SUPPLEMENTAL MATERIAL

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

THIS MATERIAL SUPPLEMENTS THE MAIN PAPER, "DEVELOPMENT AND VALIDATION OF PATIENT-REPORTED OUTCOME MEASURES FOR SLEEP DISTURBANCE and sleep-related impairments." The background for the Patient-Reported Outcomes Measures Information System (PROMIS) Roadmap initiative, and of the sleep-wake item bank in particular, are presented in the published paper. In this supplement, we present an expanded version of psychometric methods, details of methodology, and results. This supplement includes additional tables and figures and also refers to tables and figures published in the main paper.

The specific aims of the PROMIS sleep-wake project were (1) to develop an archive of self-report measures that assess sleep and sleep-related impairments (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 item response theory (IRT) models, and (4) to examine the validity of the new item banks against widely used existing measures (Pittsburgh Sleep Quality Index [PSQI], Epworth Sleepiness Scale [ESS]).

We present the detailed methods and results below in 2 sections. The first section describes instrument development and classical test theory (CTT) analyses. The second section describes calibration of items and item banks using item IRT analyses.

I. ITEM BANK DEVELOPMENT AND CTT ANALYSES

METHODS

Overview

Development of the sleep-wake function item banks was a single-site project conducted under the broader multisite PROMIS initiative. The methods for this process were similar to those used for the other PROMIS item banks[1-4] 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 item IRT. The entire process was iterative rather than linear.

Conceptual Framework

The PROMIS domain framework (see http://www.nihpromis.org, "Domain Framework") is drawn from the World Health Organization's tripartite framework of physical, mental, and social health.[5] Within this framework, PROMIS places sleep-wake function as a physical health measure that is also influenced by mental health. Investigators on the sleep-wake project generated a list of 17 potentially distinct conceptual categories within the broad spectrum of sleep-wake function (Table S1). These 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. In retrospect, some of the concepts included in the initial framework were not well suited to the development of item banks using CTT and IRT methods. For instance, quantitative sleep variables, although salient to the description of sleep-wake patterns, are not arrayed in linear fashion from "good" to "poor." For instance, both short and long sleep duration may be associated with increased health risks in a U-shaped function. In a different way, symptoms of some sleep disorders may be very specific and clinically relevant but occur too infrequently within the population or the individual to be used in a general instrument. Thus, some items may be diagnostically important but poorly scaled in self-reports item banks such as PROMIS.

Development of the Item Banks

Development of the item banks included 4 steps (see Figure 1, main manuscript). First, comprehensive literature searches were used to ensure broad content validity. An earlier literature review on instruments related to insomnia[6] laid the groundwork for the eventual identification of more than 100 sleep questionnaires and almost 3000 items. A team of Health Science librarians at the University of Pittsburgh then generated a second literature search to ensure a comprehensive review of the sleep-wake functioning domain. Literature searches in the Medline, Psych Info, and Health and Psychosocial Instruments databases were conducted using a list of 291 sleep-wake search terms developed by the research team. These sleep-wake search terms such as *sleepiness* were crossed with measurement terms such as *validity* and *psychometric* to narrow the search field to assessment and instrument literature. Five hundred thirty-five candidate citations were identified by the search, 126 were further examined by the content experts, and 71 sources were found to have adequate psychometric documentation. Citation history searches were run for each measure to determine its acceptance and use by the scientific community. Copies of the measures were gathered from both electronic and print sources, and the measures were then reviewed for item-bank suitability. When necessary, written permission was obtained for use of items. The comprehensive literature-search process and initial narrowing to instruments related specifically to sleep-wake function yielded a final pool of 82 questionnaires, 2529 sleep items, and a refined conceptual model of the sleep-wake function domain.

The second step was item banking. An Access database was created for the initial pool of 2529 sleep items. Items were coded into 17 sleep-wake content-conceptual categories (Table S1), and these categories were then subcategorized into 76 "bins." Each of the 2529 items within the item bank was assigned to 1 of the bins based on independent ratings by the 3 sleep-content experts (DJB, DEM, AG), with resolution of disagreements in coding by consensus conference. Qualitative Item Review based on the larger PROMIS Network protocol[3] allowed for a substantial reduction in the number of items by deleting those with redundant content. Items that were thought to be confusing, awkward, or related to multiple content areas were rewritten to be consistent with PROMIS Network standards of verb tense, time frame, and response set. Thus, 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 completing the qualitative item and expert item reviews, 310 items in 53 bins were retained for further testing (available upon request).

The third step was to conduct focus groups with patients and nonpatient participants. Five focus groups were conducted, including 2 sleep disorder groups, 2 sleep disorder and psychiatric patient groups, and 1 group of normal sleepers. Participants were recruited from sleep medicine centers, outpatient clinics, and advertisements. During these group sessions, the focus group facilitator 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. Focus groups provide essential patient input in the development of patient-reported outcomes.[7] Thirty-six volunteers participated (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). Formal qualitative analyses of the focus-group transcripts are currently being conducted in conjunction with the Qualitative Data Analysis Program (Interim Director: Donald Musa) at the University of Pittsburgh's University Center for Social and Urban Research using Atlas.ti software.[8] A preliminary qualitative review of participant comments revealed 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. From these themes in the focus groups, we generated 10 additional items for initial testing. Detailed results of focus group discussions are being submitted as a separate publication.

The fourth step consisted of cognitive interviews with patients to evaluate whether proposed items were readily understandable. Seventy-five "core" items were selected for cognitive interviewing with 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). Cognitive-interview participants completed the Wide Range Achievement Test[9] to estimate their reading levels. The mean Wide Range Achievement Test score was 48.1 (sleep disturbance = 7.2, range 31 [third grade] to 57 [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 according to participant feedback.

Initial Testing of the Item Bank

Initial testing of the item bank included pilot testing in a small sample to further refine items, followed by testing in a larger sample for psychometric analyses. The pilot study 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/). This sample completed the preliminary item bank of 310 items via the Internet.

Data from the 150 participants with a sleep disorder and 150 control participants were compared in 5 ways, with the goal of trimming the large item bank down to a more manageable size for subsequent testing and evaluation. 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. The pilot-item responses were also compared with the results of the prior cognitive interview from patients and subjected to a second round of expert item review.

