Patient-Reported Outcomes Measurement Information System
Dynamic Tools to Measure Health Outcomes From the Patient Perspective

SLEEP DISTURBANCE

A brief guide to the PROMIS Sleep Disturbance instruments:

ADULT
PROMIS Item Bank v1.0 – Sleep Disturbance
PROMIS Short Form v1.0 – Sleep Disturbance 4a
PROMIS Short Form v1.0 – Sleep Disturbance 6a
PROMIS Short Form v1.0 – Sleep Disturbance 8a
PROMIS Short Form v1.0 – Sleep Disturbance 8b

ABOUT SLEEP DISTURBANCE

The PROMIS Sleep Disturbance instruments assess self-reported perceptions of sleep quality, sleep depth, and restoration associated with sleep. This includes perceived difficulties and concerns with getting to sleep or staying asleep, as well as perceptions of the adequacy of and satisfaction with sleep. Sleep Disturbance does not focus on symptoms of specific sleep disorders, nor does it provide subjective estimates of sleep quantities (total amount of sleep, time to fall asleep, amount of wakefulness during sleep). The Sleep Disturbance short form is universal rather than disease-specific. It assesses sleep disturbance over the past seven days.

Sleep Disturbance instruments are only available for adults (ages 18+).

(For complete definition see http://nihpromis.org/measures/domainframework1*)*

INTRODUCTION TO ASSESSMENT OPTIONS

There are two administration options for assessing Sleep Disturbance: <u>short forms</u> and <u>computerized adaptive test (CAT)</u>. When administering a short form, instruct participants to answer all of the items (i.e., questions or statements) presented. With CAT, participant responses guide the system's choice of subsequent items from the full item bank (27 items in total). Although items differ across respondents taking CAT, scores are comparable across participants. Some administrators may prefer to ask the same question of all respondents or of the same respondent over time, to enable a more direct comparability across people or time. In these cases, or when paper administration is preferred, a short form would be more desirable than CAT. This guide provides information on all Sleep Disturbance short form and CAT instruments.

Whether one uses a short form or CAT, the score metric is Item Response Theory (IRT), a family of statistical models that link individual questions to a presumed underlying trait or concept of sleep disturbance represented by all items in the item bank. When choosing between CAT and a short form, it is useful to consider the demands of computer-based assessment, and the psychological, physical, and cognitive burden placed on respondents as a result of the number of questions asked.

Figure 1

Figure 1 illustrates the correlations (strength of relationship) of the full bank with CAT and with short forms of varying length. The correlation of CAT scores with the full bank score is greater than a short form of any length. A longer CAT or longer short form offers greater correlation, as well as greater precision. When evaluating precision, not all questions are equally informative. The flexibility of CAT to choose more informative questions offers more
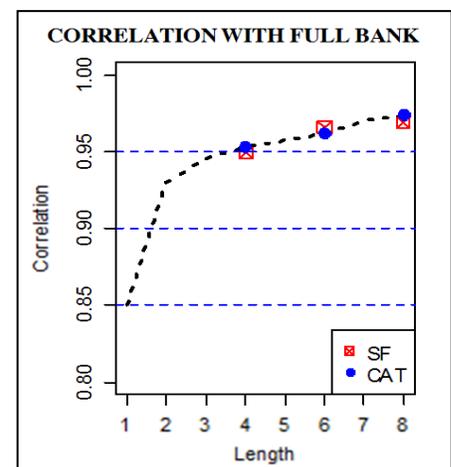
precision.

