Published in final edited form as:
Arthritis Care Res (Hoboken). 2013 October ; 65(10): 1625–1633. doi:10.1002/acr.22025.

Validity and reliability of Patient-Reported Outcomes Measurement Information System (PROMIS) Instruments in Osteoarthritis

Joan E. Broderick, PhD, **Stefan Schneider, PhD**, **Doerte U. Junghaenel, PhD**, **Joseph E. Schwartz, PhD**, and **Arthur A. Stone, PhD**
Department of Psychiatry and Behavioral Science Stony Brook University

Abstract

Objective—Evaluation of known group validity, ecological validity, and test-retest reliability of four domain instruments from the Patient Reported Outcomes Measurement System (PROMIS) in osteoarthritis (OA) patients.

Methods—Recruitment of an osteoarthritis sample and a comparison general population (GP) 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 Test (CAT) instruments that use a 7-day recall. Known group validity (comparison of OA and GP), ecological validity (comparison of aggregated daily measures with CATs), and test-retest reliability were evaluated.

Results—The recruited samples matched (age, sex, race, ethnicity) the demographic characteristics of the U.S. 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: .85). Ecological validity was also established through high correlations between aggregated daily measures and weekly CATs (≥ .86). Test-retest validity (7-day) was very good (≥ .80).

Conclusion—PROMIS CAT instruments demonstrated known group and ecological validity in a comparison of osteoarthritis patients with a general population 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 osteoarthritis patients.

Keywords

pain; fatigue; physical functioning; patient outcomes assessment; reliability and validity; osteoarthritis

Corresponding author: Joan E. Broderick, Ph.D., Department of Psychiatry and Behavioral Science, Putnam Hall, South Campus, Stony Brook University, Stony Brook, NY 11794-8790, Phone: 631-632-8083, Fax: 631-632-3165, Joan.Broderick@StonyBrook.edu.

DISCLOSURES: AAS is a Senior Scientist with the Gallup Organization and a Senior Consultant with ERT, Inc. JEB also makes these disclosures due to her relationship to AAS.

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 published. 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, Junghaenel, Schneider, Stone.

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

The Patient-Reported Outcomes Measurement Information System (PROMIS), an NIH-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 (IRT), 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, including available sets of short-forms (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 mid-range 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 makes 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).

A goal of PROMIS has been to offer common metrics for the measurement of PROs to maximize comparability 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 rheumatological diseases has 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 comparison group. Known group validity is demonstrated when the scores on a measure are significantly different between two 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," which is the standard PROMIS recall period) with scores obtained using daily short form versions of the PROMIS measures. The domains reported are pain intensity,[1] pain interference, fatigue, and physical functioning. These are among the most common self-report domains for OA (12).

[1]The PROMIS pain intensity measure is a single numerical rating scale with 7-day recall. Thus, it does not use CAT technology. For ease of communication in this paper, when the term, "PROMIS CATs," is used, it includes the pain intensity item.

Participants and Methods

Participants

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 to 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 U.S. 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^{SM} (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 short forms (SFs) for each of the next 28 consecutive days. At the end of each week (on days 7, 14, 21, and 28), the PROMIS CATs 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.

Measures

Assessment of medical comorbidities—At enrollment, participants completed 12 questions on the Assessment Center about current comorbid health conditions. Questions were drawn from the Arthritis Impact Measurement Scale (AIMS2) (14) section on comorbidities

PROMIS CATs 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. 4) The physical function item bank measures self-reported capability including upper extremity (dexterity), and lower extremity (walking or mobility) functioning, central regions (neck, back), and instrumental activities of daily living.

These four domains were measured with daily PROMIS short-forms (available from: https://www.assessmentcenter.net/PromisForms.aspx) and compared with PROMIS CATs administered at the end of each week (PROMIS CAT demonstration available from: http://www.nihpromis.org/software/demonstration). The CATs were set to administer ≥ 4 and ≤ 12 items, and to terminate when SE < 3 T-score points (> .90 score reliability) was achieved. 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 (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 short-forms (2) and were administered daily as static short-forms consisting of 1 item (pain intensity), 6 items (pain interference), 7 items (fatigue), and 10 items (physical functioning). The reporting period of each item[2] 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 IRT employing the EAP 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 CATs, 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 CATs demonstrate known group validity, differences in mean scores between the OA and GP samples were examined using 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 a mixed-effects ANOVA with Group (General Population vs. OA sample) as a between-person factor and Method (daily SF vs. 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=.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 CATs and the week average of daily SFs, separately for the two samples. Tests for differences in independent correlations were used to determine if the validity was different for the two groups (18). The study was powered to detect a difference between correlations of .80 versus .60.

