Psychological Test and Assessment Modeling, Volume 58, 2016 (2), 353-370

Measurement properties of PROMIS Sleep Disturbance short forms in a large, ethnically diverse cancer cohort

Roxanne E. Jensen[1,2], Bellinda L. King-Kallimanis[3], Eithne Sexton[4], Bryce B. Reeve[5,6], Carol M. Moinpour[7], Arnold L. Potosky[2,8], Tania Lobo[2] & Jeanne A. Teresi[9,10]

Abstract

AIMS: To evaluate model fit, differential item function (DIF), and construct validity of select short forms from the PROMIS® Sleep Disturbance item bank.

METHODS: We recruited cancer survivors who were between 6 - 13 months post diagnosis (n = 4,956), as part of the Measuring Your Health (MY-Health) study. We measured sleep disturbance using 10 items commonly found in PROMIS Sleep Disturbance short forms (Sleep 4a, Sleep 6a, Sleep 8b), and which are frequently administered in computerized adaptive testing.

We evaluated domain reliability using Cronbach's coefficient alpha and factorial validity by fitting a PROMIS Sleep Disturbance unidimensional measurement model using confirmatory factor analysis (CFA). At the item-level, we examined DIF with respect to race/ethnicity (non-Hispanic White [NHW], non-Hispanic Black [NHB], Hispanic, and Asian/Pacific Islander), age, and sex. We used a multi-group CFA and multiple indicators, multiple methods (MIMIC) analyses. We then assessed construct validity (convergent, discriminate, and known groups) for sleep short forms, and a new "best fit" 6-item sleep disturbance short form.

---

[1] *Correspondence concerning this article should be addressed to:* Roxanne E. Jensen, Ph.D., Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center, Georgetown University, 3300 Whitehaven Street NW, Suite 4100, Washington, DC 20007, USA; e-mail: rj222@georgetown.edu

[2] Department of Oncology, Georgetown University, Washington DC, USA

[3] Pharmerit International, Boston, MA, USA

[4] Royal College of Surgeons in Ireland, Dublin, Ireland

[5] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

[6] Department of Health Policy and Management, University of North Carolina, Chapel Hill, NC, USA

[7] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[8] Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center, Washington

[9] Columbia University Stroud Center and New York State Psychiatric Institute, New York, NY, USA

[10] Research Division, Hebrew Home at Riverdale, RiverSpring Health, Riverdale, NY, USA

RESULTS: We identified a satisfactory unidimensional sleep disturbance 6-item measure ($\chi^2(6)37.6$, $p < 0.001$, RMSEA = 0.031). To achieve this, we removed four items from the model with item content overlap and added residual covariances between positively worded items in order to address a method effect. We identified one instance of DIF: NHW participants were less likely to agree with the statement "I had difficulty falling asleep" compared to NHBs, Hispanics, or Asians/Pacific Islanders, who all reported the same level of sleep disturbance. After controlling for DIF, we extended this into a MIMIC model, identifying no additional DIF by age or sex. Across all race/ethnicity groups, the adjusted overall means suggest that older adults reported significantly lower sleep disturbance, and NHW, NHB, and Hispanic women reported significantly higher sleep disturbance than male survivors of the same race/ethnicity.

CONCLUSIONS: We could not fit a unidimensional measurement model for either the full 10-items, or for any combination of sleep disturbance items used in PROMIS Sleep Disturbance short forms. However, after we removed the overlapping item content and adjusted for methods effects, a 6-item measurement model for sleep disturbance fit the data well, with very little evidence of substantial DIF. This suggests this new measure (Sleep 6b) can be used in different groups across the adult lifespan, and in males and females in a heterogeneous cancer population. Our findings suggest further validation work is necessary to understand the impact of reverse-scored items, response set effects, and content overlap in this item bank.

Key words: sleep disturbance, PROMIS, differential item functioning, measurement invariance, methods effects

Introduction

Sleep problems are common for cancer patients, both during and after treatment (Garland et al., 2014). The prevalence of insomnia ranges from 30 % to 60 % (Savard, Simard, Blanchet, Ivers, & Morin, 2001); moreover, during chemotherapy, patients are three times more likely to report insomnia than the general population (Palesh et al., 2010). After treatment, insomnia symptoms can persist for up to 2 to 5 years (Savard & Morin, 2001). For cancer survivors, symptoms of insomnia frequently result in a higher risk of future physical and mental health problems, and subsequently, in a poorer quality of life (Garland et al., 2014). The availability of a valid and reliable self-report measure of sleep disturbance can help screen and identify cancer patients with clinically-relevant problems, monitor and evaluate supportive care services, and identify effective interventions; with the goal of improving the overall quality of life and functional ability of those assessed.

The Patient-Reported Outcomes Measurement Information System® (PROMIS®), a U.S. National Institutes of Health Common Fund initiative, includes a number of extensive item response theory (IRT)-calibrated item banks. Researchers can choose to administer domain items by computerized adaptive testing (CAT) or can select a subset of items for use as fixed length short forms (Buysse et al., 2010). The PROMIS Sleep Disturbance IRT-calibrated item bank includes items that measure perception of sleep quality, depth of sleep, satisfaction with sleep, and perception of difficulty getting and staying asleep (Cella et al., 2010). Qualitative methods were used to ensure content validity of the sleep disturbance

domain, including issues commonly identified by cancer patients (Flynn et al., 2010). Currently, four PROMIS Sleep Disturbance fixed-item short forms are widely available. These short forms vary in number of items (e.g., 4-, 6-, and 8- items) and are scored on the same *t*-score metric (50 = U.S. population mean score), with the longer short forms reporting higher reliability (Yu et al., 2011). Initial validation work has provided evidence that supports the reliability and validity of the PROMIS Sleep Disturbance measures in small clinical populations (Cella et al., 2010), and internet-based general samples (Buysse et al., 2010). However, to date, the PROMIS Sleep Disturbance measures have not been validated in an ethnically diverse, community-based sample, nor among oncology patients. Demonstrating the validity and reliability of the PROMIS Sleep Disturbance measure in this community-based sample of cancer patients will ensure that it is appropriate for use with study participants reporting mild to severe sleep disturbance.