SHORT FORM DIFFERENCES

You will notice that there are 4 Sleep Disturbance short forms. Items in the 4a, 6a, and 8a short forms were selected based on rankings using two psychometric criteria: (1) maximum interval information; and 2) CAT simulations. Item rankings were similar for both criteria. For the maximum interval criterion, each item information function was integrated (without weighting) for the interval from the mean to 2 SDs worse than the mean. For the CAT simulations, responses to all items in each bank were generated using a random sample of 1,000 simulees drawn separately for each bank (centered on 0.5 SD worse than the general population mean). Items were rank ordered based on their average administration rank over the simulees. Content experts reviewed the items and rankings and made cuts of 4, 6, and 8 items. For each domain, 4-item, 6-item and 8-items have been selected so that the items are nested/overlap (e.g., the 8-item form is the 6-item form plus two additional items). The 4a, 6a, and 8a short forms can be administered with short forms of similar length from other domains (physical function, depression, anxiety, pain interference, fatigue, and satisfaction with participation in social roles v1.0) as part of a PROMIS Profile (see PROMIS-29, 43 or 57 Profile v2.0), though they can also be administered individually.

The original short form (8b) was constructed by the domain team with a focus on representing the range of the trait and also representing the content of the item bank. Domain experts reviewed short forms to give input on the relevance of each item. Each domain group worked independently and the original short forms are 6-10 items long depending on the domain. Psychometric properties and clinical input were both used and likely varied in importance across domains.

In selecting between short forms, the difference is instrument length. The reliability and precision of the short forms within a domain is highly similar. If you are working with an adult sample in which you wanted the most precise measure, select the 8a short form. If you are working in an adult sample in which you expected huge variability in a domain area and wanted different subdomains covered, you should select the 8b short form. If you had little room for additional measures but really wanted to capture something as a secondary outcome, you should use one of the shorter instruments (4a, 6a).

SCORING THE INSTRUMENT

Short Forms: PROMIS instruments are scored using item-level calibrations. This means that the most accurate way to score a PROMIS instrument is to utilize scoring tools within Assessment Center or API that look at responses to each item for each participant. Data collected in either of these platforms will automatically score in this way. We refer to this as "response pattern scoring." Response pattern scoring can be used when data was collected on paper or in another software package through the Assessment Center Scoring Service. Because response pattern scoring is more accurate than the use of raw score/scale score look up tables, it is preferred. However, if you aren't able to use response pattern scoring, you can use the instructions below which rely on raw score/scale score look-up tables.

Each question has five response options ranging in value from one to five. To find the total raw score for a short form with all questions answered, sum the values of the response to each question. For example, for the adult 8-item form, the lowest possible raw score is 8; the highest possible raw score is 40 (see all short form scoring tables in Appendix).

A score can be approximated if a participant skips a question. If items are missing, first check how many items were answered. For short forms with at least 5 items, confirm that 4 or 50% of items, whichever is greater, were answered. For example, a 4-item short form can only be scored with complete data. A 5-item short form can be scored as long as 4 items were answered. A 10-item short form can be scored as long as the participant answered at least 5 items. For branched instruments (e.g., Alcohol Use), the screening question is not used in calculating the score and therefore shouldn't be counted when assessing if the minimum number of items were answered. After confirming that enough responses were provided, sum the response scores from the items that were answered (not including any screening question). Multiply this sum by the total number of items in the short form. Finally, divide by the number of items that were answered. For example, if a respondent answered 5 of 8 questions and answered all items with the second lowest response option (2), you would sum all responses (10), multiply by the number of items in the short form (8) and divide by the number of items that were answered (5). Here (10x8)/5=16. If the result is a fraction, round up to the nearest whole number. This is a pro-rated raw score.

Again, the formula is:

$$\frac{\text{(Raw sum x number of items on the short form)}}{\text{Number of items that were actually answered}}$$

Locate the applicable score conversion table in the Appendix and use this table to translate the total raw score or pro-rated score into a T-score for each participant. The T-score rescales the raw score into a standardized score with a mean of 50 and a standard deviation (SD) of 10. Therefore a person with a T-score of 40 is one SD below the mean. It is important to note that Assessment Center will convert a participant's pattern of responses to a standardized T-score after they have finished a CAT. The standardized T-score is reported as the final score for each participant.

For the adult PROMIS Sleep Disturbance 8b short form, a raw score of 10 converts to a T-score of 35.9 with a standard error (SE) of 3.3 (see scoring table for the 8b short form in appendix). Thus, the 95% confidence interval around the observed score ranges from 29.4 to 42.4 (T-score $\pm$ (1.96*SE) or 35.9 $\pm$ (1.96*3.3).

