Validity and Reliability of Patient-Reported Outcomes Measurement Information System Instruments in Osteoarthritis

JOAN E. BRODERICK, STEFAN SCHNEIDER, DOERTE U. JUNGHAENEL, JOSEPH E. SCHWARTZ, AND ARTHUR A. STONE


Methods. We recruited an OA sample and a comparison general population (GP) sample through an internet survey panel. Pain intensity, pain interference, physical functioning, and fatigue were assessed for 4 consecutive weeks with PROMIS short forms on a daily basis and compared with same-domain Computer Adaptive Testing (CAT) instruments that use a 7-day recall. Known-group validity (comparison of OA and GP), ecological validity (comparison of aggregated daily measures with CAT instruments), and test–retest reliability were evaluated.

Results. The recruited samples matched the demographic characteristics (age, sex, race, and ethnicity) of the US sample for arthritis and the 2009 Census for the GP. Compliance with repeated measurements was excellent (>95%). Known-group validity for CATs was demonstrated with large effect sizes (pain intensity 1.42, pain interference 1.25, and fatigue 0.85). Ecological validity was also established through high correlations between aggregated daily measures and weekly CATs (≥0.86). Test–retest validity (7-day) was very good (≥0.80).

Conclusion. PROMIS CAT instruments demonstrated known-group and ecological validity in a comparison of OA patients with a GP sample. Adequate test–retest reliability was also observed. These data provide encouraging initial data on the utility of these PROMIS instruments for clinical and research outcomes in OA patients.

INTRODUCTION

The Patient-Reported Outcomes Measurement Information System (PROMIS), a National Institutes of Health–directed initiative, has developed self-report measures for a variety of health experiences (www.nihpromis.org). They were developed using modern psychometric techniques in order to achieve optimally precise, yet relatively brief, measures. Specifically, item banks were developed using item response theory, which yields a comprehensive set of calibrated items to assess each patient-reported outcome (PRO) domain (1,2). An important characteristic of PROMIS item banks is their systematic coverage of very low through very high levels of the measured experience (3). As a result, they have demonstrated high reliability and measurement precision (2,4,5). For each PRO domain, measures can be administered via computerized adaptive testing (CAT), or by selecting any subset of items from the bank for use as a static short form (SF), including available sets of SFs (2). CAT is a state-of-the-art measurement methodology that enables measurement precision with presentation of very few items (6). Each respondent is initially presented with an item tapping the midrange of the latent trait. Subsequent questions address higher or lower trait levels depending upon the person’s responses to the preceding items. This allows for rapid identification of the respondent’s placement on the domain continuum and scale score (7). The brevity and ease of measurement make CAT attractive not only for assessing endpoints in clinical trials, but also for monitoring an individual’s patient status in clinical care. The potential advantages of PROMIS over traditional instruments specifically for measuring PROs in rheumatology patients have been described previously (8).

Supported by the NIH/National Institute of Arthritis and Musculoskeletal and Skin Diseases (grant 01-AR-057948). Joan E. Broderick, PhD, Stefan Schneider, PhD, Doerte U. Junghaenel, PhD, Joseph E. Schwartz, PhD, Arthur A. Stone, PhD: Stony Brook University, Stony Brook, New York.

Dr. Stone is a senior scientist with the Gallup Organization and a senior consultant with ERT, Inc.

Address correspondence to Joan E. Broderick, PhD, Department of Psychiatry and Behavioral Science, Putnam Hall, South Campus, Stony Brook University, Stony Brook, NY 11794-8790. E-mail: Joan.Broderick@StonyBrook.edu.

Submitted for publication January 17, 2013; accepted in revised form March 27, 2013.
Significance & Innovations

- The National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS) has developed state-of-the-art short forms and computer adaptive testing methods for assessing domains of relevance to rheumatology.
- Longitudinal assessment using PROMIS instruments in a sample of osteoarthritis patients was compared with a sample from the general population.
- Known-group validity and ecological validity for the PROMIS instruments, pain intensity, pain interference, physical functioning, and fatigue were demonstrated.
- Test–retest reliability (7-day) was very good.

