.50 was used to consider items for exclusion.

Content-expert review

Content experts (DJB, DEM, AG) reexamined items from the clinical perspective to eliminate items with questionable properties according to the 5 criteria described above. Conversely, items with important clinical implications were added back even if they failed to meet some of the previous 5 criteria.

Estimating Individual Scores

After the final sleep disturbance and SRI items were calibrated, each respondent's location on the corresponding sleep disturbance and SRI θ scales was estimated, i.e., a "score" for each individual was calculated.[64] Scoring under the CTT framework typically sums fixed values assigned to each response for each individual item. By contrast, scoring under the IRT framework is not based upon simple addition of item-category values, or even upon a fixed number of items. Rather, IRT-related scoring strategies mathematically estimate an individual's location on the θ scale by using that individual's pattern of item responses in conjunction with estimated item parameters. We used the maximum likelihood method provided by MULTILOG 7 for estimating individual scores on the sleep disturbance and SRI scales.

Preliminary Validity Evidence

To evaluate the face validity of the final sleep disturbance and SRI item banks, θ scores were compared between individuals who did and who did not report a previously diagnosed sleep disorder. We also compared θ scores between subjects who self-reported being treated for a sleep disorder and those who did not endorse treatment. (See Table S2 in supplemental material.)

RESULTS

Figure 1 summarizes the developmental steps and results of the PROMIS sleep-wake–function item banks. Although depicted as linear, the actual process was iterative. For instance, the conceptual framework was modified in response to item banking, focus group, and expert review steps. The development of the item banks is described in the Methods section; the remainder of this section will focus on psychometric testing of the 128 sleep-wake items administered to the YouGov Polimetrix and clinical samples.

Of the 128 items used for psychometric calibration, 19 were reverse scored, and 5 with hypothesized U-shaped responses (e.g., very short and very long sleep duration) were rescaled to yield unidirectional scales. Thirteen items were removed because their response scalings were not sufficiently unidirectional. Specifically, 8 items had responses that were clock times (e.g., S37: *What time did you usually go to bed*?), 3 could not be construed as directional (e.g., S23: *I napped*), and 2 had responses contingent upon other items (e.g., S24: *How long did your naps usually last*?). After these deletions, 115 items were used for subsequent analyses.

Initial testing addressed whether 1 underlying θ dimension might apply to the items. Internal consistency reliability of the 115 candidate items was high, as indicated by a Cronbach α of 0.96. However, this metric would be expected to be high for a measure with this many items. Item-total correlations (Table 2) were smaller than 0.40 for 39 (34%) of the 115 items, indicating that these items were not strongly related to a single underlying dimension.

Pearson correlations of the item-bank total score with the PSQI total score were 0.66 in the entire sample and 0.85 in the YouGov Polimetrix sample. Correlation with the ESS was lower (r = 0.25 for the entire sample, 0.36 for the YouGov Polimetrix sample). These results suggest convergent validity with sleep quality, as measured by the PSQI, and discriminant validity against the tendency to doze, as measured by the ESS, across the 115-item bank.

Initial EFA (EFA-1) of the 115 items yielded an RMSR of 0.10, indicating marginal fit for the 1-factor model. The scree plot of eigenvalues for this EFA revealed 1 dominant factor and an elbow after 4 factors (see Figure S1 in supplemental material). Five factors had eigenvalues greater than 2.8, and 75 items had factor loadings greater than 0.50 on at least 1 factor (see tables in supplemental material). These factors were labeled *Sleep Quality and Sleep Onset* (32 items, e.g., sleep quality, sleep restfulness, satisfaction with sleep, difficulty falling asleep), *Waking Symptoms* (24 items, e.g., had enough energy, sleep during the daytime, trouble staying awake, problems during the day because of poor sleep), *Sleep-Wake Transition* (7 items, e.g., felt alert when woke up, woke without an alarm), *Sleep-Onset Problems* (8 items, e.g., feeling tense

Table 1—Final Item Calibrations for the PROMIS Sleep Disturbance Item Bank[a]

Items	a	b1	b2	b3	b4
S20: I had a problem with my sleep.	2.80	-0.56	0.33	0.98	1.74
S42: It was easy for me to fall asleep.	2.09	-1.10	0.05	0.94	1.84
S44: I had difficulty falling asleep.	2.51	-0.46	0.31	0.98	1.72
S45: I lay in bed for hours waiting to fall asleep.	2.18	0.03	0.85	1.55	2.38
S50: I woke up too early and could not fall back asleep.	1.19	-0.98	0.33	1.76	3.30
S65: I felt physically tense at bedtime.	1.64	0.21	1.13	2.02	2.96
S67: I worried about not being able to fall asleep.	2.37	0.28	1.02	1.62	2.37
S68: I felt worried at bedtime.	1.77	0.22	1.21	1.95	2.73
S69: I had trouble stopping my thoughts at bedtime.	1.75	-0.57	0.41	1.04	1.81
S70: I felt sad at bedtime.	1.40	0.67	1.56	2.26	3.13
S71: I had trouble getting into a comfortable position to sleep.	1.52	-0.19	0.95	1.76	2.72
S72: I tried hard to get to sleep.	2.47	0.03	0.66	1.19	1.88
S78: Stress disturbed my sleep.	1.99	-0.02	0.89	1.61	2.22
S86: I tossed and turned at night.	1.85	-0.58	0.66	1.37	2.30
S87: I had trouble staying asleep.	2.19	-0.90	0.10	1.00	1.78
S90: I had trouble sleeping.	3.66	-0.61	0.16	0.96	1.62
S92: I woke up and had trouble falling back to sleep.	2.17	-0.55	0.36	1.31	2.24
S93: I was afraid I would not get back to sleep after waking up.	1.97	0.22	1.03	1.65	2.47
S105: My sleep was restful.	2.45	-1.20	-0.15	0.72	1.59
S106: My sleep was light.	1.51	-0.65	0.44	1.59	2.61
S107: My sleep was deep.	1.57	-1.52	-0.35	0.66	1.92
S108: My sleep was restless.	2.30	-0.29	0.69	1.45	2.33
S109: My sleep quality was…	3.39	-1.22	0.00	1.08	1.90
S110: I got enough sleep.	2.17	-1.56	-0.16	0.77	1.81
S115: I was satisfied with my sleep.	2.77	-1.25	-0.34	0.43	1.09
S116: My sleep was refreshing.	2.58	-1.35	-0.34	0.49	1.28
S125: I felt lousy when I woke up.	1.91	-0.14	0.68	1.32	2.07

