

Published in final edited form as:

J Psychosom Res. 2014 April ; 76(4): 300–306. doi:10.1016/j.jpsychores.2014.01.010.

Ecological Validity and Clinical Utility of Patient-Reported Outcomes Measurement Information System (PROMIS®) instruments for detecting premenstrual symptoms of depression, anger, and fatigue

Doerte U. Junghaenel, PhD, Stefan Schneider, PhD, Arthur A. Stone, PhD, Christopher Christodoulou, PhD, and Joan E. Broderick, PhD

Department of Psychiatry and Behavioral Science, Stony Brook University

Abstract

Objective—This study examined the ecological validity and clinical utility of NIH Patient Reported-Outcomes Measurement Information System (PROMIS®) instruments for anger, depression, and fatigue in women with premenstrual symptoms.

Methods—One-hundred women completed daily diaries and weekly PROMIS assessments over 4 weeks. Weekly assessments were administered through Computerized Adaptive Testing (CAT). Weekly CATs and corresponding daily scores were compared to evaluate ecological validity. To test clinical utility, we examined if CATs could detect changes in symptom levels, if these changes mirrored those obtained from daily scores, and if CATs could identify clinically meaningful premenstrual symptom change.

Results—PROMIS CAT scores were higher in the pre-menstrual than the baseline ($ps < .0001$) and post-menstrual ($ps < .0001$) weeks. The correlations between CATs and aggregated daily scores ranged from .73 to .88 supporting ecological validity. Mean CAT scores showed systematic changes in accordance with the menstrual cycle and the magnitudes of the changes were similar to those obtained from the daily scores. Finally, Receiver Operating Characteristic (ROC) analyses demonstrated the ability of the CATs to discriminate between women with and without clinically meaningful premenstrual symptom change.

Conclusions—PROMIS CAT instruments for anger, depression, and fatigue demonstrated validity and utility in premenstrual symptom assessment. The results provide encouraging initial evidence of the utility of PROMIS instruments for the measurement of affective premenstrual symptoms.

© 2014 Elsevier Inc. All rights reserved.

Corresponding author: Doerte U. Junghaenel, PhD, Department of Psychiatry and Behavioral Science, Putnam Hall, South Campus, Stony Brook University, Stony Brook, NY 11794-8790, Phone: 631-632-9787, Fax: 631-632-3165, Doerte.Junghaenel@stonybrook.edu.

Competing interest statement: AAS is a Senior Scientist with the Gallup Organization and a Senior Consultant with ERT, inc. JEB has conflicts due to her relationship with AAS.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

PROMIS®; premenstrual symptoms; depression; fatigue; anger; validity

Introduction

Most women of reproductive age experience some emotional or physical discomfort during the late luteal phase of the menstrual cycle that remits shortly after the commencement of menses (e.g., 1-3). These symptoms are usually mild and do not interfere with life. A subset of women, however, experience premenstrual syndrome (PMS) that disrupts normal physical and psychosocial functioning (e.g., 2,4). In its most severe form, women (up to 8%) present with premenstrual dysphoric disorder (PMDD) where symptoms reach a **clinically significant** level (5-10). The personal and economic impact of PMS/PMDD is substantial (10,11), and proper clinical management has been somewhat challenging (1).

There is a lack of universally accepted diagnostic criteria and measures for clinical practice (1). The most defining characteristic of PMS/PMDD is the association between symptomatology and menstrual cyclicality (12). Over 65 prospective and retrospective assessment instruments exist (see e.g., 13), yet their reliability, validity and utility for diagnostic purposes vary. Prospective charting of daily symptoms for at least two consecutive menstrual cycles is considered the gold standard for assessment and diagnosis (for review see e.g., 10,14,15). Daily reports should reflect the onset of symptoms during the premenstrual phase and their reduction during the follicular phase (10). Consecutive daily reporting, however, poses various problems that have impeded consistent and effective use in medical practice. It is time-consuming and often impractical for women (16); not surprisingly, women may not complete the daily charting for more than one cycle (e.g., 17). Finally, treatment may not be initiated until the daily recording is completed, and women may not want to wait that long (12).

Another factor hampering the clinical utility of diary assessments is the broad inclusion of all possible symptoms; it reduces diagnostic precision because the status of individual premenstrual symptom components is not isolated (18). Health providers are often interested in targeting specific symptoms that are most bothersome to a specific woman; thus, aggregates of treatment change may obscure important treatment effects on specific dimensions (12).

The combination of burdensome daily reporting and aggregation of different symptoms have made accurate and practical diagnostic assessment difficult. Patients and clinicians would benefit from the availability of more convenient measures to assist with longitudinal monitoring and diagnosis that target core symptom domains. Ideally, over the course of 1-2 months, weekly measures that accurately reflect symptoms for that week could be administered. The measures would need to be sufficiently sensitive to change to note the affective fluctuations across the menstrual cycle.

Measures developed by the National Institutes of Health Patient-Reported Outcomes Measurement Information System® (PROMIS®) could potentially remedy some of these

issues. PROMIS is a NIH initiative to develop state-of-the-art self-report instruments for common medical symptoms (www.nihpromis.org). Its goal is to provide measures of patient-reported outcomes (PROs) that are brief, precise, optimally cover a wide range of symptom severity, and can be easily incorporated in research and clinical practice. Most of the PROMIS measures ask respondents to report over the previous week.

PROMIS developed domain item banks using item response theory (IRT), which yields a broad set of calibrated items to assess PRO domains with high reliability and measurement precision (19,20). They can be administered as static short forms or via Computerized Adaptive Testing (CAT), a sophisticated computer assessment methodology that allows for a precise assessment of a person's symptomatology. In CAT, a person is first presented with an item that targets moderate symptom severity. Based on the person's response, subsequent questions are tailored to tap higher or lower symptom levels. This allows for fast identification (usually 4-8 items) of where the person scores on the domain continuum (21).

