Responsiveness and Minimal Important Change of the PROMIS Pain Interference item bank in patients presented in Musculoskeletal Practice.

Wouter Schuller , Caroline B. Terwee , Berend Terluin , Daphne Rohrich , Raymond W.J.G. Ostelo , Henrica C.W. de Vet

PII: S1526-5900(22)00439-4
DOI: https://doi.org/10.1016/j.jpain.2022.10.013
Reference: YJPAI 4192

To appear in: *Journal of Pain*

Received date: 10 December 2021
Revised date: 19 October 2022
Accepted date: 20 October 2022

Please cite this article as: Wouter Schuller , Caroline B. Terwee , Berend Terluin , Daphne Rohrich , Raymond W.J.G. Ostelo , Henrica C.W. de Vet , Responsiveness and Minimal Important Change of the PROMIS Pain Interference item bank in patients presented in Musculoskeletal Practice., *Journal of Pain* (2022), doi: https://doi.org/10.1016/j.jpain.2022.10.013

Highlights

1. The PROMIS Pain Interference item bank was responsive in patients with musculoskeletal complaints.

2. Responsiveness was comparable to other frequently used measurement instruments.

3. Scores correlated well with several disease-specific measurement instruments.

4. Our results support the generic use of the PROMIS Pain Interference item bank.

5. Minimal Important Change was 3.2.

Responsiveness and Minimal Important Change of the PROMIS Pain Interference item bank in patients presented in Musculoskeletal Practice.

Wouter Schuller [1,2,3], Caroline B. Terwee [1,2], Berend Terluin [4], Daphne Rohrich [1,2], Raymond W.J.G. Ostelo [1,2,5], Henrica C.W. de Vet [1,2]

[1] Amsterdam UMC location Vrije Universiteit, Epidemiology and Data Science, De Boelelaan 1117, Amsterdam, Netherlands

[2] Amsterdam Public Health Research Institute, Methodology, Amsterdam, Netherlands

[3] Spine Clinic, Provincialeweg 152-154, Zaandam, Netherlands

[4] Amsterdam UMC location Vrije Universiteit Amsterdam, General Practice, De Boelelaan 1117, Amsterdam, Netherlands

[5] Health Science of the faculty of earth and life sciences and the Amsterdam Public Health Research Institute, De Boelelaan 1117, Amsterdam, Netherlands

Address for correspondence

Amsterdam UMC location Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam Public Health Research Institute, De Boelelaan 1117, Amsterdam, Netherlands

Phone number 0031204449255 (Fax number not available)

Email address w.schuller@amsterdamumc.nl

Disclosures

Funding sources

This study has been conducted as part of a larger research effort funded by the Dutch Association for Musculoskeletal Medicine.

Abstract

We evaluated the responsiveness of the PROMIS Pain Interference item bank in patients with musculoskeletal pain by testing predefined hypotheses about the relationship between the change scores on the item bank, change scores on legacy instruments and Global Ratings of Change (GRoC), and we estimated Minimal Important Change (MIC). Patients answered the full Dutch-Flemish V1.1 item bank. From the responses we derived scores for the standard 8-item short form (SF8a) and a CAT-score was simulated. Correlations between the change scores on the item bank, GRoC and legacy instruments were calculated, together with Effect Sizes (ES), Standardized Response Means (SRM), and Area Under the Curve (AUC). GRoC were used as an anchor for estimating the MIC with (adjusted) predictive modeling. Of 1677 patients answering baseline questionnaires 960 completed follow-up questionnaires at three months. The item bank correlated moderately high with the GRoC (Spearman's rho 0.63) and with the legacy instruments (Pearson's R ranging from 0.45 to 0.68). It showed a high ES (0.97) and SRM (0.71), and could distinguish well between improved and not improved patients based on the GRoC (AUC 0.77). Comparable results were found for the derived SF8a and CAT-scores. The MIC was estimated to be 3.2 (CI 2.6-3.7) T-score points.

Perspective

Our study supports the responsiveness of the PROMIS-PI item bank in patients with musculoskeletal complaints. Almost all predefined hypotheses were met (94%). The PROMIS-PI item bank correlated well with several legacy instruments which supports generic use of the item bank. MIC for PROMIS-PI was estimated to be 3.2 T-score points.