[a] Column a displays the slope parameter (how well the item discriminates between respondents with low or high symptom levels). Columns b1- b4 display threshold values for individual responses (low threshold values indicate that the item is sensitive to low severity levels and high threshold values indicate that the item is sensitive to high severity levels).

and worried at bedtime, sleep disturbed by sadness), and *Sleep Disorder Symptoms* (4 items, e.g., restless legs, legs jerked or twitched). Items with factor loadings less than 0.50 included those representing extreme severity (e.g., use of medication to stay awake), time-based items, sleep-related behaviors and beliefs, other specific sleep disturbances (e.g., sleep apnea), and 2 items with hypothesized U-shaped responses (time to fall asleep, sleep length).

A second round of EFA (EFA-2) was conducted with 74 items (see Table S2 in supplemental material). These included the 75 items that loaded with a value of greater than 0.50 on 1 of the first 4 factors identified above, plus 3 additional items that were expected, on the basis of content-expert review, to load on 1 of those 4 factors (items S13: *I fell asleep while driving*, S16: *I took prescription medication to stay awake during the day (example: Ritalin)*, and S128: *How long did it take you to feel alert after waking up?*). On the other hand, the 4 items loading on *Sleep Disorder Symptoms* from EFA-1 were excluded from further analysis because 4 items were considered too few to constitute a viable scale. EFA-2 yielded an adequate RMSR value of 0.04 and 4 factors, which were labeled *Sleep-Onset Problems* (21 items, including items from Factors 1 and 4 in EFA-1), *Waking Symptoms* (26 items), *Sleep-Wake Transition* (9 items), and *Sleep Quality* (18 items, from Factor 1 in EFA-1).

Empirical review of the 4 factors from EFA-2 suggested conceptual similarities between new Factors 1 and 4 (Sleep-Onset Problems and Sleep Quality) and between new Factors 2 and 3 (Waking Symptoms and Sleep-Wake Transition). Therefore, a third EFA (EFA-3) was conducted with 74 items, combining Factors 1 and 4 and Factors 2 and 3 from EFA-2. In EFA-3, combined new Factor 1-4 had an SRMR of 0.09, and all 39 items had factor loadings greater than 0.50. Combined new Factor 2-3 had an RMSR of 0.09, and 33 of the 35 items had factor loadings greater than 0.50. Items S13 and S16 (fell asleep while driving, prescription medication to stay awake) did not have factor loadings in excess of 0.50 in this round of EFA (see Table S2 in supplemental material).

A single CFA was performed on the 2 final factors, which included 72 total items. The factor labeled *Sleep Disturbance* had an SRMR of 0.086, RMSEA of 0.140, TLI of 0.957, and CFI of 0.843. Although the indices are slightly outside the desired ranges, all 39 items in Sleep Disturbances had factor loadings greater than 0.50. The factor labeled *Sleep-Related Impairment* had an SRMR of 0.82, RMSEA of 0.157, TLI of 0.955, and CFI of 0.812. Again, although the indices are slightly outside the desired ranges, all but 2 of the 33 items in Sleep-Related Impairment had factor loadings in excess of 0.50, the exceptions being S118 (*I woke without an alarm clock*) and S122 (*I woke with an alarm clock*).

IRT Model Selection and Item Calibration

Unidimensional GRMs were fitted to the sleep disturbance and SRI item banks separately. A constrained GRM and a general GRM were each fitted to the data for model comparison. The χ^2 difference statistics between these models were 1684 and 1682 for the sleep disturbance and SRI item banks, respectively ($P < 0.001$ for each), indicating significantly better fit for the general GRM on each item bank. Consequently, general GRMs were used for item calibration on each item bank.

Item-parameter estimates were obtained using MULTILOG 7.03. Item-parameter estimates of the final items for each item bank are displayed in Tables 1 and 2. (Note that these are the items retained after the item selection process described below.) Column *a* presents slope parameters, which are an indicator of item information. Columns *b1* through *b4* represent threshold-parameter estimates for individual item-category responses. Lower threshold-parameter estimates (*b1* through *b4*) indicate items that detect lower levels of severity; higher threshold-parameter estimates indicate items that detect greater levels of severity.