In the final pilot-study step, Lexile analyses were completed to approximate the reading level of each item. The mean item Lexile score was 406.1, indicating a third-grade reading level (sleep disturbance = 260.8 or 1.5 grades, with a maximum item Lexile score of a seventh-grade reading level). Results from the pilot study provided empiric evidence for further narrowing the item bank from 310 to 128 items for subsequent analyses.

Subsequent psychometric testing was conducted using 128 items (representing 46 of the 75 initially-constructed domains), two currently-available sleep-wake measures (PSQI and ESS), and patient-reported ratings of global health (as opposed to specific elements of health or health conditions). The PROMIS 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 across domains (1 item each for general health and quality of life[10]).

Two samples were used for this round of 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); 1259 of these were a general sample of the adult population and 734 were self-identified as having 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*?" "*Did the treatment help you*?" The clinical sample included 259 patients at the University of Pittsburgh Medical Center (61% women, 2% Hispanic, 30% minority) who endorsed sleep and wake symptoms during a telephone screening interview. These individuals were re-

Table S1—Categories of sleep-wake function identified by expert consensus

Sleep quality
Sleep onset
Sleep duration
Sleep continuity
Sleep offset
Rhythms and timing
Sleep habits and behaviors
Causes of sleep disturbance
Sleep-related beliefs
Dreams and nightmares
Breathing-related sleep problems (apnea)
Movement disorders (periodic limb movements)
Parasomnias
Insomnia
Sleepiness and alertness
Fatigue and energy
Consequences of poor sleep (other)

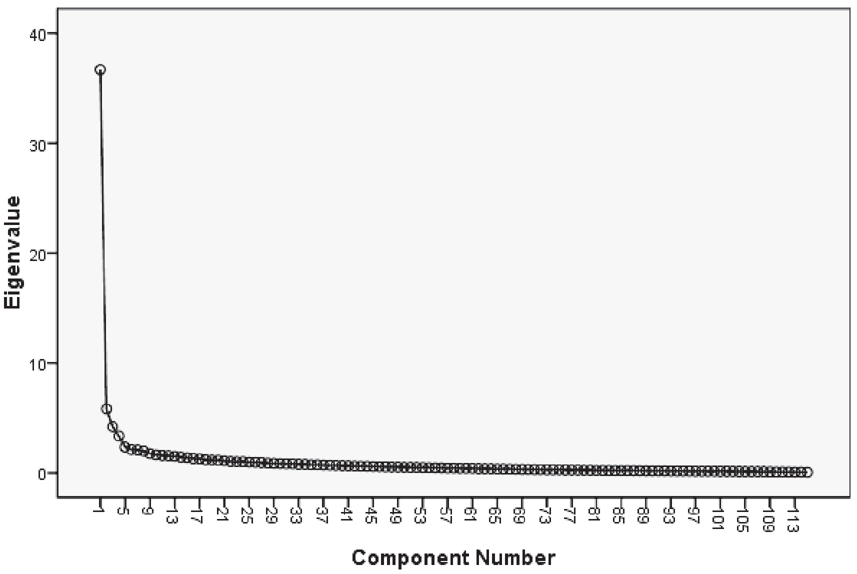

Figure S1—Scree plot for Exploratory Factor Analysis 1

cruited from sleep medicine, general medicine, and psychiatric clinics.

Psychometric analyses included descriptive statistics, internal consistency reliability (Cronbach α), and convergent validity with published measures (Pittsburgh Sleep Quality Index [PSQI] and Epworth Sleepiness Scale [ESS]), and tests of symptom-dimension unidimensionality. Unidimensionality is an essential assumption for IRT models, which were planned as the major psychometric analysis for PROMIS (see Section II below). Given the lack of measurement standardization in the sleep medicine field, we had no preconceptions regarding the most appropriate factor structure for the PROMIS sleep-wake item bank. Accordingly, we first used exploratory factor analysis (EFA) to explore the factor structure. The entire sample was randomly split into 2 subsamples, one for 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.[11] Following the guidance of previous PROMIS data-analysis plans,[2] 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.[2] Following factor analysis, additional items with low factor loadings were dropped. These reduced item banks were used for IRT analyses, which are addressed in Section II below.

RESULTS

Figure 1, in the main manuscript, presents a summary of the developmental steps and results of the PROMIS sleep-wake function item banks. Although the process is displayed as a linear one in the figure, the actual process was iterative in nature. For instance, the conceptual framework was modified in response to item banking, focus group, and expert review steps. The development of the conceptual framework and item banks is described in the Methods section above; the remainder of this section will focus on psychometric testing of the 128 sleep-wake function items administered to the YouGov Polimetrix and clinical samples. These 128 items are shown in Table S2.

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 (i.e., from *good* to *bad*), which is an assumption for both CTT and IRT methods. Thirteen items were removed from psychometric testing because their response scalings were not sufficiently unidirectional. Specifically, 8 items had responses that were actual 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. See Table S2.

Internal-consistency reliability of the 115 PROMIS items, measured by Cronbach α, was 0.96, indicating a high degree of internal consistency. However, Cronbach α is influenced by the total number of items and would be expected to be high for an item bank with this many items. Item-total correlations (Table S2) were smaller than 0.40 for 39 (34%) of the 115 items, indicating that a substantial proportion of items were not strongly related to a single underlying dimension.

Correlations of the item-bank total score with the PSQI total score, which was designed to measure nighttime sleep quality, was substantial (r = 0.66 for the entire sample, 0.85 in the YouGov Polimetrix sample), supporting convergent validity with the PSQI. On the other hand, correlation with the ESS, which focuses on daytime tendency to doze, was much lower (r = 0.25 for the entire sample, 0.36 for the YouGov Polimetrix sample).