Key words: PROMIS, pain Interference, item bank, responsiveness, minimal important change, musculoskeletal complaints

Introduction

Part of the burden of living with pain is reflected in the way in which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities. For patients with pain, it is therefore important to measure these hindrances. The Pain Interference (PI) item bank was developed within the Patient Reported Outcome Information System (PROMIS™) domain framework[11] to measure the degree to which pain limits or interferes with daily activities[1]. Several studies evaluated psychometric properties of the PROMIS-PI item bank, reporting essential unidimensionality (Omega-H 0.97-0.99, Explained Common Variance 0.81-0.95[16,40]), good reliability (>0.95 for 96% of a sample of patients with musculoskeletal pain[14]), good construct validity (correlations >0.50 with several legacy instruments[16,40]), and good cross-cultural validity for the Dutch-Flemish translation (only 1-2 out of 40 items showing differential item functioning for language[14,40]). Where most questionnaires are disease-specific, PROMIS questionnaires were designed to be generic. The PROMIS-PI item bank was shown to correlate well with questionnaires addressing neck pain[5,22], low back pain[5,22], knee complaints[28], shoulder complaints[42], and foot and ankle complaints[35]. Sensitivity to change and responsiveness was addressed in several studies, including populations with carpal tunnel syndrome[21], spinal disorders[22,41], COPD[50], osteoarthritis[12], stroke[12], low back pain[3,12,25], knee pain[26], cancer[49], and depressive disorders[3]. All studies administered the PI item bank using either fixed short forms or Computerized Adaptive Testing (CAT). Some of these studies reported Minimal Important Change (MIC) as well, ranging from 1.9 to 8.9[2,4,6,13,26,49], on a T-score scale that is standardized to have a mean of 50 and a SD of 10. These studies, however, have some limitations. Most of these studies used effect sizes (ES) and Standardized Response Means (SRM) to evaluate responsiveness, but without specifying hypotheses about the expected magnitude of the changes. ES and SRM alone are insufficient measures of responsiveness, because they are influenced by floor and ceiling effects, and they are dependent on the standard deviation of the baseline score (ES) or the standard deviation of the change score (SRM). Therefore only comparisons of ES and SRM of different measurement instruments within studies, i.e. using the same sample, are informative to compare the responsiveness of instruments. Only a small number of studies, some of which with small sample

sizes, compared the responsiveness of the PI item bank with other frequently used PROMs that measure similar constructs (called legacy instrument). Most studies presenting MIC values used mean change methods to estimate MIC values, which do not reflect a threshold for minimal improvement[46]. This threshold is important because the aim of estimating MIC values is to get an impression of the minimum change score that can be considered important to the patient. Because of these limitations, and because responsiveness and MIC values may differ between populations, more work is needed to evaluate the responsiveness and the MIC in different settings[4]. We therefore aimed to study the responsiveness and estimate the MIC of the PROMIS-PI item bank in a population of patients with musculoskeletal complaints, who were treated by musculoskeletal physicians[37]. To evaluate the responsiveness we tested predefined hypotheses about the resemblance between the ES and SRM of the PROMIS-PI item bank and the ES and SRM of legacy instruments, together with predefined hypotheses about the correlations of the PROMIS-PI change scores with the change scores of legacy instruments and Global Ratings of Change (GRoC). We tested responsiveness for the full item bank and for derived scores for the standard 8-item short form (SF8a) and a simulated CAT. We estimated MIC using (adjusted) predictive modeling[43].

Methods

Study design

To collect data for our study we used an existing web-based registry of patients who presented at the practices of 31 participating musculoskeletal (MSK) physicians in the Netherlands. MSK physicians are medical doctors who are trained to use Spinal Manipulative Treatment (SMT). They are consulted by patients with a variety of musculoskeletal complaints, most frequently of spinal origin, such as low back pain or neck pain. Specific SMT techniques are almost invariably used, but can be combined with other treatment options, such as prescription medication, or injections in the spine under X-ray guidance[37]. For our study we recruited patients who presented at the MSK practice for a first

consultation. At the first visit (baseline), the physician entered data about the age, gender, type and duration of the main complaint and the existence of concomitant complaints in a web-based register. Main complaints were recorded according to the International Classification of Primary Care. Registered patients were asked to participate in this longitudinal study. After the patients gave informed consent the physician entered the patients email address in the registry. A computer program (Readmail) was custom built to send automated invitations by email to fill in a web-based questionnaire immediately after a patients email address was entered in the registry, and after a follow-up period of three months. From October 2013 to February 2014 the data from this registry was used for the present study. The study was approved by the Medical Ethical Committee of the VU Medical Center (2013/20).