Methods

Sample. Participants were from the Measuring Your Health (MY-Health) study cohort. Overall study design, recruitment strategy, and demographic characteristics of the MY-Health sample used in these analyses are described in the companion issue (see Jensen et al., 2016). We excluded participants without cancer stage, age, or race/ethnicity information, or those who did not answer all 10 sleep disturbance items in order to ensure all reliability and validity testing was completed using a single uniform cohort (n = 4,956).

Demographic and clinical variables. We collected registry-based information on age at diagnosis, sex, date of cancer diagnosis, cancer type, and cancer stage. Participant self-report information was collected on receipt of chemotherapy and hormonal therapy, race/ethnicity, comorbid conditions (number and type), education level, current employment status, annual income, marital status, insurance coverage, and information concerning whether or not the participant was born in the U.S.

Sleep disturbance. We evaluated the psychometric properties of three previously established sleep disturbance forms (4a, 6a and 8b forms). The 10 item form administered in this study also includes items that are frequently selected in the online PROMIS CAT assessment (Table 1). It was hypothesized that each of the four forms would be unidimensional and have satisfactory psychometric properties. PROMIS Sleep Disturbance 6b is a custom form, created for the analysis presented below. Higher scores reflect more sleep disturbance for all forms tested and all PROMIS sleep disturbance items are administered on a 5-point Likert scale. (Table 1)

Survey measures. In addition to the sleep disturbance domain, seven PROMIS domains were included in the MY-Health study based on their impact in cancer patient populations: emotional distress – anxiety (11 items); emotional distress – depression (10 items); fatigue (14 items); pain interference (11 items); physical function (16 items); cognitive function v.2 (8 items); and ability to participate in social roles and activities v.2 (10 items; Cella et al., 2007). Symptom thresholds reported for pain, anxiety, depression, and fatigue used in our known groups validity testing are defined elsewhere (Cella et al., 2014).

Table 1:
Item-Level and Short-Form Properties

PROMIS Item Identifier	Short Form					Floor (%)	Ceiling (%)	Mean	SD	Item Text
	Sleep 4a	Sleep 6a	Sleep 6b	Sleep 8b	Sleep 10					
Sleep87	--	--	X	X	X	22.1	8.0	2.7	1.2	I had trouble staying asleep.
Sleep90	--	--	--	X	X	23.7	7.6	2.6	1.2	I had trouble sleeping.
Sleep110	--	--	X	X	X	15.8	9.79	2.8	1.2	I got enough sleep.
Sleep109	X	X	--	X	X	11.0	5.0	2.7	1.0	My sleep quality was…
Sleep108	--	X	X	X	X	28.5	4.4	2.3	1.2	My sleep was restless.
Sleep115	--	--	X	X	X	15.7	17.2	3.0	1.3	I was satisfied with my sleep.
Sleep116	X	X	X	X	X	15.5	16.2	3.0	1.3	My sleep was refreshing.
Sleep44	X	X	X	X	X	37.9	6.4	2.2	1.3	I had difficulty falling asleep.
Sleep20	X	X	--	--	X	31.9	6.4	2.4	1.2	I had a problem with my sleep.
Sleep72	--	X	--	--	X	44.0	6.4	2.2	1.3	I tried hard to get to sleep.

	By Short Form					By Race/Ethnicity (Sleep 6b)				By Age (Sleep 6b)		
	Sleep 4a	Sleep 6a	Sleep 6b	Sleep 8b	Sleep 10	White	Black	Hisp.	Asian	21-49	50-64	65-84
Total Floor (%)	7.3	6.7	5.2	4.1	4.1	4.8	6.9	3.8	5.9	3.3	4.8	6.6
Total Ceiling (%)	2.0	1.5	1.2	1.1	1.1	0.8	1.7	1.6	1.1	2.6	1.0	0.7
Mean	50.3	50.8	50.4	50.4	50.8	50.2	50.7	51.6	49.2	52.9	51.2	48.3
SEM mean	3.5	3.1	3.3	2.6	2.4	3.2	3.3	3.3	3.3	3.2	3.2	3.3
SD	9.5	9.9	9.9	9.9	8.3	9.5	10.8	9.9	9.8	10.0	9.8	9.6
Cronbach's α	0.88	0.92	0.89	0.93	0.95	0.89	0.90	0.89	0.89	0.90	0.89	0.87

Non-PROMIS measures included in this study: the FACT-G Physical Well-Being (PWB) subscale, which includes seven items aimed at capturing concepts such as nausea, pain, and energy (Cella et al., 1993); the U.S. Acculturation Scale, reports adaptation of U.S. immigrants to U.S. culture (Marin, Sabogal, Marin, Otero-Sabogal, & Perez-Stable, 1987); and the Eastern Cooperative Oncology Group Performance Status Scale, which is a self-reported performance measure where 0 represents *no symptoms* and 5 equals *more than 2 hours a day* on bed rest (Oken et al., 1982).