For pro-rated scores, this calculation assumes that responses are missing at random. This isn't always true. Therefore, use caution when interpreting the final pro-rated T-score.

CAT: A minimum number of items (4) must be answered in order to receive a score for Sleep Disturbance CAT. The first item is selected because it provides the most information about the U.S. general population. The response to this item will guide the system's choice of the next item for the participant. The participant's response to this item will dictate the selection of the following question, and so on. As additional items are administered, the potential for error is reduced and confidence in the respondent's score increases. CAT will continue until either the standard error drops below a specified level, or the participant has answered the maximum number of questions (12), whichever occurs first.

For most PROMIS instruments, a score of 50 is the average for the United States general population with a standard deviation of 10 because calibration testing was performed on a large sample of the general population. However, Sleep Disturbance instruments such as these short forms were not calibrated on a national sample and so a score of 50 represents the average of the calibration sample which was generally more enriched for chronic illness. As these instruments, a score of 50 likely represents somewhat sicker people than the general population.   The T-score is provided with an error term (Standard Error or SE). The Standard Error is a statistical measure of variance and represents the "margin of error" for the T-score.

Important: *A higher PROMIS T-score represents more of the concept being measured.* For negatively-worded concepts like Sleep Disturbance, a T-score of 60 is one SD worse than average. By comparison, a Sleep Disturbance T-score of 40 is one SD better than average.

STATISTICAL CHARACTERISTICS

There are four key features of the score for Sleep Disturbance:

- **Reliability**: The degree to which a measure is free of error. It can be estimated by the internal consistency of the responses to the measure, or by correlating total scores on the measure from two time points when there has been no true change in what is being measured (for z-scores, reliability = $1 - SE^2$).
- **Precision**: The consistency of the estimated score (reciprocal of error variance).
- **Information:** The precision of an item or multiple items at different levels of the underlying continuum (for z-scores, information = $1/SE^2$).
- **Standard Error (SE):** The possible range of the actual final score based upon the scaled T-score. For example, with a T-score of 52 and a SE of 2, the 95% confidence interval around the actual final score ranges from 48.1 to 55.9 (T-score $\pm$ (1.96*SE) = 52 $\pm$ 3.9 = 48.1 to 55.9).

The final score is represented by the T-score, a standardized score with a mean of 50 and a standard deviation (SD) of 10.

In Figure 2 (adult 8b short form), the dotted horizontal line represents a degree of internal consistency reliability (i.e., .90 or .95) typically regarded as sufficient for an accurate individual score. The shaded blue region marks the range of the scale where measurement precision is comparable to the reliability of .90 for the eight-item form. Figure 2 also tells us where on the scale the form is most informative based upon the T-score. This form would typically be more informative than a Sleep Disturbance form with fewer items.
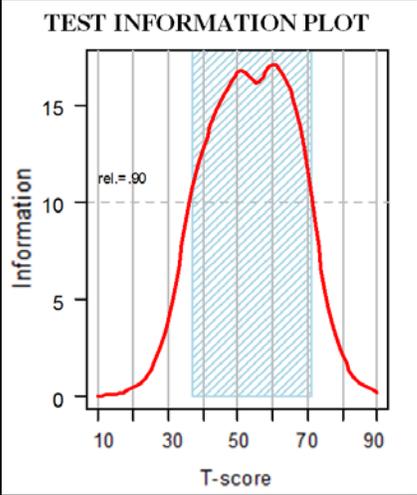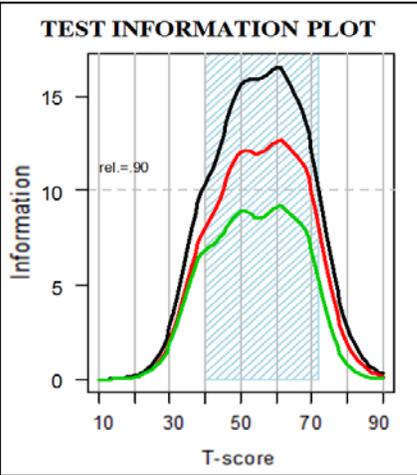

Figure 2

Figure 3

Figure 3 (adult 4a, 6a & 8a short forms) also tells us where on the scale the form is most informative based upon the T-score: the 8-item form is more informative than the 6-item form, which is more informative than the 4-item form.