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 (Figure S1) revealed 1 domi-

Table S2—Classic test theory item statistics for 128 initial items

Item #	Item Stem[a]	Item-total correlation[b] (*r* value)	Factor Loadings[c]				Reason for item removal[d]
			EFA 1	EFA 2	EFA 3	CFA	
S1	*I worked at irregular times.*	0.162					FLE
S2	*My "best time" of day was…*	N/A					NU
S3	*I drank alcohol (beer, wine, mixed drinks).*	-0.139					FLE
S4	I had enough energy. [reverse scored]	0.635	F2	F2	F2	F2	
S5	I fell asleep when I did not mean to.	0.371	F2	F2	F2	F2	
S6	I was sleepy during the daytime.	0.642	F2	F2	F2	F2	
S7	I had trouble staying awake during the day.	0.615	F2	F2	F2	F2	
S8	I fell asleep in public places (example: church, movie, work).	0.330	F2	F2	F2	F2	
S9	I felt sleepy when driving.	0.360	F2	F2	F2	F2	
S10	I had a hard time getting things done because I was sleepy.	0.679	F2	F2	F2	F2	
S11	I had a hard time concentrating because I was sleepy.	0.683	F2	F2	F2	F2	
S12	I made mistakes because I was sleepy.	0.639	F2	F2	F2	F2	
S13	*I fell asleep while driving.*	0.158		F2			FLE
S14	*I used caffeine to stay awake during the day.*	0.425					FLE
S15	*I drank coffee, tea, cola, or energy drinks to stay awake during the day.*	0.455					FLE
S16	*I took prescription medication to stay awake during the day (example: Ritalin).*	0.157		F2			FLE
S17	I was fatigued.	0.701	F2	F2	F2	F2	
S18	I felt tired.	0.715	F2	F2	F2	F2	
S19	I tried to sleep whenever I could.	0.530	F2	F2	F2	F2	
S20	I had a problem with my sleep.	0.803	F1	F1	F1	F1	
S21	*I had an urge to move my legs when I was sitting still or lying down.*	0.421	F5				FD
S22	*I had restless feelings in my legs in the evening or night.*	0.443	F5				FD
S23	I napped.	0.320	F2	F2	F2	F2	
S24	*How long did your naps usually last?*	N/A					NU
S25	I had problems during the day because of poor sleep.	0.791	F2	F2	F2	F2	
S26	I had trouble coping because of poor sleep.	0.754	F2	F2	F2	F2	
S27	I had a hard time concentrating because of poor sleep.	0.777	F2	F2	F2	F2	
S28	I had a hard time thinking clearly because of poor sleep.	0.763	F2	F2	F2	F2	
S29	My daytime activities were disturbed by poor sleep.	0.777	F2	F2	F2	F2	

EFA refers to exploratory factor analysis; CPAP, continuous positive airway pressure; BiPAP, bilevel positive airway pressure.

[a] Items in italics were removed during psychometric evaluation. See last column for reason for removal

[b] Item-total correlations were calculated based on the 115-item version.

[c] Factor loadings are shown only for items with factor loading values > 0.50. F indicates the specific factor on which the item loaded.

[d] Reason code: NU, Not unidirectional (item responses could not be defined on a unidirectional scale); TB, Clock time-based items; CI, Contingent items (item administration based on response to another item); FD, factor dropped for further round of analysis due to small number of items loaded on this factor; FLE, factor loading in exploratory factor analysis (item did not load > 0.50 on any factor in exploratory factor analysis); FLC, factor loading in confirmatory factor analysis (item did not load > 0.50 on either factor in confirmatory factor analysis).

Table S2 *(continued)*—Classic test theory item statistics for 128 initial items

Item #	Item Stem[a]	Item-total correlation[b] (*r* value)	Factor Loadings[c]				Reason for item removal[d]
			EFA 1	EFA 2	EFA 3	CFA	
S30	I felt irritable because of poor sleep.	0.775	F2	F2	F2	F2	
S31	I had a hard time getting things done because of poor sleep.	0.768	F2	F2	F2	F2	
S32	I made mistakes because of poor sleep.	0.709	F2	F2	F2	F2	
S33	I had a hard time controlling my emotions because of poor sleep.	0.702	F2	F2	F2	F2	
S34	I avoided or cancelled activities with my friends because of poor sleep.	0.657	F2	F2	F2	F2	
S35	I felt clumsy because of poor sleep.	0.670	F2	F2	F2	F2	
S36	*What time did you want to go to bed?*	N/A					TB
S37	*What time did you usually go to bed?*	N/A					TB
S38	*What time did you want to go to sleep?*	N/A					TB
S39	*What time did you usually try to sleep?*	N/A					TB
S40	I *went to bed about the same time every night.* [reverse scored]	0.333					FLE
S41	*How long did it usually take you to fall asleep?*	0.009					FLE
S42	It was easy for me to fall asleep. [reverse scored]	0.659	F1	F1	F1	F1	
S43	How often did you have difficulty falling asleep?	0.691	F1	F1	F1	F1	
S44	I had difficulty falling asleep.	0.713	F1	F1	F1	F1	
S45	I laid in bed for hours waiting to fall asleep.	0.661	F1	F1	F1	F1	
S46	*What time did you usually get out of bed to start your day?*	N/A					TB
S47	*What time did you want to wake up?*	N/A					TB
S48	*How long did you usually sleep?*	0.440					NU
S49	*What time did you usually wake up to start your day?*	N/A					TB
S50	I woke up too early and could not fall back asleep.	0.471	F1	F1	F1	F1	
S51	*What time did you want to get out of bed?*	N/A					TB
S52	*How long did you usually spend in bed, including time awake and time asleep?*	0.263					NU
S53	*I woke up about the same time every day.* [reverse scored]	0.419					FLE
S54	*My sleeping hours were different from night to night.*	0.569					FLE
S55	*My bed was comfortable.* [reverse scored]	0.337					FLE
S56	*I worked while I was in bed.*	0.112					FLE
S57	*I used my computer while I was in bed.*	0.087					FLE
S58	*Prescribed sleep medicine helped me sleep.*	0.286					FLE

EFA refers to exploratory factor analysis; CPAP, continuous positive airway pressure; BiPAP, bilevel positive airway pressure.

[a] Items in italics were removed during psychometric evaluation. See last column for reason for removal

[b] Item-total correlations were calculated based on the 115-item version.

[c] Factor loadings are shown only for items with factor loading values > 0.50. F indicates the specific factor on which the item loaded.

[d] Reason code: NU, Not unidirectional (item responses could not be defined on a unidirectional scale); TB, Clock time-based items; CI, Contingent items (item administration based on response to another item); FD, factor dropped for further round of analysis due to small number of items loaded on this factor; FLE, factor loading in exploratory factor analysis (item did not load > 0.50 on any factor in exploratory factor analysis); FLC, factor loading in confirmatory factor analysis (item did not load > 0.50 on either factor in confirmatory factor analysis).