Differential Item Functioning (DIF) hypothesis generation. The goal was to identify items that might have a different meaning or might not be understood well and/or equivalently by individuals of any of the groups referenced. This is often referred to as DIF, and it occurs when individuals from two or more groups with an equal standing on the trait of interest (i.e., sleep disturbance) have a different probability of responding a certain way to an item. Content experts (clinical or counseling psychologists, public health professionals, and a gerontologist) qualitatively reviewed the sleep items regarding potential sources of DIF. We asked the experts to rate each of the 10 sleep items with respect to gender, age, race/ethnicity, language, education, and diagnosis. They provided hypotheses in terms of presence and direction of DIF.

Experts did not identify race/ethnicity, language, or education-DIF hypotheses for any of the items. Some experts posited that, at the same level of sleep disturbance, women will report more trouble staying asleep, more trouble sleeping, poorer sleep quality, less refreshing sleep, and more problems with sleep than men.

Diagnosis- and age-DIF hypotheses were posited for most of the items. For example, conditional on sleep disturbance, raters hypothesized that cancer patients and those with chronic conditions, as well as older individuals, would be more likely to report more trouble staying asleep and with sleeping, more restless sleep, less satisfaction with their sleep, less refreshing sleep, more difficulty falling asleep, more problems with sleep, and trying harder to get to sleep. Cancer patients and those with a chronic illness were posited to be more likely to report not getting enough sleep. Older people were posited to report worse sleep quality, conditional on sleep disturbance. Given that all patients were diagnosed with cancer, hypotheses related to diagnosis could not be examined, but are given here for completeness, and for use in future work in which non-cancer patients are compared to those with cancer.

Statistical analysis. We examined distributional characteristics and missing data, including item-level floor and ceiling effects for each of the 10 sleep disturbance items and five short forms variants (three commonly used forms: 4a, 6a, 8b and two custom forms 6b, 10). We considered there to be a floor or ceiling effect when more than 20 % of responses were in the highest (ceiling) or lowest (floor) response category.

Differential Item Functioning. Items with DIF pose a threat to the validity of the scale, and conclusions drawn regarding the concept being measured may be biased. To determine whether DIF was present in the PROMIS Sleep Disturbance measure we used a three-step procedure: In Step 1, we established a measurement model using confirmatory factor analysis (CFA); this served the purpose of validating an assumption of unidimensionality for the measure. In Step 2, we investigated the assumption of measurement

invariance with respect to race/ethnicity by using a multi-group CFA model. In Step 3, we tested DIF, with respect to both age and sex, by extending the multi-group CFA, and by using the multiple indicator, multiple cause (MIMIC; Jöreskog & Goldberger, 1975; Muthén, 1984) modeling procedure. The MIMIC specification allows for multiple groups to be tested simultaneously (e.g., sex and race). These additional variables are modeled as single indicator exogenous variables in the MIMIC model and tested as possible violators of invariance. Weighted least square means and variance-adjusted estimator (WLSMV) for the estimation of all parameters, and for these analyses, were conducted using Mplus (version 6.12; Muthén & Muthén, 1998-2015).

We used the chi-square test of exact fit, the root mean square error of approximation (RMSEA) and the comparative fit index (CFI; Browne & Cudeck, 1992) to assess overall model goodness-of-fit of each measurement model tested. A RMSEA value of < 0.08 indicates satisfactory model fit and a value of < 0.05 indicates close model fit. However, when using the WLSMV estimation with a large sample as was done here, it has been suggested that cut-off scores should be more conservative (RMSEA < 0.045 and CFI > 0.95; Yu, 2002).

Step 1: Measurement model

For Step 1, we tested a unidimensional measurement model using CFA for three commonly used forms: 4a, 6a, 8b and two custom forms 6b, 10. Our aim was to determine whether these commonly used forms had satisfactory measurement properties and to determine an appropriate model for testing DIF.

Step 2. Multi-group CFA. For Step 2, we tested measurement invariance and the presence of DIF using a multi-group CFA, based on the satisfactory measurement model from Step 1. First, all parameters were freely estimated to establish model fit; next, we simultaneously constrained the factor loadings and intercepts for each item to be equal across the four groups (Black, White, Asian/Pacific Islander, Hispanic). A strong indicator of DIF is a significant deterioration in the fit of this fully constrained model when compared to the model with all parameters fully estimated. If the fit of the model deteriorated, we used the modification indices (MI) and standardized expected parameter changes (EPC) to determine which items exhibited DIF. The EPC indicates the size of the expected change in the parameter estimate were it to be freed. As there were a large (80) number of tests under consideration, to maintain a family wise Type I error rate of 5 %, a Bonferroni-adjusted critical value of 11.7 was considered significant (Holm, 1979). The EPC was evaluated for salient magnitude based on Cohen's *d* effect sizes (Cohen, 1988) with a small effect size considered substantial.

If the modification index and standardized EPC met the criteria described above for a factor loading or intercept, we removed the corresponding equality constraint and freely estimated the parameters for that ethnic group. The appropriateness and significance of a change made to the model was assessed using the chi-square difference test; the difference in the fit of the null and alternative models was tested using the Bonferroni-adjusted *p*-value in order to reduce the Type I error rate. The process of identifying DIF was an

iterative process, changing one parameter at a time, until no more modification indices and EPCs met the criteria.