We also tested whether the two measurement methods (CAT and 1-week average of daily scores) demonstrated acceptable agreement for *individual* respondents. Even though the two assessment methods may not yield completely identical scores for each individual and week, it is desirable that the difference between the two 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 two methods for each individual and week and estimated the percent of difference scores exceeding a 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 a MCID, because it just exceeded the margins attributable

[2]PROMIS physical functioning short forms and CAT do not specify a reporting period.

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 CATs 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 five imputations, which were generated from the person-period dataset of all study days and accounting for the correlated nature ("non-independence") of repeated daily measures within subjects (27). All analyses were performed using Mplus Version 7 (28).

Results

Only four 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 two 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 U.S. population; the characteristics of the OA sample were comparable to reported U.S. prevalence rates for arthritis (13). For example, the mean age reported in the Census Bureau's 2009 Population Survey is 44 years, similar to the mean in our sample; and the prevalence rate for arthritis in the general population is 21.5% (29) -- 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% general population, 12% 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 26.9 (SD = 1.55) days in the GP sample (4.0% missed) and on 26.8 (SD = 1.77) days in the OA sample (4.4% missed). Out of 392 weekly CATs per sample, 16 (4.1%; GP sample) and 21 (5.4%; OA sample) were missed on the 7th day of the week and were completed on the following day; only 2 (0.5%) were totally missed in each sample.

Known group differences: CATs

The mean scores in the two samples based on daily SF scores and CATs 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 score means 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 ps < .001) for all PROMIS CATs, with large effect sizes (Cohen's d of 1.4 for pain intensity, 1.3 for pain interference, 0.9 for fatigue), thus confirming the known-groups validity of the PROMIS CATs (see Table 3).

Ecological Validity

Our primary test of ecological validity is the correlations between CATs and aggregated SFs for each week (Table 4). For the four PRO domains and both samples, the correlations range from .84 to .95 with narrow confidence intervals (the lower confidence limit of all correlations is $r \geq .74$), showing a high correspondence between the two assessment methods. The magnitude of the correlations did not significantly differ between the GP and OA samples for any PRO domain in any week (all ps $\geq .10$).

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 two groups for the CATS 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 (ps $< .001$) than the corresponding PROMIS CATs for each PRO domain. However, the magnitude of this difference was similar for the OA and GP samples (see Table 3) with no statistically significant group-by-reporting-period interaction, suggesting ecological validity of the group difference in CATs.

Individual patient-level agreement

We next compared the CATs and aggregated SF scores for individual respondents. As shown in Figure 1, differences between the scores exceeded the threshold for MCID in less than 25% of the cases for all PRO domains, with the smallest rates (< 5%) for pain intensity, and somewhat higher rates (22%) for fatigue scores. For pain interference, individual patient agreement was significantly better ($p < .001$) for the OA sample than the GP sample.

Test-retest reliability

We examined the weekly test-retest reliability of PROMIS CATs and aggregated daily SFs for each PRO domain. The intraclass correlations, shown in Table 5, were consistently high for daily SFs (ICCs ranging from .83 to .95) and CAT scores (ICCs ranging from .80 to .92) across all PRO domains and did not significantly differ between the OA and GP samples (all ps > .10).

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 U.S. 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 CATs. These data provide strong support for the known group validity of these PROMIS CATs. Importantly, the group differences found for the CATs 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 > 80th percentile on pain intensity and pain interference, > 70th percentile on fatigue, and < 20th percentile for physical functioning.