Table S2 *(continued)*—Classic test theory item statistics for 128 initial items

Item #	Item Stem[a]	Item-total correlation[b] (*r* value)	Factor Loadings[c]				Reason for item removal[d]
			EFA 1	EFA 2	EFA 3	CFA	
S59	*I used alcohol to help me sleep.*	0.194					FLE
S60	*I used over-the-counter medicine to help me sleep (example: Tylenol PM, Nytol).*	0.249					FLE
S61	*Over-the-counter medicine helped me sleep.*	0.178					FLE
S62	*I used a CPAP or BiPAP machine during sleep.*	N/A					CI
S63	*I felt sleepy at bedtime.*	0.039					FLE
S64	*I felt tired at bedtime.* [reverse scored]	-0.072					FLE
S65	I felt physically tense at bedtime.	0.651	F4	F1	F1	F1	
S66	I felt jittery or nervous at bedtime.	0.600	F4	F1	F1	F1	
S67	I worried about not being able to fall asleep.	0.682	F1	F1	F1	F1	
S68	I felt worried at bedtime.	0.671	F4	F1	F1	F1	
S69	I had trouble stopping my thoughts at bedtime.	0.673	F4	F1	F1	F1	
S70	I felt sad at bedtime.	0.588	F4	F1	F1	F1	
S71	I had trouble getting into a comfortable position to sleep.	0.620	F1	F1	F1	F1	
S72	I tried hard to get to sleep.	0.701	F1	F1	F1	F1	
S73	I was afraid of going to bed.	0.416	F4	F1	F1	F1	
S74	I was afraid of going to sleep.	0.377	F4	F1	F1	F1	
S75	*Light disturbed my sleep.*	0.343					FLE
S76	*Noise disturbed my sleep.*	0.371					FLE
S77	*Caffeine disturbed my sleep.*	0.257					FLE
S78	Stress disturbed my sleep.	0.710	F1	F1	F1	F1	
S79	*Pain disturbed my sleep.*	0.562					FLE
S80	Worrying disturbed my sleep.	0.692	F1	F1	F1	F1	
S81	My sleep was disturbed by racing thoughts.	0.678	F1	F1	F1	F1	
S82	*My bed partner disturbed my sleep.*	0.226					FLE
S83	My sleep was disturbed by sadness.	0.552	F4	F1	F1	F1	
S84	*Medications disturbed my sleep.*	0.299					FLE
S85	*My sleep at night was affected by my menstrual cycle.*	N/A					CI
S86	I tossed and turned at night.	0.661	F4	F1	F1	F1	

EFA refers to exploratory factor analysis; CPAP, continuous positive airway pressure; BiPAP, bilevel positive airway pressure.

[a] Items in italics were removed during psychometric evaluation. See last column for reason for removal

[b] Item-total correlations were calculated based on the 115-item version.

[c] Factor loadings are shown only for items with factor loading values > 0.50. F indicates the specific factor on which the item loaded.

[d] Reason code: NU, Not unidirectional (item responses could not be defined on a unidirectional scale); TB, Clock time-based items; CI, Contingent items (item administration based on response to another item); FD, factor dropped for further round of analysis due to small number of items loaded on this factor; FLE, factor loading in exploratory factor analysis (item did not load > 0.50 on any factor in exploratory factor analysis); FLC, factor loading in confirmatory factor analysis (item did not load > 0.50 on either factor in confirmatory factor analysis).

Table S2 *(continued)*—Classic test theory item statistics for 128 initial items

Item #	Item Stem[a]	Item-total correlation[b] (*r* value)	Factor Loadings[c]				Reason for item removal[d]
			EFA 1	EFA 2	EFA 3	CFA	
S87	I had trouble staying asleep.	0.682	F1	F4	F1	F1	
S88	*I woke up to use the bathroom.*	0.238					FLE
S89	How long did it usually take you to fall back asleep after an awakening during the night?	0.565	F1	F4	F1	F1	
S90	I had trouble sleeping.	0.787	F1	F4	F1	F1	
S91	*My legs jerked or twitched repeatedly during sleep.*	0.427	F5				FD
S92	I woke up and had trouble falling back to sleep.	0.658	F1	F4	F1	F1	
S93	I was afraid I would not get back to sleep after waking up.	0.628	F1	F4	F1	F1	
S94	*I avoided sleep.*	0.408					FLE
S95	*I snored loudly.*	0.164					FLE
S96	*I stopped breathing during sleep.*	0.236					FLE
S97	*I walked during sleep.*	0.150					FLE
S98	*I screamed during sleep.*	0.289					FLE
S99	*Child-care disturbed my sleep.*	0.129					FLE
S100	*I kicked, punched, or swung my arms during sleep.*	0.302	F5				FD
S101	*I checked the clock when I was awake at night.*	0.470					FLE
S102	*How many dreams did you remember having each night?*	N/A					NU
S103	*I had trouble sleeping because of bad dreams.*	0.478					FLE
S104	*I could predict how I would sleep.*	-0.146					FLE
S105	My sleep was restful. [reverse scored]	0.728	F1	F4	F1	F1	
S106	My sleep was light.	0.537	F1	F4	F1	F1	
S107	My sleep was deep. [reverse scored]	0.528	F1	F4	F1	F1	
S108	My sleep was restless.	0.716	F1	F4	F1	F1	
S109	My sleep quality was… [reverse scored]	0.798	F1	F4	F1	F1	
S110	I got enough sleep. [reverse scored]	0.759	F1	F4	F1	F1	
S111	I wished I got more sleep each night.	0.661	F1	F4	F1	F1	
S112	I had all the sleep I needed. [reverse scored]	0.747	F1	F4	F1	F1	
S113	*I felt I needed 8 hours of sleep to function well during the day.*	0.008					FLE
S114	I was satisfied with the amount of sleep I got. [reverse scored]	0.759	F1	F4	F1	F1	

EFA refers to exploratory factor analysis; CPAP, continuous positive airway pressure; BiPAP, bilevel positive airway pressure.

[a] Items in italics were removed during psychometric evaluation. See last column for reason for removal

[b] Item-total correlations were calculated based on the 115-item version.

[c] Factor loadings are shown only for items with factor loading values > 0.50. F indicates the specific factor on which the item loaded.