Table 2—Final Item Calibrations for the PROMIS Sleep-related Impairment Item Bank[a]

Items	a	b1	b2	b3	b4
S4: I had enough energy.	1.83	-1.68	-0.11	1.17	2.19
S6: I was sleepy during the daytime.	2.24	-1.29	0.27	1.07	2.11
S7: I had trouble staying awake during the day.	2.20	-0.14	0.93	1.73	2.55
S10: I had a hard time getting things done because I was sleepy.	3.45	0.10	0.97	1.65	2.38
S11: I had a hard time concentrating because I was sleepy.	3.40	-0.09	0.88	1.58	2.28
S18: I felt tired.	2.67	-1.54	0.18	0.94	1.90
S19: I tried to sleep whenever I could.	1.43	-0.44	0.69	1.88	3.18
S25: I had problems during the day because of poor sleep.	3.76	-0.09	0.84	1.53	2.25
S27: I had a hard time concentrating because of poor sleep.	4.82	0.10	1.02	1.61	2.22
S29: My daytime activities were disturbed by poor sleep.	3.66	-0.05	0.74	1.65	2.47
S30: I felt irritable because of poor sleep.	2.92	-0.03	0.89	1.56	2.33
S33: I had a hard time controlling my emotions because of poor sleep.	2.60	0.36	1.26	1.99	2.68
S119: I felt alert when I woke up.	1.67	-1.58	-0.39	0.52	1.39
S120: When I woke up I felt ready to start the day.	1.87	-1.51	-0.48	0.39	1.19
S123: I had difficulty waking up.	1.18	-0.15	1.04	2.02	2.99
S124: I still felt sleepy when I woke up.	1.72	-1.27	0.12	0.80	1.66

[a]Column *a* displays the slope parameter (how well the item discriminates between respondents with low or high symptom levels). Columns *b1-b4* display threshold values for individual responses (low threshold values indicate that the item is sensitive to low severity levels and high threshold values indicate that the item is sensitive to high severity levels).

Item Selection

The 70 sleep disturbance and SRI items were further refined following a 6-step item-selection procedure. Items removed during this procedure are presented in the online supplement in Tables S3 and S4.

Item information function

Item information curves, which reflect overall item performance, were visually examined, and items with limited information were removed. Specifically, 4 items from sleep disturbance (S50: *I woke up too early and could not fall back asleep;* S73: *I was afraid of going to bed;* S74: *I was afraid of going to sleep;* S83: *My sleep was disturbed by sadness*) and 4 items from SRI (S5: *I fell asleep when I did not mean to;* S8: *I fell asleep in public places (example: church, movie, work);* S9: *I felt sleepy when driving;* S23: *I napped*) had discrimination-parameter estimates less than 1.0 and were removed.

Response distributions

Because the item distributions were often right skewed (i.e., more precision was obtained at higher severity levels), the top 2 response categories for each item were examined closely. One additional item from sleep disturbance (S66: *I felt jittery or nervous at bedtime*) and 6 additional items from SRI (S12: *I made mistakes because I was sleepy;* S28: *I had a hard time thinking clearly because of poor sleep;* S32: *I made mistakes because of poor sleep;* S33: *I had a hard time controlling my emotions because of poor sleep;* S34: *I avoided or cancelled activities with my friends because of poor sleep;* S35: *I felt clumsy because of poor sleep*) were removed because the 2 most-severe categories were infrequently endorsed (less than 6% in total).

Construct validity

Pearson correlations with PSQI, ESS, and PROMIS global health and fatigue scale items were reexamined. Items that showed higher correlations with the global health and/or fatigue items than with PSQI and ESS were removed, since other PROMIS scales are being developed for each of these domains. Three items from SRI (S4: *I had enough energy;* S17: *I was fatigued; S18: I felt tired*) were removed on this basis. No item from the sleep-disturbance item bank was removed at this step.

Monotonicity/scalability

Empiric item-response curves generated from TestGraf indicated that all items showed good monotonic item-response curves, and observed *H* statistic values were larger than 0.40 for all items. Both methods indicated that responses for all items were monotonic, and no items were removed at this step.

Local independence

Four items from the sleep-disturbance item bank (S80: *Worrying disturbed my sleep*; S112: *I had all the sleep I needed*; S114*: I was satisfied with the amount of sleep I got*; S115: *I was satisfied with my sleep*) and 2 additional items from SRI (S26*: I had trouble coping because of poor sleep*; S31: *I had a hard time getting things done because of poor* sleep) shared local dependence (exceeded a threshold value of > 0.50 on the Q3 statistic) with 5 or more other items. These items were therefore removed.

Content-expert review

Content experts (DJB, DEM, AG) again reviewed the items for content validity. Two items from sleep disturbance (S50: *I woke up too early and could not fall back asleep;* S115: *I was satisfied with my sleep*) and 3 items from SRI (S4: *I had enough energy;* S18: *I felt tired;* S33: *I had a hard time controlling my emotions because of poor sleep*) were added back because of their important clinical implications. Another 5 items from sleep disturbance (S43: *How often did you have difficulty falling asleep?* S81: *My sleep was disturbed by racing thoughts;* S89: *How long did it usually take you to fall back asleep after an awakening during the night?* S111: *I wished I got more sleep each night;* S117: *I felt refreshed when I woke up*) and 3 items from SRI (S121: *When I got out of bed I felt ready to start the day;* S126: *I had to force myself to get up in the morning;* S128: *How long did it take you to feel alert after waking up?*) were further removed because of conceptual redundancy with other items. The final sleep-disturbance item bank consisted of 27 items, and the final SRI item bank consisted of 16 items. The