As would be expected, about 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 PROMIS percentile, > 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 four 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 CATs ranging from .80 - .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 assessment 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 CATs 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 CATs demonstrate excellent ecological validity (32). The correlations between the aggregated SF scores and the CATs ranged from .86 to .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 CATs for individual patient scores. The difference between individual respondents' CATs 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 about 20% of the respondents having discrepancies between the aggregated SFs and the CATs 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, for example, 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 CATs. 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, ethnicity) “bins” were filled in order to structure the samples to match U.S. profiles for GP and osteoarthritis 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 General Population sample reported high blood pressure; that is almost identical to the 23% of the population reported by the CDC as being aware of their hypertension (41). Likewise, our two sample’s self-reports of heart disease were very similar to national epidemiological 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 osteoarthritis. The logistical constraints of the study precluded verifying the diagnosis. Misrepresentation is likely minimal as studies comparing self-report and physician confirmed diagnosis 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 sum, PROMIS CAT instruments for pain, interference due to pain, fatigue, and physical functioning demonstrated known group and ecological validity in a comparison of osteoarthritis patients with a general population 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 osteoarthritis 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 IRT-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 conducted the research with a high degree of rigor and great personal interactions with our participants. We are very appreciative of their important contribution: Lauren Cody, Gim Yen Toh, and Laura Wolff. We are also grateful for the assistance of Christopher Christodoulou, Ph.D. who monitored our demographic sampling.

Grant support: NIH/NIAMS 1U01AR057948

Financial support: None

References

1. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007; 45:S22–31. [PubMed: 17443115]
2. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. 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]

3. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011; 13:R147. [PubMed: 21914216]
4. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010; 150:173–82. [PubMed: 20554116]
5. Lai JS, Cella D, Choi S, Junghaenel DU, Christodoulou C, Gershon R, et al. How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example. Arch Phys Med Rehabil. 2011; 92:S20–7. [PubMed: 21958919]
6. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res. 2010; 19:125–36. [PubMed: 19941077]
7. 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]
8. 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.
9. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol. 2009; 36:2061–6. [PubMed: 19738214]
10. DeVellis, RF. Scale development: Theory and applications. 2nd ed. Sage; Thousand Oaks, CA: 2003.
11. Stone, A.; Shiffman, SS. Ecological validity for patient reported outcomes. In: Steptoe, A., editor. Handbook of Behavioral Medicine: Methods and Applications. Springer; New York: 2010. p. 99-112.
12. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol. 1997; 24:799–802. [PubMed: 9101522]
13. Bolen J, Schieb L, Hootman JM, Helmick CG, Theis K, Murphy LB, et al. Differences in the prevalence and severity of arthritis among racial/ethnic groups in the United States, National Health Interview Survey, 2002, 2003, and 2006. Preventing Chronic Disease. 2010; 7:A64. [PubMed: 20394703]
14. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis Rheum. 1992; 35:1–10. [PubMed: 1731806]
15. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. J Clin Epidemiol. 2010; 63:1169–78. [PubMed: 20688473]
16. Choi SW. Firestar: Computerized Adaptive Testing (CAT) Simulation Program for Polytomous IRT Models. Appl Psychol Meas. 2009; 33:644–5.
17. Choi SW, Swartz RJ. Comparison of CAT Item Selection Criteria for Polytomous Items. Appl Psychol Meas. 2009; 33:419–40. [PubMed: 20011456]
18. Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed. Erlbaum; Hillsdale, NJ: 1988.
19. Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: Models, issues, and tools. Journal of the American Statistical Association. 2002; 97:257–70.
20. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat. 2007; 17:529–69. [PubMed: 17613641]
21. Choudhary P. A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. Journal of Statistical Planning and Inference. 2008; 138:1102–15.