The 7-day recall period and brevity of assessment makes PROMIS CATs appealing for clinical use; however, their validity and utility for the measurement and monitoring of premenstrual affective symptoms has not been established. In this report, we examine the ecological validity and clinical utility of PROMIS measures for detecting anger, depression, and fatigue in women presenting with premenstrual symptoms. These symptoms are among the cardinal features of premenstrual syndrome (e.g., 15,22). Ecological validity measures the degree to which scores reflect experience in the natural environment. Clinical utility addresses if the measure is sensitive to changes over the course of the menstrual cycle and can detect diagnostically relevant symptom elevations.

We assessed the ecological validity of the PROMIS CATs by comparing CAT scores (which use a seven-day recall period) with the average of scores collected using daily short-form versions of the PROMIS measures. To test clinical utility, we examined if CATs could detect changes in symptom levels, if these changes mirrored those obtained from daily scores, and if the CATs could identify NIH-defined clinically meaningful menstrual cycle symptom change.

Participants and Methods

Participants

This study was approved by the Stony Brook University Institutional Review Board and is part of a larger project investigating the ecological validity of PROMIS instruments across several community and clinical groups. Data were collected from women (N = 100) reporting premenstrual symptoms who were part of a national online research panel of 1.7 million survey respondents (www.surveyspot.com). Interested women were invited to contact our research office and were then telephone screened for eligibility. General inclusion criteria were age ≥ 21 years, fluency in English, availability to participate in the study for 29 to 36 days, and high-speed home Internet access.

PMS/PMDD-specific eligibility criteria were: no hysterectomy, regular monthly menses, not pregnant, no use of hormone replacement therapy or fertility drugs. Eligibility also required

a set of self-reported physical and emotional/behavioral symptoms with onset prior to menses and remission by the end of menses (see e.g. 15,23). In addition, two or more of the following physical symptoms were necessary: abdominal bloating, weight gain due to water retention, increased appetite/food cravings, breast pain/tenderness, acne flare-ups, hot flashes, headache, constipation, dizziness, diarrhea, poor coordination, or change in sex drive. One or more of the following emotional/behavioral symptoms also needed to be present: irritability/angry outbursts, mood swings or depressed mood, poor impulse control, tension/anxiety, lethargy/sluggishness/fatigue, crying, social withdrawal, insomnia, thirst, or difficulty concentrating or thinking clearly. Symptoms needed to be rated by the respondent as at least moderately severe and needed to have interfered with the woman's daily life. Eligible women were scheduled to start assessments approximately two weeks prior to the estimated start of their next menses, because symptom levels are expected to change during the week before and after menses onset (23).

Data collection

All interactions with study participants were conducted over the phone and computer. Participants completed the symptom ratings at home by logging onto the PROMIS Assessment CenterSM (<http://www.assessmentcenter.net/>), a free online data collection tool. Research staff coached the participants over the phone in how to provide electronic informed consent and in use of the Assessment Center. Starting the following day, participants completed daily short forms between 6 PM and midnight for each of the next 28 consecutive days. At the end of each week (on days 7, 14, 21, and 28), the PROMIS 7-day recall CATs were administered in addition and prior to the daily short forms for that day. Compliance was monitored daily, and participants were contacted if they missed an assessment. Participants were compensated up to \$150 for study completion. To facilitate compliance, a lottery for another \$150 was conducted after every 25th study completer; the lottery was open only to those participants who completed every assessment.

Measures

PROMIS CATs and corresponding daily measures—Three PROMIS symptom domains (see <http://www.nihpromis.org/measures/domainframework1>) were included in the study: 1) the anger item bank assesses angry mood, negative social cognitions, verbal aggression, and efforts to control anger; 2) the depression item bank assesses negative mood, negative views of the self, negative social cognition, and decreased positive affect and engagement; 3) the fatigue item bank consists of symptoms that range from mild subjective feelings of tiredness to an overwhelming, debilitating, and sustained sense of exhaustion. The bank taps into the experience of fatigue (frequency, duration, and intensity) and the impact of fatigue on physical, mental, and social activities.

For the purpose of daily assessment, scales consisting of 8 (anger), 8 (depression), and 7 (fatigue) items from the PROMIS banks were selected consistent with the creation of PROMIS Version 1 short-forms (Cella et al., 2010). The recall period of each item was modified from “In the past 7 days...” to “In the last day...”. Scale scores for the daily measures were obtained using IRT and employing the national item parameters established

for PROMIS (<http://www.nihpromis.org>). This scoring placed the daily scores on the PROMIS T-score metric (24).

At the end of each of the four weeks, these 3 domains were assessed with PROMIS CATs that were set to administer no less than 4 and no more than 12 items and to terminate when $> .90$ score reliability ($SE < 3$ T-score points) was achieved. The scores are reported on a T-score metric (mean=50; standard deviation=10) that is anchored to the distribution of scores in the U.S. general population (25,26).

Analysis strategy

Descriptive statistics on the reliabilities and number of items administered by the CAT were examined first. The goal of CAT administration is to achieve high reliability for a respondent's score with a small number of items. The obtained reliabilities can vary between respondents and assessments, and were calculated based on the IRT-based standard errors (SE) of the T-scores, where $reliability = 1 - (SE/10)^2$ (27).

Next, the ecological validity of the PROMIS CATs was addressed by examining (a) differences in mean symptom levels on CAT and daily short forms and (b) correlations between CATs and daily scores for each week. The 7 days of daily scores that corresponded to the 7 days covered by the CAT were compared for each week in the month-long study.