Table 3—Known-groups validity of the PROMIS sleep-disturbance and wake-disturbance item banks

	No sleep disorder (N) *n* = 1342	Insomnia (I) *n* = 358	Sleep apnea (A) *n* = 504	Restless legs syndrome (R) *n* = 132	P Value for pairwise comparisons vs N
Sleep disturbance θ	-0.27 (0.97)	1.00 (0.76)	-0.06 (0.96)	0.73 (0.89)	< 0.001 (I, A, R)
Sleep-related impairment θ	-0.25 (0.91)	0.75 (0.85)	0.14 (0.91)	0.71 (0.91)	< 0.001 (I, A, R)

PROMIS refers to Patient-Reported Outcomes Information System.

final item IRT calibration values for the sleep disturbance and SRI item banks are displayed in Tables 1 and 2.

Estimating Individual Scores

Using item-parameter estimates of the final sleep disturbance and SRI item banks, sleep disturbance and SRI θ scores were estimated for each of the 2252 individuals in the calibration sample. In the θ scaling, the sample mean is given a value of 0 and the standard deviation a value of 1; higher θ values indicate greater severity. Sample sleep-disturbance θ scores ranged from -2.32 to 3.13, and SRI θ scores ranged from -2.27 to 3.29. The mean θ scores for sleep disturbance in the YouGov Polimetrix group with no sleep disorders, the YouGov Polimetrix group with self-reported sleep disorders, and the clinical group were -0.34, 0.24, and 0.77, respectively. The mean θ scores for SRI in the same respective groups were -0.33, 0.25, and 0.88.

Preliminary Validity Evidence

To evaluate the construct validity of the final sleep-disturbance and SRI item banks, θ scores were compared between self-reported sleep disorder and no sleep disorder groups. Subjects reporting each sleep disorder had higher θ values for both sleep disturbance and SRI, compared with those with no sleep disorder (Table 3). These findings suggest that the sleep disturbance and SRI item banks do, in fact, differ in expected ways among known groups, supporting their construct validity. We also compared subjects with self-reported treatment versus untreated sleep disorders. As would be expected in a clinical setting, subjects with untreated sleep disorders had significantly higher mean θ scores ($P < 0.001$) for both sleep disturbance and SRI, compared with those who had received treatment (Untreated: sleep disturbance $\theta = 0.72$, SRI $\theta = 0.61$; Treated: sleep disturbance $\theta = 0.19$, SRI $\theta = 0.27$). This broadly suggests that the sleep disturbance and SRI item banks are able to detect treatment response.

We also examined convergent validity between θ scores for the sleep disturbance and SRI item banks and PSQI and ESS. The product-moment correlation between sleep disturbance and PSQI (hypothesized to measure similar attributes) was 0.85, and, between SRI and the ESS (hypothesized to measure related but slightly different constructs), 0.45. This expected pattern of results supports the validity of sleep disturbance and SRI item banks. Contrary to expectations, the correlation between SRI θ values and the PSQI ($r = 0.70$) was larger than the SRI-ESS correlation. However the SRI-ESS correlation was larger than that of the sleep disturbance-ESS correlation ($r = 0.25$), again supporting the validity of sleep disturbance and SRI item banks.

DISCUSSION

The PROMIS sleep disturbance and SRI item banks were developed through a systematic process of literature reviews, content-expert review, qualitative research, pilot testing, and psychometric testing in more than 2000 individuals. This process narrowed an initial list of 310 items to 43 and 17 potential content categories to 2, representing overall sleep disturbances, quality, and satisfaction (sleep disturbance item bank) and daytime impairments related to sleep or sleep problems (SRI item bank). CTT assessments including internal consistency reliability, convergent validity, EFA, and CFA provided support for the 2 preliminary item banks. The 2 final item banks demonstrated unidimensionality and local independence, important determinants of validity using IRT and, therefore, adequately represent the sleep disturbance and SRI domains from a psychometric perspective. The final items also adequately represent general sleep disturbances and SRI from a clinical perspective, i.e., they have good face validity and construct validity. This conclusion is further supported by significant differences between individuals with and without self-reported sleep disorders and between those with treated and untreated sleep disorders. Taken together, these findings support the reliability and validity of the PROMIS sleep disturbance and SRI item banks.

The initial identification of 17 categories, 53 subcategories, and 310 individual sleep-wake items was intentionally broad. This breadth provided important conceptual lessons regarding the sleep-wake functional domain within the context of the PROMIS initiative. Some sleep-wake categories, such as the timing of sleep, can be readily assessed by self-report, but no sleep times are inherently "better" or "worse" than others. Likewise, both short and long sleep durations are risk factors for adverse health outcomes,[7,68] suggesting that optimal duration of sleep is not a monotonic function of the sort that is amenable to IRT analysis. Furthermore, symptoms of specific sleep disturbances such as snoring or sleep apnea may not be accurately assessed by the individual. It is possible that, in attempting to assess sleep-wake function comprehensively, the tested item banks simply did not include enough items regarding specific sleep disturbances, or enough respondents with specific sleep-wake problems, to emerge as distinct factors in factor analysis. It may also be surprising to many clinicians that constructs such as sleepiness, fatigue, and SRI were grouped into a single final factor. Although we initially included these as separate conceptual domains, empiric analyses did not support their distinction in the final item banks. These findings challenge our clinical understanding of symptom structures and invite further study.