Figure 4 is a sample of the statistical information available in Assessment Center for the Sleep Disturbance CAT.

More information is available online via Assessment Center (assessmentcenter.net).

Scaling Model Used For Calibration	Graded Response Model (GRM)
Total Number of Items	27

Sample	N	Alpha Reliability
Sleep Calibration ▼	2252	0.97

Score Distributions

	Mean	SD	P5	P10	P25	P50	P75	P90	P95
Raw	60.73	23.88	30.00	33.00	40.00	56.00	78.00	96.00	105.00
Scale	49.78	10.30	33.44	36.49	42.09	49.42	57.22	63.40	66.51

										Min	Max
Scale Score	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	10.0	90.0
SE	2.30	.90	.40	.20	.20	.20	.20	.30	.60		
Reliability	.00	.16	.87	.96	.97	.97	.97	.92	.68		

Figure 4

PREVIEW OF SAMPLE ITEM

Figure 5 shows a Sleep Disturbance item from the full item bank as it would appear to a study participant during data collection in Assessment Center. Several formats for presenting the items are available for computer-based administration through Assessment Center (see FAQ *section).*

PROMIS Assessment Center℠

In the past 7 days
I had trouble stopping my thoughts at bedtime
- ○ Not at all
- ○ A little bit
- ○ Somewhat
- ○ Quite a bit
- ○ Very much

Figure 5

Figure 6 is an excerpt from the paper version of the eight-item short form. This is the paper version format used for all Sleep Disturbance instruments. It is important to note, CAT is not available for paper administration.

	In the past 7 days…	Not at all	A little bit	Somewhat	Quite a bit	Very much
Sleep116 2	My sleep was refreshing............................	□ 5	□ 4	□ 3	□ 2	□ 1
Sleep20 3	I had a problem with my sleep	□ 1	□ 2	□ 3	□ 4	□ 5
Sleep44 4	I had difficulty falling asleep	□ 1	□ 2	□ 3	□ 4	□ 5

Figure 6

DATA REPORTS

Upon completion of a CAT for Physical Function, Sleep Disturbance, Anxiety, Pain Interference, Fatigue, Sleep Disturbance, and Satisfaction with Participation in Social Roles v1.0, a data report is available in Assessment Center. Figure 7 demonstrates some of the information available on the data reports.

Your score on the Sleep Disturbance CAT is __. The average score is 50. Your score indicates that your level of sleep disturbance is higher (worse) than:

· __ percent of people in the general population
· __ percent of (age group)
· __ percent of (gender)

Your score is shown below on the graph. The diamond (♦) is placed where we think your score lies. It represents your T-score, which is a standardized measure that is based on an average score of 50, based on responses to the same questions by the general population in the United States.