Then, we examined the clinical utility of weekly CAT scores. First, we determined if weekly CAT scores could detect symptom change across the menstrual cycle. For this purpose, the CATs were assigned to the weeks of the cycle using the self-reported first day of menses (as reported in the diaries) as the reference day. Specifically, counting backwards from the first day of menses, the recall period of the CAT that covered the majority (at least 4 of 7) of days before this day was coded as “pre-menstrual,” and the 4 weekly CATs were categorized accordingly as “baseline”, “pre-menstrual”, “menstrual”, and “post-menstrual”. To test the hypothesis that symptoms would be elevated during the pre-menstrual week, repeated measures ANOVA with *a priori* contrasts was used to test for mean differences between CAT scores for the pre-menstrual week in comparison with CAT scores for the baseline, menstrual, and post-menstrual weeks, respectively.

Second, we examined whether changes in the weekly CAT scores adequately paralleled those obtained from daily scores. For this comparison, the daily scores were averaged using the first day of menses as reference (i.e., day 0) to create aggregated daily scores for the baseline (days -14 to -8), pre-menstrual (days -7 to -1), menstrual (days 0 to 6) and post-menstrual (days 7 to 13) week. Note that the days aggregated for each of these weeks corresponds precisely to the menstrual cycle as a menstrual diary would, thus they were not completely aligned with the seven days covered by the corresponding CAT. For example, whereas the CAT was programmed to be administered on day 14 for all participants, for women starting menses on days 12, 13, 14, and 15 of the study, the days of the pre-menstrual week were days 5-11, 6-12, 7-13, and 8-14, respectively. Differences in change between CAT and averaged daily scores were tested using repeated measures ANOVA with Time (menstrual cycle week) and Method (daily diary vs. CAT) as two within-person factors, and by testing the Time \times Method interaction term; *a priori* contrasts were again

used to compare whether the methods differed in change from the pre-menstrual to each of the remaining weeks.

Finally, we examined the ability of the weekly CATs to discriminate between women who did and did not demonstrate clinically meaningful premenstrual symptom change on the daily assessments for each PRO domain. For each woman, we calculated the percent change in daily scores from the late luteal (the 6 days before menses) to the follicular (days 5–10 of the cycle).¹ Following the National Institutes of Mental Health definition (29), a 30% symptom increase in the luteal phase was considered clinically meaningful, and women were categorized accordingly into groups with and without meaningful change. Receiver Operating Characteristic (ROC) analyses were then used to examine the extent to which changes in CAT scores from the pre- to the post-menstrual week accurately discriminated between the groups. ROC curves depict the true positive (sensitivity) and false positive (1-specificity) rates for a series of incremental cut-offs in CAT change scores. The area under the ROC curve (AUC) was used to summarize classification accuracy (with values ranging from 0.5 = chance level of discrimination to 1.0 = perfect discrimination). Optimal cut-offs for clinically meaningful changes in CATs were determined by identifying the point with the highest sensitivity and specificity in the ROC analysis (30).

Handling of missing data

Multiple imputation was used to account for missing data resulting from participant noncompliance on some days and from variation between participants in the onset of menses. A set of 20 multiple imputed datasets was generated using Markov chain Monte Carlo estimation, and the results carried out for each dataset were combined using Rubin's rules to adjust the test statistics for the uncertainty about imputed values (31,32). Analyses were performed using Mplus Version 7 (33). The logistic procedure in SAS (version 9.3; Cary, NC) was used for the ROC analyses.

Results

Participant characteristics and compliance

Out of the 100 enrolled women, 7 were excluded from the analyses (5 dropped out and 2 reported the onset of menses on day 28). Table 1 shows demographic characteristics of the analyzed sample ($n = 93$). The average age was 36 years, 73% of the women were White, 15% Hispanic, 13% African American, and about half (52%) were married. Educational status ranged from high school (14%), some college (41%), college (32%), and had an advanced degree (13%). Women who were excluded were somewhat younger (mean = 33 years, $p = .23$), but otherwise did not differ from the analyzed sample.

Women in the analysis sample completed an average of 26.8 ($SD = 1.73$, Median = 27) out of 28 daily assessments; a total of 111 (4.3%) out of 2,604 assessment days were missed. Out of 372 weekly CAT assessments, only 3 were missed.²

¹The minimum possible T-score for each daily short form measure (equaling a raw score of 0) was used as zero point for calculating percent increase (28).

Measurement properties of CATs

Table 2 shows the number of items and reliabilities (based on the standard error of measurement) of the CATs. The CAT administered an average of 6.2 (anger), 5.3 (depression), and 4.2 (fatigue) items, with average reliabilities .91 for each domain. Reliabilities exceeded a level of .90 for 95% (anger), 93% (depression) and 100% (fatigue) of the assessments, and did not fall below .73 for any individual assessment.

We then evaluated the ecological validity of PROMIS CATs by examining symptom level differences and correlations between CATs and aggregated daily scores for each week of the cycle. Symptom levels on daily scores were significantly lower ($p < .0001$) than the corresponding PROMIS CATs for each PRO domain (Table 3). As shown in Table 4, the correlations ranged from .73 to .88 ($p < .0001$), suggesting moderate to high correspondence between the two assessment methods.

Mean CAT changes across the menstrual cycle

The clinical utility of changes in CAT scores was examined next. The mean CAT T-scores were not significantly different ($p > .12$) from the PROMIS U.S. population average normative score of 50 during the baseline and post-menstrual weeks (see Figure 1). For all domains, the CAT T-scores for the pre-menstrual week were higher than the baseline week ($p < .0001$) and the post-menstrual ($p < .0001$) week, indicating sensitivity of the weekly CATs to detect pre-menstrual symptom change. CAT scores for the menstrual week were somewhat elevated compared to the pre-menstrual week ($p < .05$ for anger and fatigue; $p = .14$ for depression).