SUBJECTS AND METHODS

Subjects. This study is part of a larger project examining the ecological validity of PROMIS instruments across several clinical groups. It was approved by the Stony Brook Institutional Review Board and was conducted in compliance with the Good Clinical Practice and Declaration of Helsinki principles. Data were collected from OA patients (n = 100) and a comparison GP sample (n = 100). Both samples were recruited using a national online research panel of 1.7 million respondents (www.SurveySampling.com). Inclusion criteria for both samples were age ≥21 years, fluency in English, availability for 29–36 days, and high-speed internet access. OA patients were required to have a doctor-confirmed diagnosis of OA. Sampling of population participants was structured to match the demographic composition (age, sex, race, and ethnicity) of the US in 2009 according to the Census Bureau. For OA patients, recruitment was structured to approximate the demographic composition based upon US prevalence rates for arthritis (13).

Data collection. Data for this 4-week longitudinal study were collected on a daily basis. Participants completed the assessments on a computer via the PROMIS Assessment Center (http://www.assessmentcenter.net/), a free, online data collection tool. Participants provided electronic consent and were trained over the telephone in how to use the Assessment Center. Starting on the following day, participants completed daily SFs for each of the next 28 consecutive days. At the end of each week (on days 7, 14, 21, and 28), the PROMIS CAT instruments were administered in addition and prior to the daily SFs for that day. Compliance was monitored daily and participants were contacted if they missed an assessment. Participants were compensated $150 for study completion.

Assessment of medical comorbidities. At enrollment, participants completed 12 questions via the Assessment Center about current comorbid health conditions. Questions were drawn from the Arthritis Impact Measurement Scale (14) section on comorbidities.

PROMIS CAT instruments and corresponding daily measures. Four PROMIS domains were included in the present study: 1) the single item for pain intensity that assesses respondents’ average self-reported pain, 2) the PROMIS pain interference item bank measures the consequences of pain on a person’s life, including interference with social, cognitive, emotional, physical, and recreational activities, 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, and 4) the physical function item bank measures self-reported capability, including upper extremity (dexterity) and lower extremity (walking or mobility) functioning, central

across studies and clinical diagnoses (2). For this reason, PROMIS measures have been developed to be generic rather than disease specific (8). To date, the validity and reliability of PROMIS measures specifically in patients with rheumatologic diseases have not been fully established. Some results for the physical functioning scale have been previously reported, comparing it to legacy measures and examining sensitivity to change (3,9).

In this report, we examine “known-group validity,” ecological validity, and reliability of several PROMIS measures in patients with osteoarthritis (OA) using a general population (GP) sample as a comparison group. Known-group validity is demonstrated when the scores on a measure are significantly different between 2 groups that are expected to show differences (10) and the observed difference is in the predicted direction. It is important to note that a GP sample is not a “healthy” or “pain free” sample. As its name implies, it will include a cross section of people, some of whom are very healthy and others who have any number of illnesses. Ecological validity in this context indexes the degree to which PROs based on recall over a reporting period correspond with aggregated ratings collected in close temporal proximity to the experience (momentary or daily ratings). The underlying premise is that experience that is measured proximately has greater accuracy by precluding memory and recall bias (11). Thus, a high level of ecological validity suggests that the recall PRO provides a measure that accurately reflects aggregated daily experience. We compared PROMIS CAT scores (which ask about the “past 7 days,” i.e., the standard PROMIS recall period) with scores obtained using daily SF versions of the PROMIS measures. The domains reported are pain intensity, pain interference, fatigue, and physical functioning (the PROMIS pain intensity measure is a single numerical rating scale with 7-day recall and does not use CAT technology; for ease of communication in this study, the term PROMIS CAT includes the pain intensity item). These are among the most common self-report domains for OA (12).
4 weeks. The study was powered (80%) to detect a 0.4 SD formed separately for each week and for the average of all group (GP versus OA) as a within-person factor, and by testing the pleasure (mean ± SD) that is anchored to the distribution of scores in the US general population (8,15).

To obtain daily versions of these PROMIS domains, subsets of items from the banks were selected, consistent with the creation of PROMIS, version 1, SFs (2) and were administered daily as static SFs consisting of 1 item (pain intensity), 6 items (pain interference), 7 items (fatigue), and 10 items (physical functioning). The reporting period of each item (PROMIS physical functioning SFs and CAT do not specify a reporting period) was modified from “in the past 7 days . . .” to “in the last day. . .” Apart from this change, the wording and response options for each item were left unchanged. The daily measures were scored using item response theory, employing the expected posteriori estimator of the PROMIS scoring engine (16,17). This scoring allowed for a direct comparison of daily and CAT scores on the same metric.