[d] Reason code: NU, Not unidirectional (item responses could not be defined on a unidirectional scale); TB, Clock time-based items; CI, Contingent items (item administration based on response to another item); FD, factor dropped for further round of analysis due to small number of items loaded on this factor; FLE, factor loading in exploratory factor analysis (item did not load > 0.50 on any factor in exploratory factor analysis); FLC, factor loading in confirmatory factor analysis (item did not load > 0.50 on either factor in confirmatory factor analysis).

Table S2 *(continued)*—Classic test theory item statistics for 128 initial items

Item #	Item Stem[a]	Item-total correlation[b] (*r* value)	Factor Loadings[c]				Reason for item removal[d]
			EFA 1	EFA 2	EFA 3	CFA	
S115	I was satisfied with my sleep. [reverse scored]	0.795	F1	F4	F1	F1	
S116	My sleep was refreshing. [reverse scored]	0.783	F1	F4	F1	F1	
S117	I felt refreshed when I woke up. [reverse scored]	0.778	F1	F4	F1	F1	
S118	*I woke up without an alarm clock.* [reverse scored]	0.138	F3	F3	F2		FLC
S119	I felt alert when I woke up. [reverse scored]	0.677	F3	F3	F2	F2	
S120	When I woke up I felt ready to start the day. [reverse scored]	0.723	F1	F3	F2	F2	
S121	When I got out of bed I felt ready to start the day. [reverse scored]	0.724	F3	F3	F2	F2	
S122	*I woke up with an alarm clock.*	0.055	F3	F3	F2		FLC
S123	I had difficulty waking up.	0.513	F3	F3	F2	F2	
S124	I still felt sleepy when I woke up.	0.695	F3	F3	F2	F2	
S125	I felt lousy when I woke up.	0.776	F1	F4	F1	F1	
S126	I had to force myself to get up in the morning.	0.662	F3	F3	F2	F2	
S127	*Another person woke me up in the morning.*	0.250					FLE
S128	How long did it take you to feel alert after waking up?	0.641		F3	F2	F2	

EFA refers to exploratory factor analysis; CPAP, continuous positive airway pressure; BiPAP, bilevel positive airway pressure.

[a] Items in italics were removed during psychometric evaluation. See last column for reason for removal

[b] Item-total correlations were calculated based on the 115-item version.

[c] Factor loadings are shown only for items with factor loading values > 0.50. F indicates the specific factor on which the item loaded.

[d] Reason code: NU, Not unidirectional (item responses could not be defined on a unidirectional scale); TB, Clock time-based items; CI, Contingent items (item administration based on response to another item); FD, factor dropped for further round of analysis due to small number of items loaded on this factor; FLE, factor loading in exploratory factor analysis (item did not load > 0.50 on any factor in exploratory factor analysis); FLC, factor loading in confirmatory factor analysis (item did not load > 0.50 on either factor in confirmatory factor analysis).

nant factor and an elbow after 4 factors. EFA-1 identified 5 factors with eigenvalues greater than 2.8 and included 75 items with factor loadings of greater than 0.50 on at least 1 factor (Table S2). 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 and worried at bedtime, sleep disturbed by sadness), and *Sleep Disorder Symptoms* (4 items, e.g., restless legs, legs jerked or twitched). Forty items in EFA-1 did not have factor loadings greater than 0.50 on any factor. These items included those representing extreme severity (e.g., use of medication to stay awake), time-based items, sleep-related behaviors and beliefs, and other specific sleep disturbances (e.g., sleep apnea). In addition, the 2 items with hypothesized U-shaped responses (time to fall asleep, sleep length) did not load onto any of the 5 factors.

A second round of EFA (EFA-2) was conducted with 74 items. 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?*). The 4 items loading on Sleep Disorder Symptoms 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 the initial EFA), Waking Symptoms (26 items), Sleep-Wake Transition (9 items), and Sleep Quality (18 items, from Factor 1 in EFA-1). Results of the EFA-2 are also summarized in Table S2.

Empiric 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 RMSR of 0.09, and all 39 items had factor loadings greater than 0.50. Combined new

Factor 2-3 had an SRMR 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 greater than 0.50 in this round of EFA. These findings are also summarized in Table S2.

A single confirmatory factor analysis was performed on the 2 final factors, which included 72 total items. The factor labeled Sleep Disturbances 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.50. The factor labeled SRI 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 SRI had factor loadings in excess of 0.50, the exceptions being S118 (woke without an alarm clock) and S122 (woke with an alarm clock). The 70 items and 2 factors resulting from these analyses constituted the Sleep Disturbance and SRI item banks that were subsequently used for IRT analyses described below.

II. ITEM BANK AND INDIVIDUAL ITEM CALIBRATION USING IRT

METHODS

Description of IRT

In this section, we present calibration data resulting from IRT analyses and comparisons between the measurement precision obtained from the PROMIS sleep disturbance and SRI item banks versus conventional measures.

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 (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 point of overall severity [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 with or without symptoms).[12] In the current case, the latent traits of interest were sleep disturbance and SRI. Thus, unlike CTT, IRT provides psychometric information regarding each specific sleep or wake item on the questionnaire, as well as psychometric information for the overall test. Second, IRT provides not only invariant item-parameter estimates applicable to samples and populations, but also θ estimates for specific *individuals*. 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, as a result of these properties, IRT item-parameter estimates can be placed on the same θ scale as individuals completing the questionnaire. In other words, sleep or wake bank items can be represented along the same severity spectrum as that of individuals with sleep disturbance or SRI.

Using IRT, the relationship between responses to a specific questionnaire item and an individual's overall level of severity of sleep-wake function disturbance (θ) can be described by a number of different functions. 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., *S*-shaped function) called the *item characteristic function*. This function intuitively makes sense in that it specifies that individuals with higher θ scores (i.e., greater severity of sleep disturbance or SRI) have higher expected probabilities for endorsing the more severe response categories (e.g., *often* or *always*) than do individuals with lower θ scores.