Comparison of changes based on CAT and daily scores

The pattern of weekly changes was very similar between CATs and averaged daily scores. Change scores were computed between the premenstrual week and each of the other weeks. Differences in magnitude of change between the assessment methods reached significance in 3 of 9 instances (Table 5); those differences were small (not exceeding 1.7 T-scores), suggesting that the CAT scores successfully reflect the aggregated daily reports.

ROC analyses

Finally, we examined the ability of the CATs to discriminate between women who did and did not have clinically meaningful pre-menstrual symptom change, defined as 30% or more change between the luteal and follicular phase on daily diary reports. This criterion was met by 63% of the women for anger, by 62% for depression, and by 57% for fatigue. The ROC curves in Figure 2 show the sensitivity and specificity of pre- to post-menstrual changes in CAT scores to correctly identify women who did and did not meet this criterion. The area under the curve (AUC) values were .83 for anger, .80 for depression, and .74 for fatigue, indicating moderate to high accuracy of the CATs. T-score cut-offs for CAT changes with

²The full menstrual cycle (days -14 to +13 from menses onset) could only be represented for women who had the first day of menses on day 15 of the study. This was the case for 20% of the sample, with 83% having the first day of menses between days 13 and 19 of the study. Coding the assessment days relative to the days of the menstrual cycle slightly increased the rate of missing values because assessments completed more than 2 weeks before/after menses onset could not be included in the analyses: the resulting rates of missing values were 11% for daily short forms and 5% for CATs.

optimal sensitivity and specificity were 5.5 for anger (sensitivity = 77%, specificity = 77%), 3.0 for depression (sensitivity = 72%, specificity = 70%), and 3.5 for fatigue (sensitivity = 78%, specificity = 65%).

Discussion

The results of the present study supported the ecological validity and clinical utility of the newly developed PROMIS CAT measures to assess anger, depression, and fatigue in women reporting premenstrual symptoms. Correlations between CATs and aggregated daily scores ranged from .73 to .88 and mean CAT scores showed systematic changes in accordance with the menstrual cycle. Receiver Operating Characteristic (ROC) analyses demonstrated the ability of the CATs to discriminate between women with and without clinically meaningful premenstrual symptom change.

Results for the measurement properties of the CATs across 28 days were very encouraging. In this study, reliabilities of $>.90$ were achieved in most instances using only 4 to 7 items per domain. This finding adds to the clinical utility of the PROMIS measures, since the brevity of assessment does not compromise psychometric properties.

Ecological validity measures the degree to which scores reflect experience in the natural environment. Retrospective instruments can be prone to measurement bias through cognitive heuristics and memory issues (34). Studies in PMS/PMDD have also demonstrated recall bias and amplification of symptoms in retrospective reports (e.g., 35,36). Daily assessment reduces recall bias (34), and is, therefore, considered the gold standard for the premenstrual symptom assessment and diagnosis (e.g., 10,14,15). However, due to the substantial reporting burden, daily reporting across two months often is not a viable and cost-effective strategy in clinical practice.

Examination of the correlations between CAT scores and aggregated daily short forms for each week supported the ecological validity of the PROMIS measures. Results showed moderate to high correspondence between the two assessment methods. The correlations are typical of correspondence observed between daily and weekly methods (e.g., 37,38).

The levels of daily scores were lower (lower symptom level) than the corresponding CATs for all three domains. This finding has been consistently demonstrated in previous literature where shorter recall periods produce lower symptom levels, even when the same items and response scales are used (38-41). This level difference has sometimes been interpreted as bias inherent in weekly recall measures (42). For example, differences in symptom levels can vary based on the length of the recall period; an important consideration when evaluating change due to treatment (40). We have also found that temporal trends in symptoms can influence the accuracy of recall reports in related work on the PROMIS measures (43). Nevertheless, PROMIS measures have the advantage of being calibrated relative to the U.S. general population; therefore, normative comparisons are possible and valid (20).

Changes in CAT scores across the menstrual cycle supported the clinical utility, specifically the sensitivity of the CATs to detect significant symptom change. For each domain, the CAT

scores for the pre-menstrual week were significantly higher than the baseline and the post-menstrual week. Indeed, the mean CAT scores for the baseline and the post-menstrual weeks were not significantly different than the U.S. general population average norm of a T-score of 50. For depression and anger, in particular, the CAT scores for the menstrual week were somewhat elevated compared to the aggregated daily ratings. This could in part be explained by the fact that the CAT of the menstrual week included some (up to three) pre-menstrual days for some participants depending upon when their menses started during the study.

Examination of the comparability of change observed from week to week in CAT scores versus daily scores also confirmed the CAT clinical utility. The magnitude of changes across the weeks of cycle was very similar between CATs and weekly averages of daily scores, indicating that the sensitivity to change is as pronounced for PROMIS 7-day recall CATs as for (aggregated) daily measures.

Importantly, we examined the extent to which CAT scores could detect clinically meaningful symptom change for PMS as defined by the National Institutes of Mental Health definition of 30% change from the 6 days prior to menses onset to days 5-10 after menses onset (29). For all three domains, approximately 60% of the women met this criterion based on the daily scores, and ROC analyses suggested moderate to high discrimination of the CATs. Cut-offs with optimal sensitivity and specificity for PROMIS CAT changes ranged from 3.0 to 5.5 T-scores. A previous study has identified similar cut-points for minimally important differences of PROMIS measures; however, this study was specific to cancer patients (44). Our results suggest that cut-offs for meaningful change of PROMIS measures may generalize across patient populations, which can facilitate clinical interpretation of the PROMIS T-score metric.