The PROMIS sleep disturbance and SRI item banks join a growing number of patient-reported outcomes related to sleep and wakefulness. For instance, a recent review identified 22 instruments that measure sleep dysfunction in adults and the relationship of sleep dysfunction with health-related quality of life,[20] and our own review[48] identified more than 100 instruments broadly related to sleep, wake, and circadian function. It is also important, therefore, to place the PROMIS sleep disturbance and SRI item banks within this broader scientific and measurement context. The final PROMIS sleep disturbance and SRI item banks do not measure temporal aspects or quantitative aspects of sleep, as do sleep diares,[22,23,69] nor do they measure symptoms of specific sleep disorders, as do other instruments.[24-35] Rather, the PROMIS sleep disturbance and SRI item banks quantify unidimensional latent constructs related to qualitative aspects of sleep and daytime impairment related to sleep. To use an analogy, the PROMIS item banks should be considered as a sleep "thermometer" that assess the degree of general disturbance irrespective of possible underlying causes, in keeping with the PROMIS Initiative's goals of broad-based health-status measurement. The final content of the sleep disturbance item bank includes many insomnia-like items, and the SRI item bank includes items related to SRI, including sleepiness, fatigue, and cognitive difficulties. These content areas reflect the major dimensions of sleep-wake function identified psychometrically in a large sample of respondents, rather than any intent to develop insomnia or daytime impairment instruments.

The notion of 2 broad factors that describe sleep disturbances and SRI is intuitively appealing and is supported by other qualitative and psychometric evidence. The categorical structure of sleep-wake characteristics and problems derived from our focus groups included 2 broad categories corresponding to sleep quality or disturbances and wake quality or disturbances. As previously noted, the PSQI and ESS, which assess composite sleep quality and daytime tendency to doze, are currently the 2 most widely cited sleep scales. Recent empiric evidence confirms that the PSQI and ESS are essentially uncorrelated with each other and relate in different ways to other self-report measures of mood and stress.[70] Factor analysis of the PSQI also clearly distinguishes sleep and waking symptoms.[42] Other self-report measures, ranging from the Sleep Wake Activity Inventory[71] to more recently developed scales,[72] have also identified sleep symptoms and sleep-related waking symptoms as the dominant factors. Thus, the convergence of various types of evidence suggests that broad categories of sleep disturbance and SRI are both clinically salient and psychometrically robust.

Although the sleep disturbance and SRI item banks are similar in intent to the PSQI and ESS, they are very different in form. The final sleep disturbance and SRI item banks correlated in expected ways with these 2 commonly used scales, but the SRI-ESS correlation was smaller in magnitude than the sleep disturbance-PSQI correlation. This may be explained by the item-selection process for the SRI item bank. Items used clinically to assess daytime sleepiness (e.g., S5: *I fell asleep when I did not mean to*; S8: *I fell asleep in public places*; S9: *I felt sleepy when driving*; S23: *I napped*) were excluded from the final SRI item bank because the most extreme responses were endorsed rarely or they provided little overall information relative to the entire SRI item bank. The content assessed by these excluded items resembles ESS content. Indeed, the correlation coefficient between θ scores on these 4 excluded items and the ESS was 0.65, versus the SRI-ESS correlation coefficient of 0.45. These 4 items may be appropriate for further testing in clinical samples of patients who have more severe sleep disorders. Analyses of convergent validity with other commonly used scales such as the Insomnia Severity Index[31] will be useful in future studies.

The methodology used to derive PROMIS sleep disturbance and SRI item banks differed in several respects from the methodology of previously published instruments. First was the rigorous and iterative process of collecting, sorting, and standardizing items derived from the larger PROMIS initiative. In particular, the inclusion of patient input is being increasingly recognized as a critical component of scale development for patient-reported outcomes of all types, including sleep.[73-75] Second, the samples for pilot testing and initial psychometric analysis of the scales was larger than typically reported for sleep-wake scale development (e.g., [39-41,76]). These larger samples permitted us to conduct EFA and CFA with adequate sample sizes. Finally, the use of IRT methods to characterize the test performance of individual items lends additional strength to our methods. Scales developed with IRT have several desirable attributes, including the ability to customize item selection to specific applications. Because IRT analyzes the measurement properties of each item, precise estimates of the severity of each individual's sleep disturbance or SRI can be obtained by selecting a smaller number of items from the sleep disturbance and SRI item banks. For instance, it would be possible to select items for an epidemiologic study with a normal population that measure lower levels of severity by choosing items with low discrimination parameters. A clinical trial of a new medication for insomnia may employ high discrimination items to measure change over time. An additional benefit of IRT is the possibility of computerized adaptive testing, which uses individual item measurement properties to develop a progressively more precise estimate of an individual's sleep disturbance or SRI severity, which allows the administration of fewer overall items, typically 5 to 8. Ultimately, the choice of how many items to administer represents a balance between efficiency and measurement precision. Administering 1 or 2 items offers maximal efficiency, administering the entire item bank offers maximal precision, and administering 5 to 8 items often offers a reasonable balance between the 2. Although it seems counterintuitive at first, the calibration of each item using IRT also means that administering 2 different subsets of items with similar performance characteristics will produce very similar estimates of an individual's sleep disturbance or SRI severity (θ). A common example is the administration of alternate versions of aptitude tests such as board examinations (where θ is referred to as "ability" rather than severity).