	Your Score	SE
Sleep Disturbance	52	2

10 20 30 40 50 60 70 80 90
Better   Average   Worse

The horizontal, solid black line on either side of the diamond in your profile report shows the Standard Error (SE), a statistical measure of variance that represents the possible range of your actual score around this estimated score. It is very likely that your score is in the range between these lines.
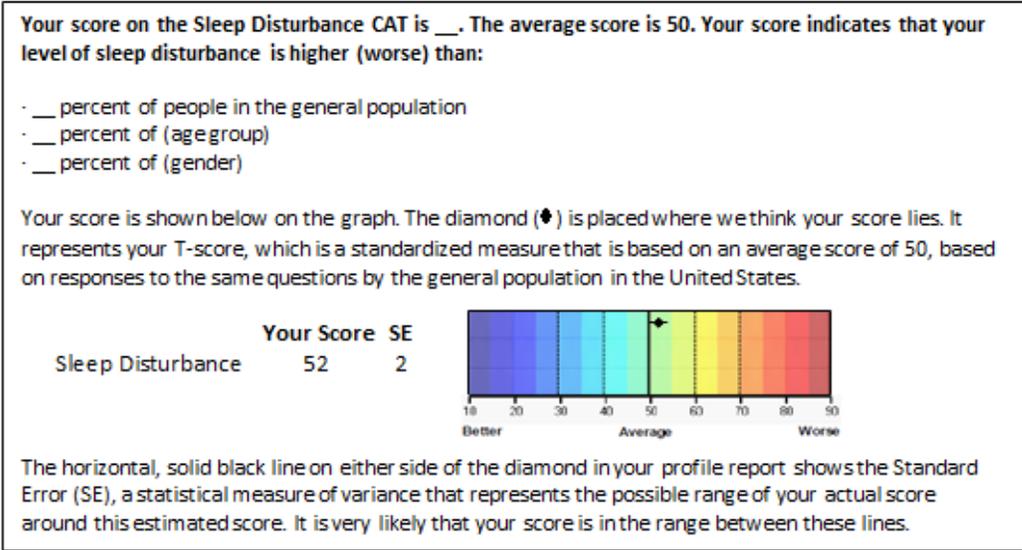
Figure 7

To access a sample report for Sleep Disturbance, complete the CAT demo at nihpromis.org. More than one CAT domain can be completed at a time; results for all domains selected will be generated and displayed within the one report.

Data reports are also available if you choose to administer a PROMIS Profile instrument, which includes a short form from seven PROMIS domains (Physical Function, Depression, Anxiety, Fatigue, Pain Interference, Satisfaction with Participation in Social Roles, and Sleep Disturbance).

FREQUENTLY ASKED QUESTIONS (FAQ)

Q: *I am interested in learning more. Where can I do that?*

All instruments are available on the PROMIS website through Assessment Center, which houses all PROMIS instruments for each domain.

Assessment Center is a free online research management tool. It enables researchers to create study-specific websites for capturing participant data securely. Studies can include measures within the Assessment Center library, as well as custom instruments created or entered by the researcher. PROMIS instruments (short forms, CATs, profiles) are a central feature of the instrument library within Assessment Center. Any PROMIS measure can be included in an online study or downloaded for administration on paper.

Detailed statistical information and development history about PROMIS items and instruments are available for review at nihpromis.org or assessmentcenter.net. To learn more, contact help@assessmentcenter.net.

Q: *Do I need to register with PROMIS to use these instruments?*

Yes, to get a copy of these instruments, we ask that you register with Assessment Center and endorse the PROMIS terms and conditions of use, so that we are better able to track who has accessed instruments for research. Assessment Center is available at assessmentcenter.net.

Q: *Are these instruments available in other languages?*

Yes, these instruments are currently available in Spanish in Assessment Center. The PROMIS group is also working to translate this form into other languages. Information on available translations is updated periodically at http://nihpromis.org/measures/translations.

Q: *Can I make my own short form?*

Yes, custom Sleep Disturbance short forms can be made by selecting any items from the item bank. Instructions for creating a custom short form in Assessment Center can be found in the Assessment Center User Manual https://www.assessmentcenter.net/UserManuals.aspx.

Q: *How do I handle multiple responses when administering a short form on paper?*

Guidelines on how to deal with multiple responses have been established. Resolution depends on the responses noted by the research participant.

- If two or more responses are marked by the respondent, and they are next to one another, then a data entry specialist will be responsible for randomly selecting one of them to be entered and will write down on the form which answer was selected. *Note: To randomly select one of two responses, the data entry specialist will flip a coin (heads - higher number will be entered; tails – lower number will be entered).To randomly select one of three (or more) responses, a table of random numbers should be used with a statistician's assistance.*

- If two or more responses are marked, and they are NOT all next to one another, the response will be considered missing.

Q: *What is the minimum change on a PROMIS instrument that represents a clinically meaningful difference?*

This question is related to an area of active research in the PROMIS network, namely the determination of the "minimally important difference" or "MID" for a PROMIS instrument. A manuscript in the *Journal of Clinical Epidemiology* outlines the process for MIDs for adult PROMIS measures and estimates the MIDs for six PROMIS-Cancer scales: Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six PROMIS-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64(5)*, 507-16.