Several limitations should be noted. First, female participants were recruited from a national Internet panel. Participants with very low education were not well represented. (45). Second, the sample was comprised of women who self-reported premenstrual symptoms; the results may, therefore, not fully generalize to women with a health provider diagnosis of PMS/PMDD. It is also important to note that context effects (47), in this case, ratings obtained in healthcare settings versus ratings obtained in a patient's home can impact the reporting of premenstrual symptoms. However, in this instance, daily diaries completed at home by women have been the standard for diagnostic evaluation. Thus, the context of ratings in this study is comparable to the situation for most PMS/PMDD assessments. Indeed, our sample may have included women with less severe PMS/PMDD than is generally seen in the clinic, thus resulting in conservative estimates (a potential strength of the study). Third, this study did not include a comparison group of women without PMS/PMDD. Thus, this report is unable to demonstrate that the changes in the three domains distinguish women with and without PMS/PMDD. Fourth, this study only assessed one month, the PROMIS daily measures have not previously been employed in the premenstrual symptoms literature, and the PRO measures only focused on three domains. As such, our methodology, while promising as an initial screening tool or way to measure treatment response, cannot be considered a substitute for standard medical procedures and criteria to diagnose PMS/PMDD. A diagnosis requires multiple symptoms of sufficient severity whose ebb and flow are entrained in the menstrual cycle as evident in prospective charting of at least two

consecutive cycles (for review see 15). Subsequent work could examine other important PMS/PMDD domains with PROMIS measures, for example, anxiety, sleep disturbance, pain intensity and pain interference and their relationship with women's social and occupational functioning.

Clinical researchers may find use of four weekly PROMIS CATs of affective symptoms to be a useful screening tool. These results suggest that the clinical application of PROMIS CATs provides not only brief, but also precise and ecologically valid information about affective symptom changes in individual patients. The emergence of patient portals in electronic medical record systems creates an opportunity for convenient patient symptom monitoring and review by the clinical team (47). In the absence of a standardized research protocol, reminders, for example through automated emails, to complete the weekly assessments should be developed between the patient and the clinician. This is particularly important when the administration of weekly PROMIS measures as a screen precedes consecutive daily reporting in that a routine of daily ratings could facilitate compliance with weekly assessments. In sum, PROMIS CAT instruments for anger, depression, and fatigue demonstrated ecological validity and sensitivity to change in the assessment of premenstrual symptoms. The results provide encouraging initial evidence of the utility of these instruments for clinical and research outcomes. Subsequent research could broaden the number of symptoms assessed, increase the length of assessment, and engage women with and without PMS diagnosis. If successful, women and their providers may have a much easier method for quantifying the pattern of affective symptoms over time for diagnosis and treatment outcome.

Acknowledgments

This research was supported by a grant from the National Institutes of Health (1 U01AR057948-01). We thank our participants and our research assistants, Laura Wolff, Gim Yen Toh, and Lauren Cody, for their assistance in completing the study. PROMIS® was funded with cooperative agreements from the National Institutes of Health (NIH) Common Fund Initiative (U54AR057951, U01AR052177, U54AR057943, U54AR057926, U01AR057948, U01AR052170, U01AR057954, U01AR052171, U01AR052181, U01AR057956, U01AR052158, U01AR057929, U01AR057936, U01AR052155, U01AR057971, U01AR057940, U01AR057967, and U01AR052186). The contents of this article use data developed under PROMIS. These contents do not necessarily represent an endorsement by the US Federal Government or PROMIS. See www.nihpromis.org for additional information on the PROMIS initiative.

Grant support: NIH/NIAMS 1U01AR057948

Financial support: None

References

1. Halbreich U, Backstrom T, Eriksson E, O'Brien S, Calil H, Ceskova E, Dennerstein L, Douki S, Freeman E, Genazzani A, Heuser I, Kadri N, Rapkin A, Steiner M, Wittchen HU, Yonkers K. Clinical diagnostic criteria for premenstrual syndrome and guidelines for their quantification for research studies. *Gynecol Endocrinol*. 2007; 23:123–30. [PubMed: 17454164]
2. Halbreich U. The diagnosis of premenstrual syndromes and premenstrual dysphoric disorder--clinical procedures and research perspectives. *Gynecol Endocrinol*. 2004; 19:320–34. [PubMed: 15724807]
3. Sveindottir H, Backstrom T. Prevalence of menstrual cycle symptom cyclicality and premenstrual dysphoric disorder in a random sample of women using and not using oral contraceptives. *Acta Obstet Gynecol Scand*. 2000; 79:405–13. [PubMed: 10830769]