Information contained in the item characteristic function can also be plotted as an item characteristic curve (ICC), which describes the probability of choosing each specific item-response category for individuals with different levels of severity (θ). An ICC can also be transformed into an item information curve, indicating the amount of discriminating psychometric information a single item contains at all points along the severity (θ) scale (see Figure S2a and S2b). All of the individual item information curves can also be combined to form a test information curve, which indicates the amount and accuracy of psychometric information the entire test contains at every point of θ. Test information is the sum of individual item information, so test information increases when test length increases. The item-information function depends on the slope of the item-response function and the conditional variance at each level on the θ scale. Therefore, the greater the slope and the smaller the variance at a particular θ level, the greater the information provided at that level. Test-information curves with a steep slope, high and broad peak, and low standard error discriminate well among respondents across the range of θ values under that peak. The psychometric information contained in the test-information curve is also called *measurement precision*. An important feature of IRT models is that the amount of information provided by a test may vary depending on the level of a respondent's severity of sleep disturbance or SRI (θ). These curves can be used to compare 2 or more scales on measurement precision through a procedure called IRT linking (for technical details on this procedure, refer to[13]). In this section, we compare the test-information curves for the PROMIS sleep disturbance and SRI item banks in relationship to the PSQI and ESS.

IRT models permit investigators to evaluate the performance of a single item or subsets of items, as well as the entire test. One practical application based on this feature, which classical test theory cannot provide, is the ability to construct short forms or tailored assessments using a subset of items selected to provide adequate precision across the entire range of θ and to maximize precision along clinically important segments of severity. Another attractive advantage of IRT is that individuals' θ estimates are independent of the specific items administered from a calibrated item bank. This feature allows IRT to be applied to computerized adaptive testing (CAT), a method that provides a unique sequence of items tailored to the individual's severity (θ). CAT avoids administering test items that add little information to an individual's assessment. For more information regarding the technical issues in IRT methodology, see Embretson and Reise[14] and Reeve et al.[2] for a description of the PROMIS analysis framework.

Although IRT has many advantages, it is based on several statistical assumptions. One such assumption is unidimensionality. The underlying θ manifested by a set of items is assumed to explain individual test performance (e.g., different levels of

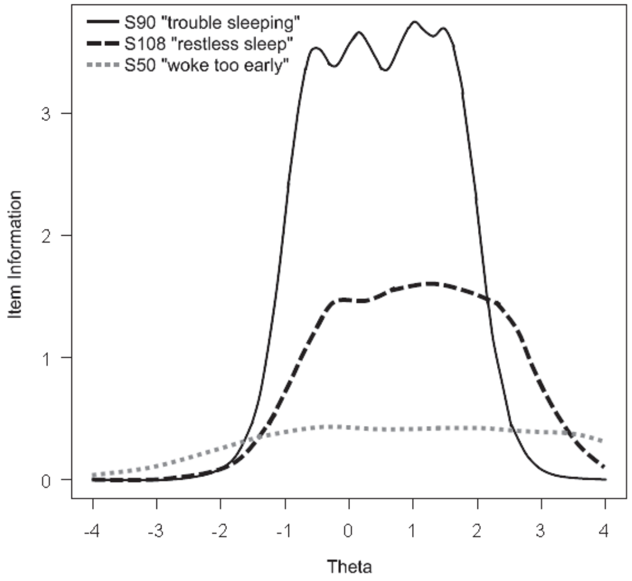

Figure S2a—Selected item-information curves for items on the sleep-disturbance item bank (S90: *I had trouble sleeping*; S108: *My sleep was restless*; S50: *I woke up too early and could not fall back asleep*). Curves with higher information along the θ scale, indicated by the height of the curve, have better measurement precision. The location of the peak reflects the discrimination parameter, a measure of which level of severity is best assessed by the item. S90 is an example of a high-information item centered in the middle of the sleep-disturbance severity scale. S108 provides less information and assesses a higher range of severity. S50 provides little information at any point of the severity scale and was dropped from the final calibration. See Table 1 in main manuscript for item-parameter estimates.

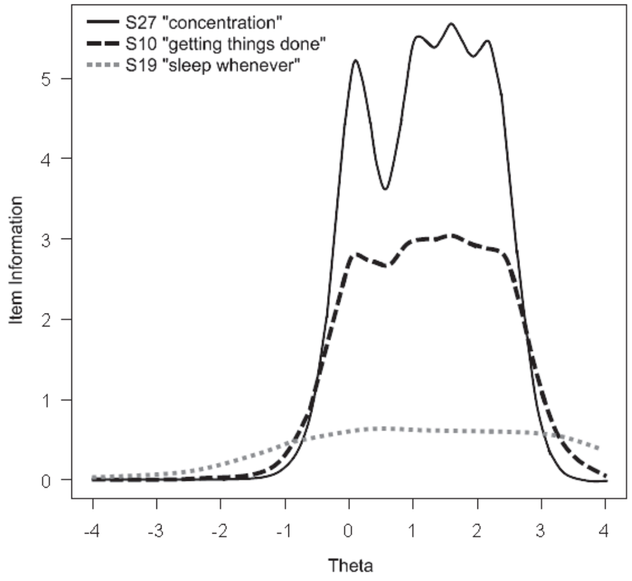

Figure S2b—Selected item-information curves for items on the sleep-related impairment item bank (S27: *I had a hard time concentrating because of poor sleep*; S10: *I had a hard time getting things done because I was sleepy*; S19: *I tried to sleep whenever I could*). Explanation per Figure 2a.

sleep disturbance or SRI) along a normal distribution of the underlying trait in the population. *Unidimensional* IRT models are those in which a single latent construct θ is responsible for overall test performance. Other important assumptions include monotonicity (i.e., that greater severity is actually reflected by the more "severe" item responses) and local independence (i.e., the absence of stronger than expected correlation among individual items), both of which are further described below.

Model selection and item calibration

The most commonly used IRT model for polytomous items (e.g., items with 5-point response scales) is the Graded Response Model (GRM[15]). GRM has 1 slope parameter and *n*-1 threshold parameters for each item, where *n* is the number of response categories. The slope parameter, a measure of item discrimination, indicates the shape of the category-response curves; the higher the slope parameter, the steeper the curves. More narrow and peaked curves indicate that the response categories differentiate well among different θ levels (sleep disturbance or SRI severity). Therefore, useful items have large slope parameters. The threshold parameter, a measure of item difficulty, indicates the location of the item on the θ scale and represents the θ level necessary to respond above the corresponding threshold with 0.50 probability. Items with overall lower threshold parameters identify lower levels of sleep disturbance or SRI severity (estimated by θ), and items with higher threshold parameters identify higher levels of severity. A constrained GRM was fit to the data with the discrimination parameters constrained to be equal across items (i.e., only 1 discrimination parameter was estimated for the entire item pool). χ^2 Tests were used to compare constrained models to general models, which permit different discrimination parameters across items. A nonsignificant χ^2 value indicates that the constrained model performed as well as the general model and the constrained model was preferred because of model parsimony. A significant χ^2value indicates that the general model fit the data better. Items were calibrated using MULTILOG 7[16] based on the model selected.