As general indicators, the sleep disturbance and SRI item banks may be useful in a variety of applications, such as clinical trials, epidemiologic studies, or routine patient care. Their utility is enhanced by their ability to measure a wide range of sleep disturbance or SRI severity. Our preliminary data suggest that the item banks can readily discriminate patient and control groups, but whether the item banks further differentiate among specific sleep disorders, such as chronic insomnia versus obstructive sleep apnea, remains to be determined. Our data also suggest that the sleep disturbance and SRI item banks may well be sensitive to treatment, but this also requires further validation. The general nature of the sleep disturbance and SRI scales may be best suited to population studies or studies in medical or psychiatric conditions. In sleep research, they may provide useful metrics to allow comparisons across different samples or different sleep disorders and will likely be most useful when combined with disease-specific measures. The sleep disturbance and SRI item banks may also be useful for epidemiologic studies, in that fixed short forms or computerized adaptive testing administration would provide high measurement precision with minimal respondent burden. IRT calibration makes it possible and feasible to select items that best capture the intended severity range of the sample being assessed. Another potential application of the sleep disturbance and SRI item banks is to administer them concurrently with PROMIS measures of emotional distress, fatigue, pain, physical function, and other health-related constructs to form a broad health profile.

Scoring of IRT-calibrated item banks differs from scoring conventional self-report scales derived by CTT, which typically sum the fixed values corresponding to each response for each question. Scoring of CTT-derived scales requires administration of the entire scale and may result in ambiguity. For instance, a score of 14 on the ESS can result from many different combinations of responses to the 8 individual items of the scale, not all of which are equivalent in severity. The benefit of IRT-derived scoring is that it provides a precise estimate for each individual based on his or her specific pattern of item responses, and it permits the use of efficient administration methods such as computerized adaptive testing. IRT scoring uses computer programs such as MULTILOG to estimate an individual's θ value based on his or her response patterns in conjunction with item-parameter values. For this reason, however, IRT-based item banks such as sleep disturbance and SRI are most efficiently administered via computer interface with direct data entry and scoring, which may prove difficult in some research and clinical settings. A fully functional online data-collection and scoring program for the sleep disturbance and SRI item banks is available at PROMIS Assessment Center http://www.assessmentcenter.net/ac1/. Moreover, static sleep-disturbance and SRI short forms can be administered in simpler formats, including pencil and paper. In this case, the individual items for a respondent can be summed, and the corresponding θ scores or T-scores can be estimated from a nonlinear transformation contained in a conversion table. Sleep disturbance and SRI short forms and computerized adaptive testing simulations will be described in a subsequent manuscript.

We acknowledge several limitations of the current work. First, the aims of our study led us to develop item banks that would be useful for assessing sleep and wake function in multiple contexts and health conditions. Conversely, our item banks may prove less useful for assessing the symptoms of any specific sleep-wake disorder. Second, these item banks do not include items assessing sleep quantities and clock times. The place of such information in sleep-wake measurements remains unsettled. Third, our initial studies did not include some common components of validation and psychometric analysis, such as test-retest reliability, sensitivity and specificity for identifying individuals with sleep-wake disorders, or responsiveness to change. This was related to our goal of developing dimensional measures of sleep-wake function, rather than disease- or treatment-specific measures, but such analyses will be valuable components of further validation studies. Fourth, the majority of our test sample consisted of an Internet panel and was characterized in less detail than many samples used in validation studies. However, this limitation was offset by rapid data acquisition, and demonstration of the feasibility of Internet data collection, and the practicalities of studying a large initial calibration sample. Finally, the present calibration will be a springboard for further work to determine whether age, sex, or other respondent characteristics lead to different item functioning.[77]

In conclusion, the development and calibration of the PROMIS sleep disturbance and SRI item banks using CTT and IRT methods supports their validity. These item banks will permit researchers and clinicians to assess and integrate qualitative aspects of sleep and SRI with other PROMIS measures, in a variety of clinical and research samples and settings, and across a variety of health conditions.

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, PhD, 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 their colleagues at the Pittsburgh PROMIS site (Jill Cyranowski, PhD, Mary Amanda Dew, PhD, Frank Ghinassi, PhD, Mary Lou Klem, PhD, David Kupfer, MD), and the national PROMIS network (particularly David Cella, PhD, and Nan Rothrock, PhD). Supported by AR052155.

DISCLOSURE STATEMENT

This was not in industry supported study. Dr. Buysse has consulted for Actelion, Arena, Cephalon, Eli Lilly, GlaxoSmithKline, Merck, Neurocrine, Neurogen, Pfizer, Respironics, Sanofi-Aventis, Sepracor, Servier, Somnus Therapeutics, Stress Eraser, Takeda, and Transcept Pharmaceuticals.

REFERENCES

1. Ohayon MM. Epidemiology of insomnia: What we know and what we still need to learn. Sleep Med Rev 2002;6:97-111.
2. Ohayon MM. From wakefulness to excessive sleepiness: what we know and still need to know. Sleep Med Rev 2008;12:129-41.
3. Cappuccio FP, Stranges S, Kandala NB, et al. Gender-specific associations of short sleep duration with prevalent and incident hypertension: the Whitehall II Study. Hypertension 2007;50:693-700.