4. Yonkers KA, Pearlstein T, Rosenheck RA. Premenstrual disorders: bridging research and clinical reality. *Arch Womens Ment Health*. 2003; 6:287–92. [PubMed: 14628181]
5. Pearlstein T, Yonkers KA, Fayyad R, Gillespie JA. Pretreatment pattern of symptom expression in premenstrual dysphoric disorder. *J Affect Disord*. 2005; 85:275–82. [PubMed: 15780697]
6. Wittchen HU, Becker E, Lieb R, Krause P. Prevalence, incidence and stability of premenstrual dysphoric disorder in the community. *Psychol Med*. 2002; 32:119–32. [PubMed: 11883723]
7. Angst J, Sellaro R, Merikangas KR, Endicott J. The epidemiology of perimenstrual psychological symptoms. *Acta Psychiatr Scand*. 2001; 104:110–6. [PubMed: 11473504]
8. Merikangas KR, Foeldenyi M, Angst J. The Zurich Study XIX. Patterns of menstrual disturbances in the community: results of the Zurich Cohort Study. *Eur Arch Psychiatry Clin Neurosci*. 1993; 243:23–32. [PubMed: 8399407]
9. Ramcharan S, Love EJ, Fick GH, Goldfien A. The epidemiology of premenstrual symptoms in a population-based sample of 2650 urban women: attributable risk and risk factors. *J Clin Epidemiol*. 1992; 45:377–92. [PubMed: 1569434]
10. Pearlstein T, Steiner M. Premenstrual dysphoric disorder: burden of illness and treatment update. *J Psychiatry Neurosci*. 2008; 33:291–301. [PubMed: 18592027]
11. Borenstein JE, Dean BB, Endicott J, Wong J, Brown C, Dickerson V, Yonkers KA. Health and economic impact of the premenstrual syndrome. *J Reprod Med*. 2003; 48:515–24. [PubMed: 12953326]
12. Yonkers KA, O'Brien PM, Eriksson E. Premenstrual syndrome. *Lancet*. 2008; 371:1200–10. [PubMed: 18395582]
13. Budeiri DJ, LiWan PA, Dornan JC. Clinical trials of treatment of premenstrual syndrome: entry criteria and scales for measuring treatment outcomes. *British J Obstet Gynaecol*. 1994; 101:689–695.
14. Endicott J, Amsterdam J, Eriksson E, Frank E, Ffreeman E, Hirschfeld R, Ling F, Parry B, Pearlstein T, Rosenbaum J, Rubinow D, Schmidt P, Severino S, Steiner M, Stewart DE, Thys-Jacobs S. Is premenstrual dysphoric disorder a distinct clinical entity? *J Womens Health Gend Based Med*. 1999; 8:663–79. [PubMed: 10839653]
15. Freeman EW. Premenstrual syndrome and premenstrual dysphoric disorder: definitions and diagnosis. *Psychoneuroendocrinol*. 2003; 28(Suppl 3):25–37.
16. Salamat S, Ismail KMK, O'Brien S. Premenstrual syndrome. *Obstet Gynaecol Reprod Med*. 2008; 18:29–32.
17. Sternfeld B, Swindle R, Chawla A, Long S, Kennedy S. Severity of premenstrual symptoms in a health maintenance organization population. *Obstet Gynecol*. 2002; 99:1014–24. [PubMed: 12052592]
18. Freeman EW, Halberstadt SM, Rickels K, Legler JM, Lin H, Sammel MD. Core symptoms that discriminate premenstrual syndrome. *J Womens Health (Larchmt)*. 2011; 20:29–35. [PubMed: 21128818]
19. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu HH, Gershon R, Reise SP, Lai JS, Cella D. 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:S22–S31. [PubMed: 17443115]
20. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, Dewalt D, Fries JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. 2010; 63:1179–94. [PubMed: 20685078]
21. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007; 16:133–41. [PubMed: 17401637]
22. Freeman EW, Derubeis RJ, Rickels K. Reliability and validity of a daily diary for premenstrual syndrome. *Psychiatry Res*. 1996; 65:97–106. [PubMed: 9122290]

23. Dickerson LM, Mazyck PJ, Hunter MH. Premenstrual syndrome. *Am Fam Physician*. 2003; 67:1743–52. [PubMed: 12725453]
24. Schneider S, Choi SW, Junghaenel DU, Schwartz JE, Stone AA. Psychometric characteristics of daily diaries for the Patient-Reported Outcomes Measurement Information System (PROMIS®): a preliminary investigation. *Qual Life Res*. in press.
25. Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). *Arthritis Care Res (Hoboken)*. 2011; 63:S486–90.
26. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. *J Clin Epidemiol*. 2010; 63:1169–78. [PubMed: 20688473]
27. Thissen, D. Reliability and measurement precision. In: Wainer, H., editor. *Computerized Adaptive Testing: A Primer*. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
28. Cohen P, Cohen J, Aiken LS, West SG. The problem of units and the circumstance for POMP. *Multivariate Behav Res*. 1999; 34:315–346.
29. Connolly M. Premenstrual syndrome: an update on definitions, diagnosis and management. *Adv Psychiatr Treat*. 2001; 7:469–477.
30. Zweig MH, Campbell G. Receiver-Operating Characteristic (ROC) plots - a fundamental evaluation tool in clinical medicine. *Clinical Chem*. 1993; 39:561–577. [PubMed: 8472349]
31. Rubin, DB. *Multiple imputation for Nonresponse in Surveys*. New York, NY: J. Wiley & Sons; 1987.
32. Schafer, JL. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall; 1997.
33. Muthén, LK.; Muthén, BO. *Mplus user's guide*. 7th. Los Angeles, CA: Muthén & Muthén; 1998-2012.
34. Stone, A.; Shiffman, SS. Ecological validity for patient reported outcomes. In: Steptoe, A., editor. *Handbook of Behavioral Medicine: Methods and Applications*. New York: Springer; 2010.
35. Marvan ML. Women's beliefs about the prevalence of premenstrual syndrome and biases in recall of premenstrual changes. *Health Psych*. 2001; 20:276–280.
36. Endicott J, Halbreich U. Retrospective report of premenstrual depressive changes: factors affecting confirmation by daily ratings. *Psychopharmacol Bull*. 1982; 18:109–112.
37. Broderick JE, Junghaenel DU, Schneider S, Piloni J, Stone AA. Pittsburgh and Epworth Sleep Scale Items: accuracy of ratings across different reporting periods. *Behav Sleep Med*. 2013b; 11:173–88. [PubMed: 23205491]
38. Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. *Pain*. 2008; 139:146–157. [PubMed: 18455312]
39. Stone A, Broderick J, Shiffman S, Schwartz J. Understanding recall of weekly pain from a momentary assessment perspective: Absolute accuracy, between- and within-person consistency, and judged change in weekly pain. *Pain*. 2004; 107:61–69. [PubMed: 14715390]
40. Stone AA, Schwartz JE, Broderick JE, Shiffman S. Variability of momentary pain predicts recall of weekly pain: A consequence of the peak (or salience) memory heuristic. *Pers Soc Psych Bull*. 2005; 31:1340–1346.
41. Keller SD, Bayliss MS, Ware JE Jr, Hsu MA, Damiano MA, Goss TF. Comparison of responses to SF-36 Health Survey questions with one-week and four-week recall periods. *Health Serv Res*. 1997; 32:367–384. [PubMed: 9240286]
42. Redelmeier DA, Kahneman D. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain*. 1996; 66:3–8. [PubMed: 8857625]
43. Schneider S, Broderick JE, Junghaenel DU, Schwartz JE, Stone AA. Temporal trends in symptom experience predict the accuracy of recall PROs. *J Psychosom Res*. 2013; 75:160–166. [PubMed: 23915773]
44. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiol*. 2011; 64:507–516.