Item selection

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

Item information function: Items with low item information were considered to be poor items in IRT calibration. Two features that affect an item's total information functions are discrimination parameter estimates and the range of threshold parameter estimates. The higher the discrimination parameter, the more peaked the item information function. The wider the range of the threshold parameters, the flatter the item information function. Figures S2a and S2b show examples of items from the sleep-disturbance and SRI item banks with different discrimination and threshold parameters. Items with discrimination parameter estimates less than 1.0 were considered for exclusion.

Response distributions: Although it is common that the item frequency distribution is skewed for questionnaire data, items with sparse cell distributions can be particularly problematic. It is not, in principle, possible to obtain good estimates of parameters for response categories with very few observations. Given that the response distributions of sleep disturbance and SRI were skewed to the right (i.e., more precision could be obtained at higher severity levels), items with sparse cells in the top 2 response categories (*often* or *always*) were considered for exclusion.

Construct validity: Correlations with conventional measures used in the field, PSQI and ESS, were examined. Given the established literature about these scales and the content of the sleep-disturbance and SRI item banks, items demonstrating high correlation with the PSQI but low correlation with ESS were desired for the sleep-disturbance item bank; the opposite pattern of correlations was desired for the SRI item bank. To ensure that the sleep-disturbance and SRI item banks could be differentiated from other health item banks developed across the PROMIS network, correlations with the global item (*In general, would you say your health is? Excellent*, *Very good*, *Good*, *Fair*, *Poor*) and the fatigue item (*How would you rate your fatigue on average? None*, *Mild*, *Moderate*, *Severe*, *Very severe*) were examined. Items showing high correlations with these more-general dimensions were considered for exclusion.

Monotonicity/scalability: The monotonicity assumption is important for scales with ordered response categories, such as the sleep-disturbance and SRI items. 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. Taking sleep disturbance as an example, the probability of endorsing or selecting an item response indicating more severe sleep disturbance should increase as the underlying level of sleep disturbance (estimated by θ) increases. Two nonparametric methods were used to evaluate the monotonicity assumption. First, a Gaussian kernel smoothed model for nonparametric analysis, incorporated into the TestGraf program,[17] was used to visualize the empiric probability-curve estimates. The other method was to calculate *H* coefficient using Mokken scale analysis for polytomous items (MSP;[18]). *H* coefficient values less than 0.30 indicate that the corresponding item does not form a monotonic scale, which indicates scalability failure,[19] and such items were considered for exclusion.

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.[14] In essence, local dependence refers to an even stronger observed relationship among test items than would be expected based on the fact that they are part of a questionnaire developed to measure a single latent construct. A computer program for local dependence indices for polytomous items (LDIP[20]) was used to evaluate local independence. This program used the item-parameter estimations from MULTILOG to conduct item pair-wise comparison. The statistic Q3 was used to evaluate local dependence. As Yen[21] proposed, if the absolute value of Q3 is larger than 0.30, the corresponding item pair should be investigated further. However, realistic datasets typically consist of 1 or more item clusters. Local dependency, therefore, may be expected to arise in item clusters.[22] The nature of the sleep-disturbance and SRI item-bank development process may be expected to produce far more locally dependent item pairs than would conventional cognitive tests. A less restrictive Q3 of 0.50 was therefore used to consider items for exclusion.

Content expert review: Items were then reexamined from the clinical perspective by content experts (DJB, DEM, AG) 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, the next step was to estimate each respondent's location on the corresponding sleep-disturbance and SRI θ scales, i.e., to provide a "score" for each individual.[14] Unlike scoring under the CTT framework, which often simply sums the fixed values assigned to each response for each individual item, scoring under the IRT framework is affected by item-response patterns and individual item parameters. All 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. Two commonly used scoring strategies, which are also provided in MULTILOG 7, are maximum likelihood and maximum a posteriori. We used maximum likelihood for estimating individual scores on the sleep-disturbance and SRI θ 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. 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. We also compared θ scores between subjects who self-reported being treated for a sleep disorder and those who did not endorse treatment.

RESULTS

Model Selection and Item Calibration

Unidimensional GRMs were fitted to the sleep-disturbance and SRI item banks separately. A constrained GRM (i.e., a model in which the discrimination parameters were constrained to be equal across items, Rasch model) and a general GRM (i.e., a model in which the discrimination parameters were allowed to vary across items) were fitted to the data for model comparison. The χ^2 difference statistics were 1684 and 1682 for the sleep-disturbance and SRI item banks, respectively (P < 0.001 for each). This indicates that the general GRM fit the data better than did the constrained GRM for each item bank. Consequently, the general GRM was 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 in the main manuscript. Within these columns, *a* represents the slope parameter, which is an indicator of item information. Higher numbers indicate

Table S3—Items removed from the Patient-Reported Outcomes Measures Information System sleep-disturbance item bank during item response theory calibration

Items	Information Function	Response Distributions	Construct Validity	Local Dependence	Content Validity
S43: How often did you have difficulty falling asleep?					Removed
S50: I woke up too early and could not fall back asleep.	X				Added Back Based on Expert Item Review
S66: I felt jittery or nervous at bedtime.		X			
S73: I was afraid of going to bed.	X	X			
S74: I was afraid of going to sleep.	X	X			
S80: Worrying disturbed my sleep.				X	
S81: My sleep was disturbed by racing thoughts.					Removed
S83: My sleep was disturbed by sadness.	X	X			
S89: How long did it usually take you to fall back asleep after an awakening during the night?					Removed
S111: I wished I got more sleep each night.					Removed
S112: I had all the sleep I needed.				X	
S114: I was satisfied with the amount of sleep I got.				X	
S115: I was satisfied with my sleep.				X	Added Back Based on Expert Item Review
S117: I felt refreshed when I woke up.					Removed

that an item contains more information, i.e., discriminates better among individuals with high and low overall θ values. Parameters $b1$ through $b4$ represent threshold values for the individual responses. Lower values for $b1$ through $b4$ indicate items that detect lower levels of severity; higher values for $b1$ through $b4$ indicate items that detect greater levels of severity.