As described in that manuscript, the MID is a tool to enhance the interpretability of patient-reported outcomes and is often defined as the "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989; 10(4):407-415).

APPENDIX-SCORING TABLES

Sleep Disturbance 8b
Short Form Conversion Table

Raw Score	T-score	SE*
8	28.9	4.8
9	33.1	3.7
10	35.9	3.3
11	38.0	3.0
12	39.8	2.9
13	41.4	2.8
14	42.9	2.7
15	44.2	2.7
16	45.5	2.6
17	46.7	2.6
18	47.9	2.6
19	49.0	2.6
20	50.1	2.5
21	51.2	2.5
22	52.2	2.5
23	53.3	2.5
24	54.3	2.5
25	55.3	2.5
26	56.3	2.5
27	57.3	2.5
28	58.3	2.5
29	59.4	2.5
30	60.4	2.5
31	61.5	2.5
32	62.6	2.5
33	63.7	2.6
34	64.9	2.6
35	66.1	2.7
36	67.5	2.8
37	69.0	3.0
38	70.8	3.2
39	73.0	3.5
40	76.5	4.4

*SE = Standard Error on T-score metric

Sleep Disturbance 4a
Short Form Conversion Table

Raw Score	T-score	SE*
4	32.0	5.2
5	37.5	4.0
6	41.1	3.7
7	43.8	3.5
8	46.2	3.5
9	48.4	3.4
10	50.5	3.4
11	52.4	3.4
12	54.3	3.4
13	56.1	3.4
14	57.9	3.3
15	59.8	3.3
16	61.7	3.3
17	63.8	3.4
18	66.0	3.4
19	68.8	3.7
20	73.3	4.6

*SE = Standard Error

Sleep Disturbance 6a
Short Form Conversion Table

Raw Score	T-score	SE*
6	31.7	5.1
7	36.9	3.9
8	40.1	3.5
9	42.5	3.3
10	44.6	3.2
11	46.4	3.1
12	48.0	3.0
13	49.5	3.0
14	50.9	3.0
15	52.3	2.9
16	53.6	2.9
17	54.8	2.9
18	56.1	2.9
19	57.3	2.9
20	58.5	2.9
21	59.7	2.9
22	61.0	2.9
23	62.3	2.9
24	63.6	2.9
25	65.0	2.9
26	66.5	3.0
27	68.1	3.1
28	70.0	3.3
29	72.4	3.6
30	76.1	4.4

*SE = Standard Error

Adult version

Patient-Reported Outcomes Measurement Information System
Dynamic Tools to Measure Health Outcomes From the Patient Perspective

PROMIS

Sleep Disturbance 8a
Short Form Conversion Table

Raw Score	T-Score	SE*
8	28.9	4.8
9	33.1	3.7
10	35.9	3.3
11	38.0	3.0
12	39.8	2.9
13	41.4	2.8
14	42.9	2.7
15	44.2	2.7
16	45.5	2.6
17	46.7	2.6
18	47.9	2.6
19	49.0	2.6
20	50.1	2.5
21	51.2	2.5
22	52.2	2.5
23	53.3	2.5
24	54.3	2.5
25	55.3	2.5
26	56.3	2.5
27	57.3	2.5
28	58.3	2.5
29	59.4	2.5
30	60.4	2.5
31	61.5	2.5
32	62.6	2.5
33	63.7	2.6
34	64.8	2.6
35	66.1	2.7
36	67.5	2.8
37	69.0	3.0
38	70.8	3.2
39	73.0	3.5
40	76.5	4.4

*SE= Standard Error on T-score metric
Adult version

Note: The Sleep Disturbance 8a table was revised on 5/22/2014. Instruments scored prior to this date should be re-scored using this table.

Conversion table applies only when ALL items on the short form have been answered. T-score metric is a linear transformation from the IRT theta scale: T-score=10*theta+50.