45. Zickuhr, K.; Smith, A. Digital Differences. Washington, DC: Pew Research Center; 2012. Online Available <http://pewinternet.org/Reports/2012/Digital-differences.aspx>
46. Schwarz, N. Retrospective and concurrent self-reports: the rationale for real-time data capture. In: Stone, AA.; Shiffman, SS.; Atienza, A.; Nebeling, L., editors. The science of realtime data capture: self-reports in health research. New York: Oxford University Press; 2007. p. 11-26.
47. Broderick JE, Dewitt EM, Rothrock N, Crane P, Forrest C. Advances in patient reported outcomes: NIH PROMIS measures. eGEMS (Generating Evidence & Methods to improve patient outcomes). 2013a; 1(1) Available at: <http://repository.academyhealth.org/egems/vol1/iss1/12>.

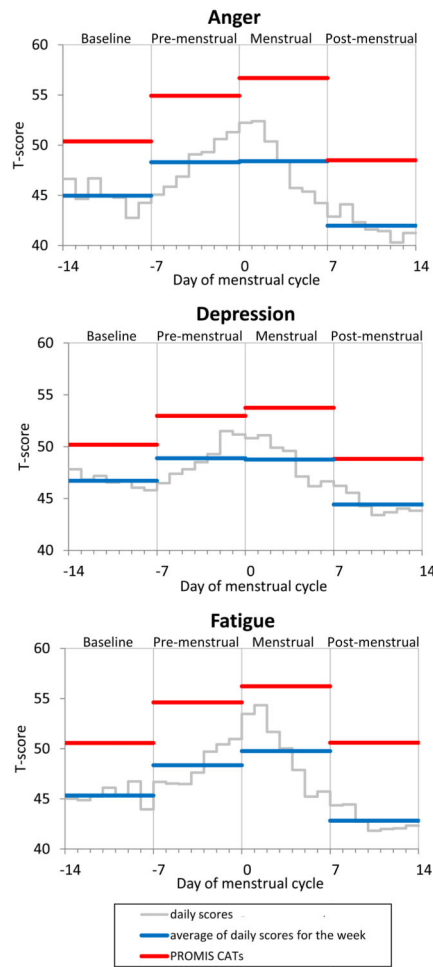


Figure 1. Daily and CAT scores over the course of the menstrual cycle for each PRO domain. Grey lines represent the mean daily scores for each day. Blue lines represent the mean of daily averages for each week. Red lines represent the mean CAT scores for each week. Day 0 is the day of menses onset.