Item Selection

Sleep disturbance and SRI items were further refined following a 6-step item-selection procedure. Items removed from the sleep-disturbance and SRI item banks, and the reason(s) for their removal, are summarized in Tables S3 and S4.

Item information function

Item-information curves were used to visually examine the item performance, 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. In the samples evaluated, these items did not adequately discriminate individuals with higher and lower overall severity, estimated by θ.

Response distributions

Items with sparse endorsement in any of 5 response categories were examined. Because the item distributions were often skewed to the right (more precision could be obtained at higher severity levels), the top 2 response categories were examined in particular. 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 response percentages in the 2 most severe categories had few observations (less than 6% in total).

Construct validity

Correlations with conventional measures used in the field, PSQI and ESS, were examined first, followed by correlations with the global health and PROMIS fatigue scale items developed at another site. Items that showed higher correlations with

Table S4—Items removed from the Patient-Reported Outcomes Measures Information System sleep-related impairment item bank during item response theory calibration

Items	Information Function	Response Distributions	Construct Validity	Local Dependence	Content Validity
S4: I had enough energy.			X		Added Back Based on Expert Item Review
S5: I fell asleep when I did not mean to.	X				
S8: I fell asleep in public places (example: church, movie, work).	X	X			
S9: I felt sleepy when driving.	X	X			
S12: I made mistakes because I was sleepy.		X			
S17: I was fatigued.			X		
S18: I felt tired.			X		Added Back Based on Expert Item Review
S23: I napped.	X				
S26: I had trouble coping because of poor sleep.				X	
S28: I had a hard time thinking clearly because of poor sleep.		X		X	
S31: I had a hard time getting things done because of poor sleep.				X	
S32: I made mistakes because of poor sleep.		X		X	
S33: I had a hard time controlling my emotions because of poor sleep.		X			Added Back Based on Expert Item Review
S34: I avoided or cancelled activities with my friends because of poor sleep.		X			
S35: I felt clumsy because of poor sleep.		X			
S121: When I got out of bed I felt ready to start the day.					Removed
S126: I had to force myself to get up in the morning.					Removed
S128: How long did it take you to feel alert after waking up?					Removed

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 the TestGraf program indicated that all items showed good monotonic item-response curves, meaning that the probability of endorsing a more "severe" item response increased as the underlying θ increased. The observed H statistic values were in fact larger than 0.40 for all items (< 0.30 is the threshold for a problematic item), indicating that responses for all items followed a monotonic scale of increasing severity. Therefore, 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.*) were removed because they shared local dependence (exceeded a threshold value of > 0.50 on the Q3 statistic) with 5 or more other items. These locally dependent items were more strongly correlated with other items than would be expected by an individual's overall level of severity, indicating that they contributed little independent information to an individual's severity ranking.

Content expert review

Finally, 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.

Items removed from each of the above 6 steps are summarized in Table S3 for sleep disturbance and Table S4 for SRI. The final sleep-disturbance item bank consisted of 27 items, and the final SRI item bank consisted of 16 items. The final item IRT calibration values of the sleep-disturbance and SRI item banks are displayed in Tables 1 and 2 of the main manuscript.

Estimating Individual Scores

Based on the item-parameter values 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. The θ scale has a mean of 0 and a standard deviation of 1, with higher θ values corresponding to more severe disturbances. Sleep-disturbance θ scores ranged from -2.32 to 3.13, and SRI θ scores ranged from -2.27 to 3.29 for the entire calibration sample. 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 YouGov Polimetrix group with no sleep disorders, the YouGov Polimetrix group with self-reported sleep disorders, and the clinical group were -0.33, 0.25, and 0.88, respectively.

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. As hypothesized, subjects reporting each sleep disorder had higher θ values for both sleep disturbance and SRI, compared with those with no sleep disorder (Table 3, main manuscript). 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 responsive to treatment.

SUMMARY (from the main manuscript)

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.

REFERENCES

1. 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.
2. 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.
3. 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.
4. Food and Drug Administration. Guidence for industry: Computerized systems used in clinical trials. http://www.fda gov/cder/guidance/index.htm, 2007.
5. Basic documents, 46th ed. Geneva, Switzerland: World Health Organization; 2007.
6. Moul DE, Hall M, Pilkonis PA, et al. Self-report measures of insomnia in adults: rationales, choices, and needs. Sleep Med Rev 2004;8:177-98.
7. 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.
8. Muhr T, Freise S. User's Manual for ATLAS.ti 5.0, 2nd ed. Berlin, Germany: Scientific Software Development; 2004.
9. Wilkinson GS, Robertson GR. Wide-Range Achievement Test 4th ed. Lutz, FL: Psychological Assessment Resources, Inc.; 2006.
10. 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.

11. Muthen LK, Muthen BO. Mplus User's Guide, 4th ed. Los Angeles, CA: Muthen & Muthen; 2007.
12. Lord FM. Applications of Item Response Theory to Practical Testing Problems. New York, NY: Erlbaum Associates; 1980.
13. Kolen MJ, Brannan RL. Test Equating, Scaling, and Linking. Methods and Practices, 2nd ed. New York, NY: Springer-Verlag; 2004.
14. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
15. Samejima F. Estimation of latent ability using a repsonse pattern of graded scores. Psychometrika Monograph 1969;17.
16. Thissen D. MULTILOG 7: Multiple Categorical Item Analysis and Test Scoring Using Item Response Theory. Chicago, IL: Scientific Software, 2003.
17. Ramsay JO. TestGraf: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data. Montreal, QE: McGill University; 1995.
18. Molenaar IW, Sijtsma K. Users Manual MSP5 for Windows: A Program for Mokken Scale Analysis for Polytomous Items. Groningen, the Netherlands: iec ProGAMMA; 2000.
19. Mokken RJ. A theory and Procedure of Scale Analysis With Applications in Political Research. New York-Berlin: de Gruyter (Mouton); 1971.
20. 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.
21. Yen W. Effects of local dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement 1984;8:125-45.
22. Ip EH-S. Testing for local dependency in dichotomous and polytomous item response models. Psychometrika 2001;66:109-32.