MIMIC modeling (Age and Sex). Using the final multi-group CFA model from Step 2, we extended it to the MIMIC model in Step 3 by including age and sex as additional exogenous variables. Both age and sex were regressed on the latent variable, but all direct effects of age and sex on the observed items were fixed to zero. DIF is indicated by a significant direct effect of an exogenous variable on an observed variable. We used the same strategy for detecting DIF with respect to race/ethnicity; the critical value for the modification indices was 10.8. This MIMIC model is described in the methods overview article in this series (Teresi & Jones, 2016).

Reliability and Validity Testing Step. In this additional step, we used classical test theory psychometric procedures to evaluate the reliability and validity (Nunnally, 1978) of each PROMIS Sleep Disturbance short form across three age (21 - 49, 50 - 64, 65 - 84) and four race/ethnicity (NHW, NHB, Hispanic, Asian/Pacific Islander) groups.

Reliability. We evaluated overall and item-level performance. First, we estimated internal consistency using Cronbach's coefficient alpha (Cronbach, 1951), with $\alpha > 0.70$ and $\alpha > 0.90$: the thresholds for reliable group and individual level (inter-individual comparisons at a single time point) measurement, respectively.

Validity. Although we did not perform formal tests of discriminant and convergent validity by constructing multi-trait, multi-method matrices, we provide some preliminary evidence for validity by examining the correspondence between hypothetical and observed relationships of the sleep measure to other measures. We examined convergent and discriminant construct validity by calculating Pearson correlations between the sleep disturbance measures and six other PROMIS measures included in this study. In terms of convergent validity, we hypothesized that symptoms (anxiety, depression, fatigue, pain interference) would show the highest correlations with sleep disturbance, reflecting common symptom clusters for cancer patients (Fan, Filipczak, & Chow, 2007). We expected lower, but moderate correlations (0.3 - 0.7) of sleep measures with function measures (physical function and ability to participate in social roles). Among other validated measures, we expected a moderate association with the physical well-being (PWB), and no meaningful correlations with an acculturation scale for U.S. immigrants.

We created known groups, the process of assessing the extent to which scores can distinguish among groups, to reflect findings from previous research in cancer patients. The groups were created based on demographic and clinical variables, symptoms, and probability of sleep disturbance. Three demographic groups were created that reflect differences in the non-cancer, general population, as females, those of younger age, and of lower education level should report higher sleep disturbance (Grandner et al., 2012; Grandner et al., 2010). Clinical and symptom groups were formed based on research findings indicating higher sleep disturbance scores reported for younger patients (Davidson, MacLean, Brundage, & Schulze, 2002), and current indication or history of clinically-relevant symptoms (e.g., pain, depression, fatigue; Bower, 2008; Irwin, 2013; Irwin, Olmstead, Ganz, & Haque, 2013). Other studies have identified lower mean sleep disturbance scores for cancer survivors reporting regular exercise (Tomlinson, Diorio,

Beyene, & Sung, 2014). For probability of sleep disturbance, two groups were created that represent patients most likely to report a high sleep disturbance score (history of sleep problem, moderate or higher pain interference or fatigue, and on bed rest) or a low sleep disturbance score (no clinically meaningful fatigue or pain, vigorous exercise five times a week or more, reporting no symptoms). T-tests were used to evaluate group differences in sleep disturbance.

Results

Overall, the MY-Health cohort was diverse with respect to age, race/ethnicity and educational status. Less than half of survey respondents (42 %) were White; 60 % were female; 58 % were under 64 years of age; and 18 % reported less than a high school diploma. A relatively large proportion of Hispanic and Asian/Pacific Islander participants were born outside the US (42 % and 17 %, respectively), and 15 % of the sample reported a history of a diagnosed sleep disorder.

Step 1: Measurement Model

The mean sleep disturbance item scores were between 2.2 and 3.0 (range: 1 - 5), with six of the 10 items showing a floor effect (Table 1). However, the fit for our measurement model to establish unidimensionality across all 10 items was unsatisfactory (χ^2 8722.60 (35), $p < 0.001$, RMSEA 0.227; Table 2). Model fit was also unsatisfactory for the 4a, 6a, and 8b sleep short forms (Table 2).

Examination of descriptive statistics revealed two issues that affected the analyses. The first issue was a method effect, when item responses reflect not just the latent factor and residual error, but also item characteristics such as wording direction (Marsh, 1996). Upon further inspection, the reversed items for some participants appeared to have inaccurate endorsement. For example, of the 495 patients who chose the response option of *never* on "I had trouble staying asleep," 34 % also chose the response option of *never* on "I got enough sleep," a positively worded item. The modification indices were suggestive of a negative and positive factor rather than measures of two separate constructs. We addressed this effect by allowing residuals of items with the same wording direction (positive) to covary. We considered an alternative method of adding a latent method factor, but chose not to do so, due to the possibility of creating an under-identified model (Kenny & Kashy, 1992).

A second methods issue affecting the analyses was the presence of high inter-correlations among a subset of sleep items, violating the assumption of local independence. For example, the items "My sleep quality was [poor to excellent]," "I tried hard to get to sleep," "I had trouble sleeping," and "I had a problem with my sleep," evidenced high inter-correlations with other items in the scale ranging from 0.8 to 0.9. This item overlap was reflected in the poor localized fit associated with each of these items, which was evidenced by large modification indices. Items for which the assumption of local independence was violated were examined and removed one at a time. The first item

Table 2:
Overall goodness-of-fit and chi-square difference test for complete sample and sub-sample likely to experience sleep disturbance