Due to response burden concerns in the full study where the GP sample was serving as a comparison group for other clinical samples, the GP sample provided daily and weekly assessments for 4 domains, but not physical functioning. Therefore, ecological validity, but not known-group validity, will be reported for physical functioning.

Data analysis. To generate a 7-day summary score from daily SFs for comparison with PROMIS CAT instruments, the daily SF scores were averaged for each participant and week. In some cases, the CAT was completed a day late in which case the daily SFs for those 7 days were averaged.

To test the hypothesis that PROMIS CAT instruments demonstrate known-group validity, differences in mean scores between the OA and GP samples were examined using analysis of variance (ANOVA). To test the ecological validity of these group level differences, we examined whether the CAT score group differences were mirrored in the daily ratings. These hypotheses were addressed using mixed-effects ANOVA with “group” (GP versus OA sample) as a between-person factor and “method” (daily SF versus CAT) as a within-person factor, and by testing the group × method interaction term. The analysis was performed separately for each week and for the average of all 4 weeks. The study was powered (80%) to detect a 0.4 SD difference between the groups at each week (α = 0.05).

To further explore ecological validity, we examined the correspondence between daily and weekly CAT measures by computing for each week the between-person correlation of the CAT instruments and the weekly average of daily SFs, separately for the 2 samples. Tests for differences in independent correlations were used to determine if the validity was different for the 2 groups (18). The study was powered to detect a difference between correlations of 0.80 versus 0.60.

We also tested whether the 2 measurement methods (CAT and 1-week average of daily scores) demonstrated acceptable agreement for individual respondents. Even though the 2 assessment methods may not yield completely identical scores for each individual and week, it is desirable that the difference between the 2 scores lie within acceptable boundaries for most individuals. The proportion of difference scores within the limits of a minimum clinically important difference (MCID) is known as “coverage probability” (19,20). We computed a difference score between the 2 methods for each individual and week and estimated the percent of difference scores exceeding an MCID value, assuming a normal distribution of the difference scores (20). The variance of the difference scores was estimated for all 4 weeks simultaneously in a multivariate analysis, accounting for the repeated measures on the same individuals (21). For pain interference, fatigue, and physical functioning CAT scores, a value of ±6 points around the mean difference on the T score metric was chosen as criterion for an MCID, because it just exceeded the margins attributable to a 95% error margin of the CAT scores. Preliminary work on PROMIS measures has suggested similar thresholds for MCID (22). Several studies have indicated a value of ±1.7 points on the 0–10 numerical rating scale as MCID for pain intensity (23,24); it appears to be largely invariant across clinical conditions (25) and has also been suggested as appropriate MCID for patients with OA (24).

To examine the test–retest reliability of the measures, we calculated the intraclass correlation coefficient (ICC) across the 4 assessment weeks for aggregated daily SFs and weekly CAT instruments in each PRO domain.

Handling of missing data. Multiple imputations were used to account for missing assessments, wherein each missing value is replaced with a set of plausible values representing the uncertainty about the values to be imputed. Following recommendations (26), we used a set of 5 imputations, which were generated from the person-period data set of all study days and accounted for the correlated nature (“nonindependence”) of repeated daily measures within subjects (27). All analyses were performed using Mplus, version 7 (28).

RESULTS

Only 4 participants (2 in the OA sample and 2 in the GP sample) dropped out of the study and were not included in the analyses. Demographic characteristics of the 2 groups (n = 98 in each group) are shown in Table 1. Participants in the OA sample were significantly older, more likely to be receiving disability benefits, and had lower income than those in the GP sample. Our sampling strategy was successful in achieving a GP sample that was demographically comparable (age, sex, ethnicity/race) to the 2009 US population; the characteristics of the OA sample were comparable to reported US prevalence rates for arthritis.
For example, the mean age reported in the Census Bureau 2009 Population Survey is 44 years, which is similar to the mean in our sample, and the prevalence rate for arthritis in the general population is 21.5% (29), which is very close to the 19% in our GP sample. Education level was not used in recruitment matching of target samples, since very low education levels in the general population (15% not completing high school) were low frequency in our internet panel. The samples differed in other diseases, as would be expected based on the mean age difference (e.g., heart disease: 3% in GP, 12% in OA; high blood pressure: 22% and 46%, respectively).

Compliance with the 28-day daily protocol was high in both samples. On average, daily SFs were completed on mean ± SD 26.9 ± 1.55 days in the GP sample (4.0% missed) and on mean ± SD 26.8 ± 1.77 days in the OA sample (4.4% missed). Out of 392 weekly CAT instruments per sample, 16 (4.1%; GP sample) and 21 (5.4%; OA sample) were missed on the seventh day of the week and were completed on the following day; only 2 (0.5%) were totally missed in each sample.