4. Cappuccio FP, Taggart FM, Kandala NB, et al. Meta-analysis of short sleep duration and obesity in children and adults. Sleep 2008;31:619-26.
5. Hall MH, Muldoon MF, Jennings JR, et al. Self-reported sleep duration is associated with the metabolic syndrome in midlife adults. Sleep 2008;31:635-43.
6. Cohen S, Doyle WJ, Alper CM, et al. Sleep habits and susceptibility to the common cold. Arch Intern Med 2009;169:62-7.
7. Gangwisch JE, Heymsfield SB, Boden-Albala B, et al. Sleep duration associated with mortality in elderly, but not middle-aged adults in a large US sample. Sleep 2008;31:1087-96.
8. Breslau N, Roth T, Rosenthal L, et al. Sleep disturbance and psychiatric disorders: A longitudinal epidemiological study of young adults. Biol Psychiatry 1996;39:411-8.
9. Riemann D, Voderholzer U. Primary insomnia: a risk factor to develop depression? J Affect Disord 2003;76:255-9.
10. Buysse DJ, Angst J, Gamma A, et al. Prevalence, course, and comorbidity of insomnia and depression in young adults. Sleep 2008;31:473-80.
11. Van Dongen HP, Maislin G, Mullington JM, et al. The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003;26:117-26.
12. Spiegel K, Knutson K, Leproult R, et al. Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes. J Appl Physiol 2005;99:2008-19.
13. Walker MP, Stickgold R. Sleep-dependent learning and memory consolidation. Neuron 2004;44:121-33.
14. Yoo SS, Gujar N, Hu P, et al. The human emotional brain without sleep--a prefrontal amygdala disconnect. Curr Biol 2007;17:R877-8.
15. Irwin MR, Wang M, Campomayor CO, et al. Sleep deprivation and activation of morning levels of cellular and genomic markers of inflammation. Arch Intern Med 2006;166:1756-62.
16. Nofzinger EA. Neuroimaging and sleep medicine. Sleep Med Rev 2005;9:157-72.
17. Dang-Vu TT, Desseilles M, Petit D, et al. Neuroimaging in sleep medicine. Sleep Med 2007;8:349-72.
18. Ancoli-Israel S. Actigraphy. In: Kryger MH, Roth T, Dement WC, eds. Principles and Practice of Sleep Medicine, 4th ed. Philadelphia, PA: Saunders; 2005:1459-67.
19. Morgenthaler T, Alessi C, Friedman L, et al. Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007. Sleep 2007;30:519-29.
20. Devine EB, Hakim Z, Green J. A systematic review of patient-reported outcome instruments measuring sleep dysfunction in adults. Pharmacoeconomics 2005;23:889-912.
21. Lichstein KL, Durrence HH, Riedel B, et al. A review of epidemiological studies of insomnia and sleep. In: Lichstein KL, Durrence HH, Riedel BW, Taylor DF, Bush AJ. Epidemiology of Sleep: Age, Gender and Ethnicity. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2004:9-41.
22. Monk TH, Reynolds CF, Kupfer DJ, et al. The Pittsburgh Sleep Diary. J Sleep Res 1994;3:111-20.
23. Monk TH, Buysse DJ. The Sleep Timing Questionnaire (STQ). Sleep 2003;26:A384-5.
24. Wunderlich GR, Evans KR, Sills T, et al. An item response analysis of the international restless legs syndrome study group rating scale for restless legs syndrome. Sleep Med 2005;6:131-9.
25. Allen RP, Kushida CA, Atkinson MJ. Factor analysis of the International Restless Legs Syndrome Study Group's scale for restless legs severity. Sleep Med 2003;4:133-5.
26. Maislin G, Pack AI, Kribbs NB, et al. A survey screen for prediction of apnea. Sleep 1995;18:158-66.
27. Netzer NC, Stoohs RA, Netzer CM, et al. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. Ann Intern Med 1999;131:485-91.
28. Douglass AB, Bornstein R, Nino-Murcia G, et al. The Sleep Disorders Questionnaire. I: creation and multivariate structure of SDQ. Sleep 1994;17:160-7.
29. Levine DW, Kaplan RM, Kripke DF, Bowen DJ, Naughton MJ, Shumaker SA. Factor structure and measurement invariance of the Women's Health Initiative Insomnia Rating Scale. Psychol Assess 2003;15:123-36.
30. Hoelscher TJ, Ware JC, Bond T. Initial validation of the insomnia impact scale. Sleep Res 1993;22:149.
31. Bastien CH, Vallieres A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001;2:297-307.
32. Soldatos CR, Dikeos DG, Paparrigopoulos TJ. Athens Insomnia Scale: validation of an instrument based on ICD-10 criteria. J Psychosom Res 2000;48:555-60.
33. Spielman AJ, Saskin P, Thorpy MJ. Treatment of chronic insomnia by restriction of time in bed. Sleep 1987;10:45-56.
34. Hays RD, Martin SA, Sesti AM, et al. Psychometric properties of the Medical Outcomes Study Sleep measure. Sleep Med 2005;6:41-4.
35. Roth T, Zammit G, Kushida C, et al. A new questionnaire to detect sleep disorders. Sleep Med 2002;3:99-108.
36. Yi H, Shin K, Shin C. Development of the sleep quality scale. J Sleep Res 2006;15:309-16.
37. Zhang L, Zhao ZX. Objective and subjective measures for sleep disorders. Neurosci Bull 2007;23:236-40.
38. Wyatt JK, Bootzin RR, Allen JJ, et al. Mesograde amnesia during the sleep onset transition: replication and electrophysiological correlates. Sleep 1997;20:512-22.
39. Buysse DJ, Reynolds CF, Monk TH, et al. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989;28:193-213.
40. Johns MW. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. Sleep 1991;14:540-5.
41. Johns MW. Reliability and factor analysis of the Epworth Sleepiness Scale. Sleep 1992;15:376-81.
42. Cole JC, Motivala SJ, Buysse DJ, et al. Validation of a 3-factor scoring model for the Pittsburgh Sleep Quality Index in older adults. Sleep 2006;29:112-6.
43. Chervin RD, Aldrich MS, Pickett R, et al. Comparison of the results of the Epworth Sleepiness Scale and the Multiple Sleep Latency Test. J Psychosom Res 1997;42:145-55.
44. Olson LG, Cole MF, Ambrogetti A. Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms. J Sleep Res 1998;7:248-53.
45. Miletin MS, Hanly PJ. Measurement properties of the Epworth Sleepiness Scale. Sleep Med 2003;4:195-9.
46. Martin JL, Ancoli-Israel S. Assessment and diagnosis of insomnia in non-pharmacological intervention studies. Sleep Med Rev 2002;6:379-406.
47. Morin CM. Measuring outcomes in randomized clinical trials of insomnia treatments. Sleep Med Rev 2003;7:263-79.
48. Moul DE, Hall M, Pilkonis PA, et al. Self-report measures of insomnia adults: Rationales, choices, and needs. Sleep Med Rev 2004;8:177-98.
49. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1):S3-11.
50. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):S22-31.
51. DeWalt DA, Rothrock N, Yount S, et al. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45(5 Suppl 1):S12-21.
52. Food and Drug Administration. Guidance for industry: computerized systems used in clinical trials. http://www.fda.gov/cder/guidance/index.htm; 2007.
53. World Health Organization. Basic documents. 46th ed. Geneva: World Health Organization; 2007.
54. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH). Guidance for Industry: Patient-reported outcome measures: Use in medical product development to support labeling claims, 2006.
55. Wilkinson GS, Robertson GR. Wide-Range Achievement Test. 4 ed. Lutz, FL: Psychological Assessment Resources, Inc.; 2006.
56. Hays RD, Bjorner J, Revicki DA, et al. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009;18:873-80.
57. Muthen LK, Muthen BO. Mplus User's Guide, 4th ed. Los Angeles, CA: Muthen & Muthen; 2007.
58. Lord FM. Applications of Item Response Theory to Practical Testing Problems. New York, NY: Erlbaum Associates; 1980.

59. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph 1969;17.
60. Thissen D. MULTILOG 7: Multiple Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago, IL: Scientific Software; 2003.
61. Ramsay JO. TestGraf: a program for the graphical analysis of multiple choice test and questionnaire data. Montreal, QE, Canada: McGill University; 1995.
62. Molenaar IW, Sijtsma K. Users Manual MSP5 for Windows: a Program for Mokken Scale Analysis for Polytomous Items. Groningen, the Netherlands: iec ProGAMMA; 2000.
63. Mokken RJ. A Theory and Procedure of Scale Analysis With Applications in Political Research. New York-Berlin: de Gruyter (Mouton); 1971.
64. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
65. Kim S-H, Cohen AS, Lin Y-H. LDIP: a computer program for local dependence indices for polytomous items. Applied Psychological Measurement 2006;30:509-10.
66. Yen W. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement 1984;8:125-45.
67. Ip EH-S. Testing for local dependency in dichotomous and polytomous item response models. Psychometrika 2001;66:109-32.
68. Hall MH, Muldoon MF, Jennings JR, et al. Self-reported sleep duration is associated with the metabolic syndrome in midlife adults. Sleep 2008;31:635-43.
69. Keklund G, Akerstedt T. Objective components of individual differences in subjective sleep quality. J Sleep Res 1997;6:217-20.
70. Buysse DJ, Hall M, Strollo PJ, et al. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and clinical/polysomnographic measures in a community sample. J Clin Sleep Med 2008;4:563-71.
71. Rosenthal L, Roehrs TA, Roth T. The Sleep-Wake Activity Inventory: a self-report measure of daytime sleepiness. Biol Psychiatry 1993;34:810-20.
72. Koffel E, Watson D. The two-factor structure of sleep complaints and its relation to depression and anxiety. J Abnorm Psychol 2009;118:183-94.
73. Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development in support for labeling claims. http://www. fda.gov/cder/guidance/index.htm, 2006.
74. Carey TJ, Moul DE, Pilkonis P, et al. Focusing on the experience of insomnia. Behav Sleep Med 2005;3:73-86.
75. Harvey AG, Stinson K, Whitaker KL, et al. The subjective meaning of sleep quality: A comparison of individuals with and without insomnia. Sleep 2008;31:383-93.
76. Weaver TE, Laizner AM, Evans LK, et al. An instrument to measure functional status outcomes for disorders of excessive sleepiness. Sleep 1997;20:835-43.
77. Pine SM. Applications of item characteristic curve theory to the problem of test bias. In: Weiss DJ, ed. Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association. Minneapolis, MN, 1977.