Model	**Chi-square (*df*)**	***p*-value**	**RMSEA 95 % CI**	**CFI**	
Sleep 4a	782.21 (2)	<0.0001	0.283 0.267 ; 0.300	0.99	
Sleep 6a	2571.89 (9)	<0.0001	0.242 0.234 ; 0.250	0.98	
Sleep 8b	5374.75 (20)	<0.0001	0.235 0.229 ; 0.240	0.97	
Sleep 10 (custom)	8722.60 (35)	<0.001	0.227 0.223 ; 0.231	0.96	
Sleep 6b (custom)	34.31 (6)	<0.001	0.031 0.022 ; 0.042	0.99	
Sub-group – Likely to experience sleep disturbance group *n* = 307					
Sleep 4a	44.14 (2)	<0.001	0.262 0.198 ; 0.332	0.98	
Sleep 6a	131.44 (9)	<0.001	0.211 0.180 ; 0.243	0.97	
Sleep 8b	500.23 (20)	<0.001	0.280 0.259 ; 0.301	0.94	
Sleep 10 (custom)	724.12 (35)	<0.001	2.531 0.237 ; 0.269	0.93	
Sleep 6b (custom)	3.88 (6)	0.693	0.000 0.000 ; 0.057	0.99	
Multiple Group – CFA Models and MIMIC models					
Model	**Chi-square (*df*)**	***p*-value**	**RMSEA 95 % CI**	**Chi-square difference**	**CFI**
Free parameter	63.36 (24)	<0.0001	0.037 0.026 ; 0.048	NA	0.99
Fixed	362.13 (90)	<0.0001	0.050 0.045 ; 0.055	292.67 (66) <0.001	0.99
“I had a problem falling asleep”	309.26 (88)	<0.0001	0.046 0.040 ; 0.051	40.12 (2) <0.001	0.97
“I got enough sleep”	306.60 (87)	<0.001	0.046 0.040 ; 0.051	4.31 (1) 0.038	0.96
MIMIC	192.40 (128)	0.0002	0.020 0.014 ; 0.026	NA	0.99

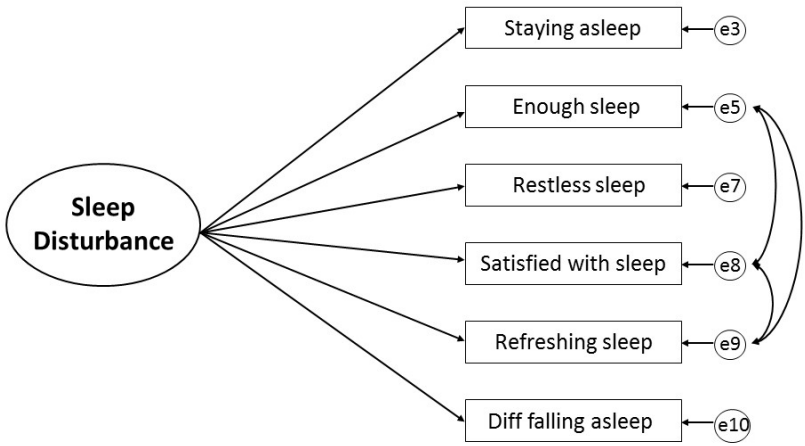

Figure 1:
PROMIS Sleep Disturbance Measurement Model

removed was "I tried hard to get to sleep" (MI = 2,775), followed by "I had trouble sleeping" (MI = 1,566), "I had a problem with my sleep" (MI = 822), and "My sleep was restless" (MI = 249). After removing items that violated assumptions of local independence, and after controlling for a methods effect due to positively worded items, we identified a 6-item measurement model (sleep 6b). The sleep 6b form evidenced a satisfactory model fit (χ^2 34.31 (6), $p < 0.001$, RMSEA 0.031, CFI = 0.99) using the RMSEA and CFI cutoffs (see Table 2 and Figure 1).

Cross-validation of the results was performed by estimating the measurement models for the short forms and custom forms using a sub-group of the sample that were likely to experience sleep disturbance (using our known groups definition; Table 2). As can be seen from the results, the overall model fit results followed a pattern similar to that of the entire cohort.

Step 2: Multi-group CFA

Using the 6-item model with covaried residuals from Step 1, we investigated the possibility of DIF across four race/ethnic groups. We found that constraining the factor loadings and intercepts to equality across the four groups led to a significant deterioration in model fit (χ^2 difference, 292.67 (66), $p < 0.001$, Table 2), suggesting the presence of DIF.

We identified DIF for the factor loading (MI = 35.94, standardized EPC = -0.12) for the item, difficulty falling asleep. Thus, the equality constraints for this item for non-Hispanic Whites were removed. This change improved significantly the overall model fit (χ^2 difference, 40.12 (2), $p < 0.001$, Table 2), with no further indications of DIF.

Step 3: MIMIC model

Extending our final multi-group CFA model accounting for DIF in the item, difficulty falling asleep from Step 2 into a MIMIC model resulted in the identification of no instances of DIF by age or sex.

Reliability

Performance of the PROMIS short forms was generally consistent with mean sleep scores ranging from 50.3 – 50.8, and Cronbach's α scores between 0.88 - 0.95 (Table 1).

Validity

We found convergent and discriminant construct validity to be generally consistent with expectations, and to be stable across short form versions. Notably, all symptom and function domains reported a similar magnitude of association (r = 0.4 - 0.6). Findings also supported the discriminate validity of the short form with the U.S. acculturation measure (r = 0.05 - 0.06) (Table 3).