Known-group differences: CAT instruments. The mean scores in the 2 samples based on daily SF scores and CAT instruments are shown in Table 2 (separated by week) and Table 3 (combined across weeks). The GP sample mean pain intensity level (2.5 across all weeks) was comparable to the PROMIS general population average of 2.6 (2), and the mean CAT scores for pain interference (51.3) and fatigue (48.8) were not significantly different from the PROMIS norm scores of 50. The mean levels of the OA sample significantly exceeded those of the GP sample (all \( P < 0.001 \)) for all PROMIS CAT instruments, with large effect sizes (Cohen’s d of 1.4 for pain intensity, 1.3 for pain interference, and 0.9 for fatigue), thus confirming the known-group validity of the PROMIS CAT instruments (Table 3).

**Ecological validity.** Our primary test of ecological validity is the correlation between CAT instruments and aggregated SFs for each week (Table 4). For the 4 PRO domains and both samples, the correlations range from 0.84 to 0.95 with narrow confidence intervals (the lower confidence limit of all correlations is \( r \geq 0.74 \)) showing a high correspondence between the 2 assessment methods. The magnitude of the correlations did not significantly differ between the GP and OA samples for any PRO domain in any week (\( P \geq 0.10 \) for all).

**Known-group and ecological validity.** Another test examined the ecological validity of the known-groups comparison, extending the CAT known-groups test. The idea is...
that the difference between the 2 groups for the CAT instruments should be similar to the difference when measured with the ecologically valid daily SFs. For both samples, the mean scores for each week of daily SFs were significantly lower ($P < 0.001$ for all) than the corresponding PROMIS CAT instruments for each PRO domain. However, the magnitude of this difference was similar for the OA and GP samples (Table 3) with no statistically significant group-by-reporting-period interaction, suggesting ecological validity of the group difference in CAT instruments.

**Individual patient-level agreement.** We next compared the CAT instruments and aggregated SF scores for individual respondents. As shown in Figure 1, differences between the scores exceeded the threshold for MCID in <25% of the cases for all PRO domains, with the smallest rates for pain intensity (<5%), and somewhat higher rates for fatigue scores (22%). For pain interference, individual patient agreement was significantly better ($P < 0.001$) for the OA sample than the GP sample.

**Test-retest reliability.** We examined the weekly test-retest reliability of PROMIS CAT instruments and aggregated daily SFs for each PRO domain. The ICCs, shown in Table 5, were consistently high for daily SFs (ICC range 0.83–0.95) and CAT scores (ICC range 0.80–0.92) across all PRO domains and did not significantly differ between the OA and GP samples ($P > 0.10$ for all).

### DISCUSSION

The purpose of this study was to examine the validity and reliability of the newly developed PROMIS CAT measures of pain intensity, pain interference, physical functioning,
and fatigue in OA patients, using a GP sample as a comparison. The samples were recruited using an innovative strategy of a national internet survey panel. Each sample was designed to reflect the US demographic characteristics of the targeted group, and the resulting samples match very closely.

As expected, OA patients showed significantly higher mean pain intensity, pain interference, and fatigue levels than the GP sample on the CAT instruments. These data provide strong support for the known-group validity of these PROMIS CAT instruments. Importantly, the group differences found for the CAT instruments were of the same magnitude as those found for aggregated daily SFs, confirming the ecological validity of the group differences.

Whereas the GP CAT scores were very close to the general population PROMIS norms, where a T score of 50 is “average,” our OA participants scored >70th percentile on pain intensity and pain interference, >70th percentile on fatigue, and <20th percentile for physical functioning.

As would be expected, approximately 20% of our GP sample reported that they had arthritis; however, we knew neither the type of arthritis nor if it was doctor diagnosed. We selected these 19 participants and looked at their CAT scores. Their average pain intensity score was >75th and >70th percentile for pain interference, and >70th percentile for fatigue. These scores are slightly lower than those in the OA sample, and provide further validity support.

Another positive finding was the test–retest reliability of CAT scores across 4 sequential assessment weeks. Some of our earlier research has reported the natural day-to-day variability in pain and other health experiences in rheumatology patients (30,31). Nevertheless, barring introduction of new treatment, injury, or other events, we would expect that a reliable measure would result in a respondent’s score being very similar across repeated measurements within a reasonable retest period. We found very good weekly reliabilities for the PROMIS CAT instruments ranging from 0.80–0.92 for both samples.