22. 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. J Clin Epidemiol. 2011; 64:507–16. [PubMed: 21447427]
23. Farrar JT, Young JP Jr. LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain. 2001; 94:149–58. [PubMed: 11690728]
24. Tubach F, Ravaud P, Martin-Mola E, Awada H, Bellamy N, Bombardier C, et al. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: Results from a prospective multinational study. Arthritis Care Res (Hoboken). 2012; 64:1699–707. [PubMed: 22674853]
25. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005; 113:9–19. [PubMed: 15621359]
26. Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: A data analyst's perspective. Multivariate Behavioral Research. 1998; 33:545–71.
27. Graham J. Missing data analysis: Making it work in the real world. Annual Review of Psychology. 2009; 60:549–76.
28. Muthén, LK.; Muthén, BO. M*plus* user's guide. 7th ed.. Muthén & Muthén; Los Angeles, CA: 1998-2012.
29. Helmick CG, Felson DT, Lawrence RC, Gabriel S, Hirsch R, Kwoh CK, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. Arthritis Rheum. 2008; 58:15–25. [PubMed: 18163481]
30. Broderick JE, Schwartz JE, Schneider S, Stone AA. Can End-of-day reports replace momentary assessment of pain and fatigue? Journal of Pain. 2009; 10:274–81. [PubMed: 19070550]
31. Schneider S, Junghaenel DU, Keefe FJ, Schwartz JE, Stone AA, Broderick JE. Individual differences in the day-to-day variability of pain, fatigue, and well-being in patients with rheumatic disease: associations with psychological variables. Pain. 2012; 153:813–22. [PubMed: 22349917]
32. Shrout PE. Measurement reliability and agreement in psychiatry. Statistical Methods In Medical Research. 1998; 7:301–17. [PubMed: 9803527]
33. 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; 139:146–57. [PubMed: 18455312]
34. Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. Qual Life Res. 2008; 17:179–93. [PubMed: 18175207]
35. 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; 107:61–9. [PubMed: 14715390]
36. 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. Personality and Social Psychology Bulletin. 2005; 31:1340–6. [PubMed: 16143666]
37. Keller SD, Bayliss MS, Ware JE Jr. Hsu MA, Damiano AM, Goss TF. Comparison of responses to SF-36 Health Survey questions with one-week and four-week recall periods. Health Services Research. 1997; 32:367–84. [PubMed: 9240286]
38. 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]
39. Zickuhr, K.; Smith, A. Digital Differences. Pew Research Center; Washington, DC: 2012. http://pewinternetorg/Reports/2012/Digital-differencesaspx
40. Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol. 2007; 17:643–53. [PubMed: 17553702]
41. Yoon, S.; Burt, V.; Louis, T.; Carroll, M. Hypertension among adults in the United States, 2009-2010. brief, Nd, editor. National Center for Health Statistics; Hyattsville, MD: 2012.

42. Fang, J.; Shaw, K.; Keenan, N. Morbidity and Mortality Weekly Report (MMWR): Prevalence of Coronary Heart Disease --- United States, 2006--2010. Prevention DfHDaS. , editor. National Center for Chronic Disease Prevention and Health Promotion; 2011. p. 1377-81.

43. Bombard JM, Powell KE, Martin LM, Helmick CG, Wilson WH. Validity and reliability of self-reported arthritis: Georgia senior centers, 2000-2001. Am J Prev Med. 2005; 28:251–8. [PubMed: 15766612]

Significance and Innovations

The NIH Patient Reported Outcomes Measurement System has developed state-of-the-art short form 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.