Known group comparisons identified significant mean score differences (Table 4). Among demographic variables, patients who were younger, female, and those with lower education levels reported higher (between 2.4 to 4.5 points) sleep disturbance (all $p <$ 0.001) as contrasted with the reference group. In contrast to the reference groups, groups created based on clinical characteristics showed clinically meaningful differences (5.4

Table 3:
Convergent and Discriminant Validity by Sleep Disturbance Short Form

Score Correlations	Sleep [10]	Sleep [4a]	Sleep [6a]	Sleep [6b]	Sleep [8b]	Hypothesized Association
PROMIS Domains						
Ability to Participate in Social Roles	-0.49	-0.48	-0.50	-0.47	-0.48	+
Fatigue	0.54	0.53	0.54	0.53	0.54	+
Anxiety	0.54	0.53	0.54	0.52	0.53	+
Depression	0.51	0.51	0.52	0.50	0.51	+
Pain Interference	0.46	0.45	0.46	0.44	0.45	*
Cognitive Function	-0.45	-0.44	-0.45	-0.44	-0.45	*
Physical Function	-0.41	-0.40	-0.41	-0.39	-0.40	*
Validated Measures						
FACT – Physical Well-Being Subscale	-0.51	-0.51	-0.52	-0.50	-0.51	*
Acculturation**	0.05	0.06	0.06	0.06	0.06	-

+Strong ($r \geq 0.70$); * Moderate ($0.30 < r < 0.70$); - Weak ($r \leq 0.30$); **Non U.S. born only

points) due to multi-morbidity (self-report count: two or more conditions vs. no conditions), cancer type (prostate vs. lung, 3.7 points), and specific comorbid conditions (depression, sleep disorder: 6.6 and 6.7 points, respectively). Mean differences on the PROMIS Sleep Disturbance measure between those receiving and not receiving cancer treatment was statistically significant, but small (1.1 - 2.7), and patients receiving a diagnosis of advanced stage cancer did not show a statistically significant difference from those without advanced stage cancer in sleep disturbance (0.7; $p = 0.13$). Participants with either severe levels of depression or anxiety reported mean sleep disturbance scores that were above 60 (all $p = <0.001$), a full standard deviation above the general U.S. population.

Table 4:
Known Group Comparisons (Sleep 6b)

Known Groups Comparisons	**Group 1:**		**Group 2:**		**Mean Group Difference**	***p*-value**
	Mean	***SD***	**Mean**	***SD***		
Demographic						
Sex: Women vs. Men	51.3	10.0	49.0	9.8	2.4	< 0.001
Age: Young (21 - 49) vs. Old (65 - 84)	52.9	10.0	48.3	9.6	4.5	< 0.001
Education: Low (<HS) vs. High (College)	52.1	10.3	48.5	9.5	3.5	< 0.001
Clinical						
Cancer Stage: Advanced vs. Localized	51.0	10.1	50.3	9.9	0.7	0.13
Cancer Site: Lung vs. Prostate	51.6	10.2	48.0	9.5	3.7	< 0.001
Number of Comorbid Conditions: 2+ vs. 0	53.2	9.7	47.8	9.7	5.4	< 0.001
Treatment: Chemotherapy (Yes vs. No)	51.8	10.0	49.1	9.8	2.7	< 0.001
Treatment: Radiation (Yes vs. No)	51.0	9.9	49.9	9.9	1.1	0.002
History of a Sleep Disorder (Yes vs. No)	56.0	9.6	49.2	9.7	6.7	< 0.001
History of Depression (Yes vs. No)	55.5	9.5	49.0	9.7	6.6	< 0.001
Vigorous Exercise: None vs. 5+ / Week	51.3	10.0	46.5	9.2	4.8	< 0.001
Symptoms						
Pain Interference (Severe vs. None)	60.5	8.9	46.2	9.3	14.3	< 0.001
Fatigue (Severe vs. None)	62.3	10.0	45.5	8.8	16.8	< 0.001
Depression (Severe vs. None)	61.6	8.5	47.7	9.3	13.9	< 0.001
Probability of Sleep Disturbance						
Low Probability (Floor) (No vs. Yes)	50.7	9.9	45.4	8.8	5.3	< 0.001
High Probability (Ceiling) (Yes vs. No)	56.7	9.1	49.2	9.7	7.4	< 0.001

Note: Groups with an a priori hypothesis of higher sleep disturbance are listed in group 1

Discussion

Overall, this study provided evidence in support of the reliability and validity of a new custom PROMIS Sleep Disturbance measure (6b) in a large, diverse U.S. cancer patient cohort. We found ceiling and floor effects to be minimal. The form demonstrated construct validity, supporting convergent and discriminant construct validity for sleep disturbance reported in other clinical populations (Khanna et al., 2012).

Unlike previous evaluations of the sleep disturbance domain (Buysse et al., 2010; Cella et al., 2010), local independence and unidimensionality assumptions were not met in this patient sample for all 10 sleep items, or for any short form. A lack of unidimensional model fit when measuring health domains is not uncommon (Cook, Kallen, & Amtmann, 2009). However, the model fit issues we identified were substantial, and could also be due to a method or administration effect. Because findings in this study sample do not reflect past validation work in smaller samples with sleep disorders, further work is needed to understand why findings could not be replicated. Potential reasons include specific characteristics of this patient population, as cancer patients may report a greater degree of sub-clinical sleep disturbance which may not be measured as well by this sleep disturbance measure. Additionally, the measure has not been examined in ethnically diverse groups, with adequate representation of individuals of older age and lower education; such cultural and demographic characteristics may have contributed to variation in interpretation and in item response. Alternatively, survey design, and/or administration method (paper survey vs. electronic system), may have contributed to the model fit issues identified in this study. However, studies of other short form PROMIS measures did not identify DIF with respect to mode of administration (Bjorner et al., 2014).