Known-group validity and reliability are important characteristics of a good PRO measure. We wanted to extend this examination to include ecological validity, since measurement error can be introduced into a recall score through memory errors and recall bias. Daily assess-

| Table 4. Correlations between PROMIS CAT instruments and aggregated daily SFs* |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|                             | Week 1                      | Week 2                      | Week 3                      | Week 4                      | Pooled across weeks         |
| Pain intensity              |                             |                             |                             |                             |                             |
| GP sample                   | 0.95 (0.92–0.96)            | 0.93 (0.88–0.96)            | 0.94 (0.90–0.96)            | 0.94 (0.90–0.97)            | 0.94 (0.91–0.96)            |
| OA sample                   | 0.91 (0.85–0.95)            | 0.92 (0.86–0.95)            | 0.95 (0.92–0.96)            | 0.95 (0.92–0.96)            | 0.93 (0.90–0.95)            |
| Pain interference           |                             |                             |                             |                             |                             |
| GP sample                   | 0.90 (0.85–0.93)            | 0.88 (0.84–0.91)            | 0.89 (0.83–0.93)            | 0.88 (0.83–0.92)            | 0.89 (0.85–0.91)            |
| OA sample                   | 0.87 (0.79–0.92)            | 0.86 (0.79–0.91)            | 0.91 (0.86–0.94)            | 0.87 (0.82–0.91)            | 0.88 (0.83–0.91)            |
| Fatigue                     |                             |                             |                             |                             |                             |
| GP sample                   | 0.88 (0.84–0.91)            | 0.90 (0.88–0.93)            | 0.88 (0.83–0.91)            | 0.89 (0.85–0.92)            | 0.89 (0.86–0.91)            |
| OA sample                   | 0.89 (0.83–0.93)            | 0.85 (0.74–0.91)            | 0.88 (0.78–0.93)            | 0.84 (0.74–0.91)            | 0.86 (0.78–0.91)            |
| Physical functioning        |                             |                             |                             |                             |                             |
| OA sample                   | 0.91 (0.87–0.94)            | 0.89 (0.86–0.92)            | 0.91 (0.87–0.94)            | 0.89 (0.84–0.92)            | 0.90 (0.87–0.92)            |

* Values are the correlation (95% confidence interval). PROMIS = Patient-Reported Outcomes Measurement Information System; CAT = computer adaptive testing; SFs = short forms; GP = general population; OA = osteoarthritis.

<table>
<thead>
<tr>
<th>Table 5. Test–retest (7-day) reliabilities*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week-to-week reliability (ICC)</td>
</tr>
<tr>
<td>CAT</td>
</tr>
<tr>
<td>Pain intensity</td>
</tr>
<tr>
<td>GP sample</td>
</tr>
<tr>
<td>OA sample</td>
</tr>
<tr>
<td>Pain interference</td>
</tr>
<tr>
<td>GP sample</td>
</tr>
<tr>
<td>OA sample</td>
</tr>
<tr>
<td>Fatigue</td>
</tr>
<tr>
<td>GP sample</td>
</tr>
<tr>
<td>OA sample</td>
</tr>
<tr>
<td>Physical functioning</td>
</tr>
<tr>
<td>OA sample</td>
</tr>
</tbody>
</table>

* ICC = intraclass correlation coefficient; CAT = computer adaptive testing; SFs = short forms; GP = general population; OA = osteoarthritis.

Figure 1. Percentage differences between Patient-Reported Outcomes Measurement Information System 7-day recall and aggregated daily assessments exceeding a threshold for minimum clinically important differences (MCIDs) in the general population and osteoarthritis (OA) samples. Error bars represent 95% confidence intervals.
ment is a method for reducing those measurement errors (11) and aggregating those scores yields the average experience for the week. Scores generated by PROMIS SFs and CAT instruments are expected to be very similar (2). Thus, ideally, the average of daily measurements of the domains for a week with SFs should correspond well with a 7-day recall CAT. The results support the conclusion that PROMIS CAT instruments demonstrate excellent ecological validity (32). The correlations between the aggregated SF scores and the CAT instruments ranged from 0.86 to 0.94. Importantly, the correlations in the OA sample were not lower than those found in the GP sample. This suggests good correspondence in OA patients who were older on average and for whom one may have speculated that poorer memory may result in less accurate recall. Furthermore, these indices of ecological validity for the PROMIS measures are higher than we have found in previous work examining other instruments measuring these domains (33). However, the prior work examined single-item daily and recall measures, which might be expected to have lower reliability.