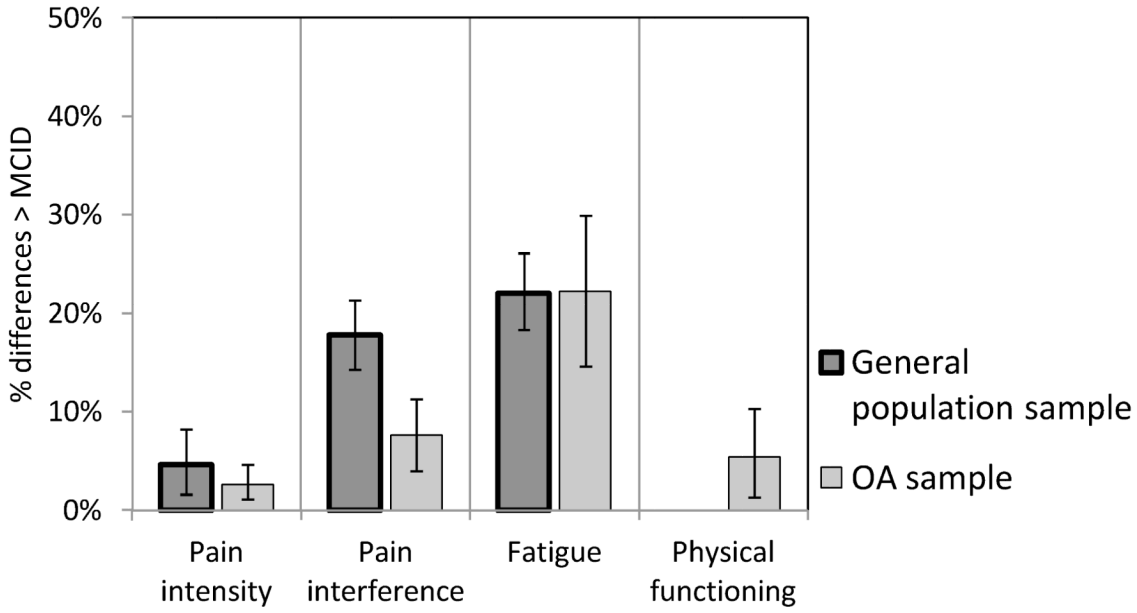

Figure 1.
Percent differences between PROMIS 7-day recall and aggregated daily assessments exceeding a threshold for Minimal Clinically Important Differences (MCID) in the general population and osteoarthritis (OA) samples. Error bars represent 95% confidence intervals.

Table 1

Demographic characteristics of the study samples

	GP sample n = 98	OA sample n = 98	p for difference between groups
Age, mean ± SD (range) years	43.9 ± 14.8 (21 -77)	56.9 ± 10.0 (29 - 81)	<.001
Arthritis diagnosis	19 (19.4)	98 (100.0)	<.001
Age categories, n (%)			<.001
21 - 44	49 (50.0)	9 (9.2)	
45 - 64	40 (40.8)	66 (67.4)	
65+	9 (9.2)	23 (23.5)	
Women, n (%)	50 (51.0)	59 (60.2)	.20
Race, n (%)			.31
White	69 (70.4)	77 (78.6)	
African American	15 (15.3)	15 (15.3)	
Asian	6 (6.1)	1 (1.0)	
Native American	2 (2.0)	1 (1.0)	
Other/multiple	6 (6.1)	4 (4.1)	
Hispanic, n (%)	14 (14.3)	11 (11.2)	.52
Married, n (%)	45 (45.9)	46 (46.9)	.89
Education, n (%)			.43
Less than high school	1 (1.0)	1 (1.0)	
High school graduate	17 (17.4)	10 (10.2)	
Some college	42 (42.9)	54 (55.1)	
College graduate	28 (28.6)	23 (23.5)	
Advanced degree	10 (10.2)	10 (10.2)	
Family income, n (%)[a]			.003
$0 - 19,999	6 (6.1)	23 (23.7)	
$20,000 - 34,999	22 (22.5)	28 (28.9)	
$35,000 - 49,999	28 (28.6)	21 (21.7)	
$50,000 - 74,999	22 (22.5)	12 (12.4)	
$75,000 and higher	20 (20.4)	13 (13.4)	
Employed, n (%)	70 (71.4)	31 (31.6)	<.001
Disability benefits, n (%)	10 (10.2)	30 (30.6)	<.001

Note:

GP=general population sample; OA=osteoarthritis sample.

[a] Income was not reported by one participant in the OA sample.

Arthritis Care Res (Hoboken). Author manuscript; available in PMC 2014 October 01.

Table 2

Means ± SDs for CATs and aggregated daily short forms by study sample and week

	CAT score				Daily short form score			
	Week 1	**Week 2**	**Week 3**	**Week 4**	**Week 1**	**Week 2**	**Week 3**	**Week 4**
Pain intensity *								
GP sample	2.60 ± 2.4	2.62 ± 2.5	2.39 ± 2.6	2.31 ± 2.3	2.23 ± 2.2	2.07 ± 2.3	2.00 ± 2.3	1.97 ± 2.3
OA sample	5.62 ± 1.9	5.54 ± 2.0	5.44 ± 2.1	5.43 ± 2.0	5.41 ± 1.8	5.16 ± 2.0	5.14 ± 2.1	5.10 ± 2.2
Pain interference								
GP sample	51.5 ± 9.2	51.4 ± 9.1	51.3 ± 10.2	51.0 ± 9.8	49.0 ± 8.1	48.5 ± 8.0	48.5 ± 8.3	48.3 ± 8.2
OA sample	61.0 ± 6.4	61.4 ± 6.4	60.4 ± 7.0	60.8 ± 6.9	59.0 ± 6.1	58.1 ± 6.5	58.0 ± 7.2	57.8 ± 7.3
Fatigue								
GP sample	49.2 ± 9.6	48.8 ± 10.2	48.9 ± 10.9	48.2 ± 10.1	45.9 ± 10.1	43.9 ± 10.0	43.8 ± 10.6	43.3 ± 10.0
OA sample	56.9 ± 7.9	56.2 ± 8.1	55.4 ± 8.6	56.2 ± 8.5	53.9 ± 8.9	51.8 ± 9.5	51.6 ± 10.1	51.4 ± 9.9
Physical functioning **								
OA sample	37.5 ± 6.7	37.5 ± 6.8	37.8 ± 7.8	37.1 ± 6.8	37.0 ± 6.5	36.9 ± 6.8	36.8 ± 6.7	36.8 ± 6.7