We identified reverse-scored items as a method effect, which was an important source of model misfit in this patient population. The sleep disturbance domain included four reverse-scored items out of the 10 investigated. Among PROMIS domains, few contain reverse-scored items. For example, in this study, only sleep disturbance and fatigue contain reverse-scored items, while all other PROMIS items are worded and scored in the same direction. As with sleep disturbance, PROMIS fatigue also showed poor model fit when including a reverse-scored item (Reeve et al., 2016). Method effects similar to those identified here are frequently identified in questionnaires with reverse-scored items (Abbott et al., 2006; Wood, Taylor, & Joseph, 2010). Reverse-scored items are typically included to minimize acquiescence-bias. However, it has been argued that their use is based more on convention rather than evidence for necessity or effectiveness (van Sonderen, Sanderman, & Coyne, 2013). The evidence presented here adds to arguments that such items may confuse some participants and complicate the interpretation of scores. Because PROMIS item banks allow for a high degree of item customization and comparison in health domains, these findings suggest that investigators should consider the necessity of including reverse-scored items, due to specific clinical content relevance and response set changes within a domain, in order to ensure minimization of measurement error. Further work is needed to see if these issues are consistent across other patient populations.

A possible limitation is that correlated residuals were included in the model and one item was found to have DIF, albeit of low impact. The effects of such modeling on the practical use of a short form scale is an area that requires further research. Another limitation is the inability to examine different ethnic subgroups within major categories, e.g., Hispanic and Asian.

After accounting for model fit issues, we found one instance of DIF for non-Hispanic White patients and no DIF by age or sex; despite hypotheses that DIF would be observed for sex and/or age for many items. The lack of DIF by age, group, or sex is consistent with previous evaluations of the PROMIS Sleep Disturbance measure in other clinical populations (Cook, Bamer, Amtmann, Molton, & Jensen, 2012). However, no previous study to date has evaluated PROMIS sleep disturbance DIF by race/ethnicity. While there were no a priori hypotheses for findings of DIF by race/ethnicity groups, DIF was identified for one item; however, there was little impact of DIF on the estimated sleep disturbance mean scores per group. Additional validation is necessary to establish whether the DIF identified in this study is meaningful across all administrations, or if it is specific to this participant sample.

Overall, the small magnitude of DIF, together with the strong and consistent evidence of validity and reliability, supports the use of the new 6 item PROMIS Sleep Disturbance short form in research settings. Because of the various methods effects identified, investigators should consider carefully the potential implications of selection and use of reverse-scored items and response set changes within these short forms.

Funding

U01AR057971 (National Institute of Arthritis & Musculoskeletal & Skin Diseases), P30CA051008 (National Cancer Institute), KL2TR000102 (National Center for Research Resources (NCRR) and the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through the Clinical and Translational Science Awards Program (CTSA))

Special thanks

Charlene Kuo, Kevin Camstra, Lindsay Wright

References

Abbott, R. A., Ploubidis, G. B., Huppert, F. A., Kuh, D., Wadsworth, M. E., & Croudace, T. J. (2006). Psychometric evaluation and predictive validity of Ryff's psychological well-being items in a UK birth cohort sample of women. *Health and Quality of Life Outcomes, 4*, 76. doi:10.1186/1477-7525-4-76

Bjorner, J. B., Rose, M., Grandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. (2014). Difference in method of administration did not significantly impact item response: an

IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. *Quality of Life Research, 23*, 217-227. doi: 10.1007/s11136-013-0451-4

Bower, J. E. (2008). Behavioral symptoms in patients with breast cancer and survivors. *Journal of Clinical Oncology, 26*(5), 768-777. doi:10.1200/JCO.2007.14.3248

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*(2), 230-258. doi:10.1177/0049124192021002005

Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., . . . Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep, 33*(6), 781-792.

Cella, D., Choi, S., Garcia, S., Cook, K. F., Rosenbloom, S., Lai, J. S., . . . Gershon, R. (2014). Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. *Quality of Life Research, 23*(10), 2651-2661. doi:10.1007/s11136-014-0732-6

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Group, P. C. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology, 63*(11), 1179-1194. doi:10.1016/j.jclinepi.2010.04.011

Cella, D., Tulsky, D. S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., . . . Brannon, J. (1993). The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure. *Journal of Clinical Oncology, 11*(3), 570-579.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of a NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11. doi:10.1097/01.mlr.0000258615.42478.55

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cook, K. F., Bamer, A. M., Amtmann, D., Molton, I. R., & Jensen, M. P. (2012). Six Patient-Reported Outcome Measurement Information System short form measures have negligible age- or diagnosis-related differential item functioning in individuals with disabilities. *Archive of Physical Medicine and Rehabilitation, 93*(7), 1289-1291. doi:10.1016/j.apmr.2011.11.022

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*(4), 447-460. doi:10.1007/s11136-009-9464-4

Cronbach, L. J. (1951). Coefficient alpha and the internal study of tests. *Psychometrika, 16*, 297-334. doi:10.1007/BF02310555

Davidson, J. R., MacLean, A. W., Brundage, M. D., & Schulze, K. (2002). Sleep disturbance in cancer patients. *Social Science & Medicine, 54*(9), 1309-1321. doi:10.1016/S0277-9536(01)00043-0

Fan, G., Filipczak, L., & Chow, E. (2007). Symptom clusters in cancer patients: A review of the literature. *Current Oncology, 14*(5), 173-179.