Finally, since use of PROs for individual patient assessment in clinical settings is becoming more common (34), we wanted to drill down further to explore the ecological validity of PROMIS CAT instruments for individual patient scores. The difference between individual respondents’ CAT instruments and aggregated SFs generally supported our conclusions. Less than 8% of the OA patients had CAT and aggregated SF scores for pain intensity, pain interference, and physical functioning that differed by more than the MCID. The fatigue measures for both OA and GP samples did not fare as well, with approximately 20% of the respondents having discrepancies between the aggregated SFs and the CAT instruments at least as large as the MCID. In our prior research, we also found lower within-subject ecological validity for fatigue measures (30). Overall, this is good news for practical applications of PROMIS measures that require accurate and (ecologically) valid scores on the level of individual patients, e.g., when these measures are incorporated in patients’ electronic medical records to monitor individual patient status.

A careful examination of these data reveals a systematic pattern of the daily SF scores being lower than the CAT instruments. This is very consistent with prior research showing lower mean symptom ratings for shorter compared to longer recall periods (33,35–37). This phenomenon is attributed to recall bias in which a number of factors, such as salience of high symptom episodes, may influence how respondents recall symptom levels (38). From an applied perspective, the implications of this level difference in the absolute levels of the scales are minimal for most intended uses of the instrument. For PROMIS, the measures have been calibrated using a single (7-day recall) period and the intended use of the instruments will allow valid norm-based comparisons between studies.

There are several caveats and limitations that should be noted. First, as mentioned earlier, the participants were recruited from a national internet panel. Enrollment into the study proceeded until demographic characteristic (age, sex, race, and ethnicity) “bins” were filled in order to structure the samples to match US profiles for GP and OA patients. Since internet access was required to participate in the study, participants with very low education were not well represented (39). However, this group is typically not well represented in studies, and reading-level challenges often further impede participation (40). Importantly, 22% of our GP sample reported high blood pressure, which is almost identical to the 23% of the population reported by the Centers for Disease Control and Prevention as being aware of their hypertension (41). Likewise, our 2 samples’ self-reports of heart disease were very similar to national epidemiologic prevalence rates (42). Thus, the study results can be viewed as generalizing to all but those with very low education or without internet access.

The OA sample was comprised of people who self-reported a physician diagnosis of OA. The logistical constraints of the study precluded verifying the diagnosis. Misrepresentation is likely minimal as studies comparing self-report and physician-confirmed diagnoses have found agreement (43). The known-group differences that were observed convey confidence in the results. Indeed, it is possible that patients recruited from clinics might show even larger group differences.

Finally, these data were collected from people who, on average, were in an overall steady state regarding their medical conditions. Results could be different, especially for ecological validity, in the context of clinical change due to disease flare or treatment initiation. This study was conducted in steady state in order to examine validity and reliability in a controlled context. Subsequent work should examine these psychometric parameters in situations involving symptom change.

In conclusion, PROMIS CAT instruments for pain, interference due to pain, fatigue, and physical functioning demonstrated known-group and ecological validity in a comparison of OA patients with a GP sample. Good test–retest reliability was also observed. These data provide encouraging initial data on the utility of these PROMIS instruments for clinical and research outcomes in OA patients. Going forward, it will be important to examine sensitivity to change in clinical outcome trials. Furthermore, there is some expectation that the measurement precision of item response theory–based PROMIS instruments improves responsiveness across a wider range of symptom severity, which may reduce sample size requirements in clinical trials (3).

ACKNOWLEDGMENTS

We gratefully acknowledge our study participants. Our research assistants, Lauren Cody, Gim Yen Toh, and Laura Wolff, conducted the research with a high degree of rigor and great personal interactions with our participants. We are very appreciative of their important contribution. We are also grateful for the assistance of Christopher Christodoulou, PhD, who monitored our demographic sampling.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors...
approved the final version to be submitted for publication.
Dr. Broderick had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Broderick, Schwartz, Stone.

Acquisition of data. Broderick, Schneider, Junghaenel, Stone.

Analysis and interpretation of data. Broderick, Schneider, Junghaenel, Schwartz, Stone.

REFERENCES