Note.

GP=general population sample; OA=osteoarthritis sample.T

* Pain intensity is measured with 1 item, a 0-10 numerical rating scale.

** Physical functioning was not measured in the general population sample.

Table 3

Means ± SDs for PROMIS CAT's and aggregated daily short forms (SFs) by sample, averaged across all 4 weeks

	GP sample	OA sample	Difference between groups	Effect size d
Pain intensity				
7-day recall *	2.48 ± 2.4	5.51 ± 1.9	3.03 ± 2.1	1.42
Daily item	2.07 ± 2.2	5.23 ± 1.9	3.16 ± 2.1	1.53
Pain interference				
CAT	51.3 ± 9.0	60.9 ± 6.1	9.58 ± 7.7	1.25
Daily SF	48.6 ± 7.7	58.2 ± 6.4	9.67 ± 7.1	1.37
Fatigue				
CAT	48.8 ± 9.6	56.2 ± 7.8	7.41 ± 8.8	0.85
Daily SF	44.2 ± 9.7	52.2 ± 9.1	7.96 ± 9.4	0.84
Physical **				
CAT	---	37.5 ± 6.8		
Daily SF	---	36.9 ± 6.5		

Note.

GP=general population sample; OA=osteoarthritis sample. Effect size d = group mean difference divided by the pooled standard deviation.

* Pain intensity is measured with 1 item, a 0-10 numerical rating scale.

** The general population sample did not complete the Physical Functioning measures.

Table 4

Correlations (95% confidence intervals) between PROMIS CATs and aggregated daily short forms

	Correlations				
	Week 1	**Week 2**	**Week 3**	**Week 4**	**Pooled across weeks**
Pain intensity					
GP sample	.95 (.92 -.96)	.93 (.88 −.96)	.94 (.90 −.96)	.94 (.90 −.97)	.94 (.91 −.96)
OA sample	.91 (.85 −.95)	.92 (.86 −.95)	.95 (.92 −.96)	.95 (.92 −.96)	.93 (.90 −.95)
Pain interference					
GP sample	.90 (.85 −.93)	.88 (.84 −.91)	.89 (.83 −.93)	.88 (.83 −.92)	.89 (.85 −.91)
OA sample	.87 (.79 −.92)	.86 (.79 −.91)	.91 (.86 − .94)	.87 (.82 − .91)	.88 (.83 −.91)
Fatigue					
GP sample	.88 (.84 − .91)	.90 (.88 −.93)	.88 (.83 −.91)	.89 (.85 −.92)	.89 (.86 −.91)
OA sample	.89 (.83 −.93)	.85 (.74 −.91)	.88 (.78 −.93)	.84 (.74 −.91)	.86 (.78 −.91)
Physical functioning					
OA sample	.91 (.87 −.94)	.89 (.86 −.92)	.91 (.87 −.94)	.89 (.84 −.92)	.90 (.87 −.92)

Note: GP=general population sample; OA=osteoarthritis sample.

Arthritis Care Res (Hoboken). Author manuscript; available in PMC 2014 October 01.

Table 5

Test-retest (7-day) reliabilities

	Week-to-week reliability (intraclass correlation)	
	CAT	**Aggregated daily short forms**
Pain intensity		
GP sample	.89	.93
OA sample	.83	.84
Pain interference		
GP sample	.84	.87
OA sample	.80	.83
Fatigue		
GP sample	.84	.86
OA sample	.85	.85
Physical functioning		
OA sample	.92	.95

Note: GP=general population sample; OA=osteoarthritis sample.