Flynn, K. E., Shelby, R. A., Mitchell, S. A., Fawzy, M. R., Hardy, N. C., Husain, A. M., . . . Weinfurt, K. P. (2010). Sleep-wake functioning along the cancer continuum: Focus group results from the Patient-Reported Outcomes Measurement Information System (PROMIS™). *Psycho-Oncology, 19*(10), 1086-1093. doi:10.1002/pon.1664

Garland, S. N., Johnson, J. A., Savard, J., Gehrman, P., Perlis, M., Carlson, L., & Campbell, T. (2014). Sleeping well with cancer: A systematic review of cognitive behavioral therapy for insomnia in cancer patients. *Neuropsychiatric Disease and Treatment, 10*, 1113-1124. doi:10.2147/NDT.S47790

Grandner, M. A., Martin, J. L., Patel, N. P., Jackson, N. J., Gehrman, P. R., Pien, G., . . . Gooneratne, N. S. (2012). Age and sleep disturbances among American men and women: Data from the U.S. Behavioral Risk Factor Surveillance System. *Sleep, 35*(3), 395-406. doi:10.5665/sleep.1704

Grandner, M. A., Patel, N. P., Gehrman, P. R., Xie, D., Sha, D., Weaver, T., & Gooneratne, N. (2010). Who gets the best sleep? Ethnic and socioeconomic factors related to sleep complaints. *Sleep Medicine, 11*(5), 470-478. doi:10.1016/j.sleep.2009.10.006

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. doi:10.2307/4615733

Irwin, M. R. (2013). Depression and insomnia in cancer: Prevalence, risk factors, and effects on cancer outcomes. *Current Psychiatry Reports, 15*(11), 404. doi:10.1007/s11920-013-0404-1

Irwin, M. R., Olmstead, R. E., Ganz, P. A., & Haque, R. (2013). Sleep disturbance, inflammation and depression risk in cancer survivors. *Brain, Behavior, and Immunity, 30*(Suppl), S58-67. doi:10.1016/j.bbi.2012.05.002

Jensen, R. E., Moinpour, C. M., Keegan, T. H. M., Cress, R. D., Wu, X.-C., Paddock, L. A., . . . Potosky, A. L. (2016). The Measuring Your Health Study: Leveraging community-based cancer registry recruitment to establish a large, diverse cohort of cancer survivors for analyses of measurement equivalence and validity of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short form items. *Psychological Test and Assessment Modeling, 58*, 99-117.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351a), 631-639. doi:10.1080/01621459.1975.10482485

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*(1), 165-172.

Khanna, D., Maranian, P., Rothrock, N., Cella, D., Gershon, R., Khanna, P. P., . . . Hays, R. D. (2012). Feasibility and construct validity of PROMIS and "legacy" instruments in an academic scleroderma clinic. *Value Health, 15*(1), 128-134. doi:10.1016/j.jval.2011.08.006

Marin, G., Sabogal, F., Marin, B. V., Otero-Sabogal, R., & Perez-Stable, E. J. (1987). Development of a short acculturation scale for Hispanics. *Hispanic Journal of Behavioral Sciences, 9*(2), 183-205. doi:110.1177/07399863870092005

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810-819. doi:10.1037/0022-3514.70.4.810

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132. doi: 10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Nunnally, J. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Oken, M. M., Creech, R. H., Tormey, D. C., Horton, J., Davis, T. E., McFadden, E. T., & Carbone, P. P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology, 5*(6), 649-655. doi:10.1097/00000421-198212000-00014

Palesh, O. G., Roscoe, J. A., Mustian, K. M., Roth, T., Savard, J., Ancoli-Israel, S., . . . Morrow, G. R. (2010). Prevalence, demographics, and psychological associations of sleep disruption in patients with cancer: University of Rochester Cancer Center-Community Clinical Oncology Program. *Journal of Clinical Oncology, 28*(2), 292-298. doi:10.1200/JCO.2009.22.5011

Reeve, B. B., Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., … Chen, W-H. (2016). Psychometric evaluation of the PROMIS® fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling*, *58(1)*, 119-139.

Savard, J., & Morin, C. M. (2001). Insomnia in the context of cancer: A review of a neglected problem. *Journal of Clinical Oncology, 19*(3), 895-908.

Savard, J., Simard, S., Blanchet, J., Ivers, H., & Morin, C. M. (2001). Prevalence, clinical characteristics, and risk factors for insomnia in the context of breast cancer. *Sleep, 24*(5), 583-590.

Teresi, J. A., & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, "Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) short forms". *Psychological Test and Assessment Modeling*, *58,* 37-78.

Tomlinson, D., Diorio, C., Beyene, J., & Sung, L. (2014). Effect of exercise on cancer-related fatigue: A meta-analysis. *American Journal of Physical Medicine & Rehabilitation, 93*(8), 675-686. doi:10.1097/PHM.0000000000000083

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PLoS ONE, 8*(9). doi:10.1371/journal.pone.0068967

Wood, A. M., Taylor, P. J., & Joseph, S. (2010). Does the CES-D measure a continuum from depression to happiness? Comparing substantive and artifactual models. *Psychiatry Reseach, 177*(1-2), 120-123. doi:10.1016/j.psychres.2010.02.003

Yu, C. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* (Doctoral Dissertation). University of California, Los Angeles, Los Angeles, California.

Yu, L., Buysse, D. J., Germain, A., Moul, D. E., Stover, A., Dodds, N. E., . . . Pilkonis, P. A. (2011). Development of short forms from the PROMIS sleep disturbance and sleep-related impairment item banks. *Behavioral Sleep Medicine, 10*(1), 6-24. doi:10.1080/15402002.2012.636266