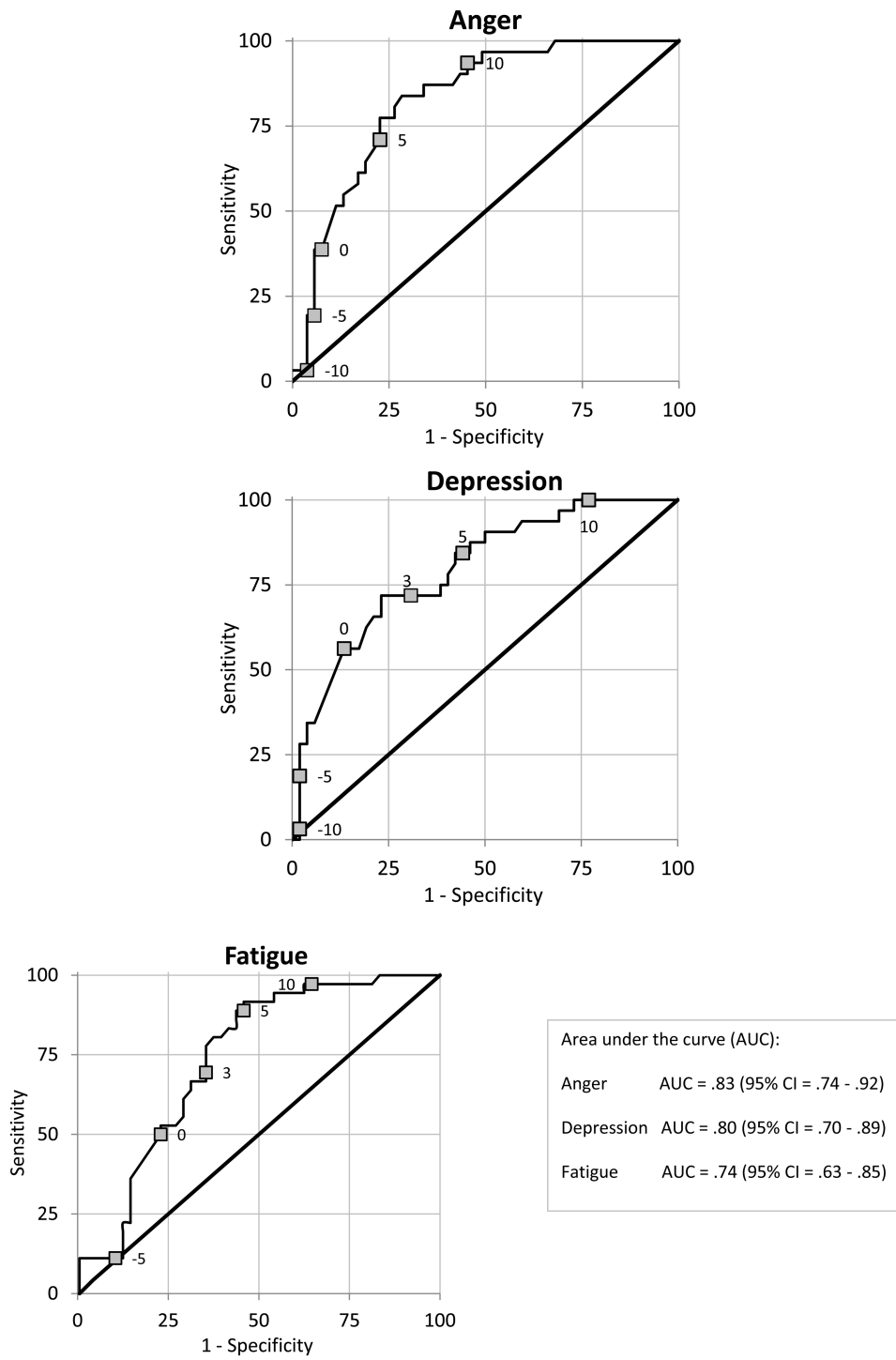


Figure 2. Receiver Operating Characteristic (ROC) curves for the prediction of clinically relevant premenstrual symptom increase from change in CAT scores for each PRO domain. Labeled points on each curve represent selected changes in CAT scores, where positive values indicate higher scores in the pre-menstrual compared to the post-menstrual week.

Table 1
Demographic characteristics of study participants ($n = 93$)

	Frequency (%)
Age (mean = 36.4, $SD = 7.9$)	
Age categories	
21 - 30	24 (25.8)
31 - 40	38 (40.8)
41 - 54	31 (33.3)
Race	
White	68 (73.1)
African American	12 (12.9)
Asian	5 (5.4)
Other/multiple	8 (8.6)
Hispanic	14 (15.1)
Married	52 (55.9)
Education	
High school	13 (14.0)
Some college	38 (40.9)
College graduate	30 (32.3)
Advanced degree	12 (12.9)
Family income ^a	
less than \$20,000	8 (8.7)
\$20,000 - 34,999	19 (20.7)
\$35,000 - 49,999	24 (26.1)
\$50,000 - 74,999	20 (21.7)
\$75,000 and higher	21 (22.8)

Note:

^aIncome not reported by one participant.

Table 2

Characteristics of Computerized Adaptive Testing (CAT) scores: reliability and number of items administered.

	Anger	Depression	Fatigue
Reliability of individual CAT scores			
Mean (<i>SD</i>)	.91 (0.03)	.93 (0.05)	.94 (0.01)
Median	.92	.94	.95
Range	.74 - .93	.73 - .96	.91 - .96
Number of CAT items administered			
Mean (<i>SD</i>)	6.24 (1.90)	5.25 (2.44)	4.17 (0.77)
Median	5.00	4.00	4.00
Range	5 - 12	4 - 12	4 - 9

Note: Reliabilities were derived from IRT-based standard errors (SE) for individual T-scores, calculated as reliability = $1 - (SE/10)^2$.

Table 3
Means (standard errors) and mean differences between PROMIS CATs and aggregated daily short forms by week of the menstrual cycle

	CAT mean	Mean of daily scores	Mean difference (CAT - daily)
<i>Anger</i>			
Baseline week	50.5 (.87)	45.2 (.81)	5.3 (.50)
Pre-menstrual week	54.9 (.81)	48.5 (.87)	6.4 (.55)
Menstrual week	56.7 (.88)	48.3 (.94)	8.4 (.62)
Post-menstrual week	48.3 (.99)	41.7 (.87)	6.6 (.72)
Average (all weeks)	52.6 (.72)	45.9 (.74)	6.7 (.45)
<i>Depression</i>			
Baseline week	50.2 (.84)	46.6 (.80)	3.5 (.45)
Pre-menstrual week	53.0 (.80)	49.0 (.85)	4.0 (.42)
Menstrual week	53.8 (.87)	48.9 (.89)	4.8 (.44)
Post-menstrual week	48.9 (.85)	44.3 (.80)	4.6 (.52)
Average (all weeks)	51.4 (.72)	47.2 (.75)	4.6 (.32)
<i>Fatigue</i>			
Baseline week	50.5 (.90)	46.1 (.91)	4.5 (.52)
Pre-menstrual week	54.6 (.80)	48.7 (1.00)	6.0 (.52)
Menstrual week	56.2 (.81)	49.6 (.95)	6.7 (.65)
Post-menstrual week	50.3 (.80)	42.3 (1.00)	8.0 (.63)
Average (all weeks)	52.9 (.67)	46.6 (.85)	6.3 (.40)

Note: All mean differences are significant at $p < .0001$.

Table 4
Correlations between PROMIS CATs and aggregated daily short forms by week of the menstrual cycle

	Correlations				
	Baseline week	Pre-menstrual week	Menstrual week	Post-menstrual week	Pooled across weeks
Anger	.83	.79	.78	.73	.78
Depression	.85	.87	.88	.81	.85
Fatigue	.83	.86	.74	.81	.81

Note: All correlations are significant at $p < .0001$.

Table 5
Measurement method differences (standard error) in change scores between PROMIS
aggregated daily scores and CATs

	Method difference: Change from pre-menstrual week (CAT minus aggregated daily)
Anger	
Baseline week	-1.22 (.61)*
Menstrual week	+1.67 (.70)*
Post-menstrual week	-0.11 (.85)
Depression	
Baseline week	-0.62 (.56)
Menstrual week	+0.89 (.52)
Post-menstrual week	+0.30 (.62)
Fatigue	
Baseline week	-1.02 (.73)
Menstrual week	+0.21 (.59)
Post-menstrual week	+1.52 (.71)*

Note:

* $p < .05$;