Measures

Our study population responded to the full Dutch-Flemish V1.1 PI item bank (www.healthmeasures.net), and to several legacy instruments. The PROMIS-PI item bank consists of 40 items with a temporal context of 7 days (e.g. “in the past seven days, how much did pain interfere with your enjoyment in life”). Each item has five possible response options; three sets of response options are used to correspond to the different items: (1) not at all, a little bit, somewhat, quite a bit, very much, (2) never, rarely, sometimes, often, always, and (3) never, once a week or less, once every few days, once a day, every hour. T-scores for the PROMIS-PI item bank for each patient were calculated based on the US item parameters using the online Health Measures Scoring Service program, provided by the US Assessment Center. Higher scores represent higher trait levels, in this case more pain interference. The Dutch-Flemish PROMIS-PI item bank was validated in Dutch populations with chronic pain[14,40].

We calculated T-scores based on the full item bank. In addition we calculated scores using only the items from the standard 8-item short form (SF8a) and a simulated CAT. Post-hoc CAT simulations

were performed with the R-package catR (v3.16) using the standard PROMIS CAT starting and stopping rules and the original US item parameters, which were obtained from HealthMeasures.

In addition to the PROMIS-PI item bank our study sample responded to one out of five disease-specific legacy instruments, tailored to the main complaint. Patient with low back pain completed the Roland-Morris Disability Questionnaire (RDQ), a 24 item questionnaire measuring disability as a result of low back pain[36]. Total score ranges from 0-24, with higher scores indicating more disability. Patients with neck pain completed the Neck Disability Index (NDI), a 10 item questionnaire measuring self-reported pain intensity and limitations in daily activities[48]. Total score ranges from 0-50, with higher scores indicating more disability. Patients with upper extremity complaints completed the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire, a 30 item questionnaire measuring complaints and functional limitations of the whole upper extremity[20], with higher scores indicating more disability. Patients with lower extremity complaints completed the Lower Extremity Function Scale (LEFS), a 20 item questionnaire measuring functional disability in patients with musculoskeletal conditions of the lower extremity[9]. Total score ranges from 0-80, with lower scores indicating higher disability. Patients with headache or migraine completed the Headache Impact Test (HIT-6), a 6 item questionnaire measuring the impact of headache on daily activities[27]. Total score ranges from 36-78, with higher scores indicating a higher impact. All legacy instruments are frequently used in research and their validity was evaluated in Dutch populations[7,8,19,23,24,29,47].

At three months follow-up patients were asked to rate their perceived change in pain interference on a Retrospective Global Ratings of Change (GRoC) instrument (`Compared to three months ago, how much do you think that the limitations that you experience due to your pain have changed`). Patients answered a single item question about their perceived change with the following response options: (1) much improved, (2) improved, (3) slightly improved, (4) unchanged, (5) slightly worse, (6) worse, or (7) much worse. A recently published paper, evaluating the reliability of transition ratings, reported the reliability of this GRoC to be relatively high[18].

Statistical analyses

Descriptive analyses were presented for the complete sample of baseline responders, and for the group of patients who did or did not answer the follow-up measurement. Possible selective loss to follow-up was evaluated by comparing baseline characteristics between the groups of patients who did or did not answer the follow-up measurement. Responsiveness was defined by the COSMIN initiative as the ability of an instrument to detect changes over time in the construct to be measured[33]. We used various approaches to test responsiveness[17,31]. First the PROMIS-PI measures and legacy instruments were correlated with the GRoC. Measuring comparable, but not precisely the same construct, we expected the PROMIS-PI item bank to correlate at least moderately with the GRoC and with the legacy instruments, with a correlation coefficient of at least 0.50[34]. As a second approach, the Area Under the Receiver Operating Characteristics Curve (AUC) was calculated for all instruments, after dividing the patient population in a group of patients considered to have improved and a group of patients considered not to have improved, based upon the GRoC scores. Patients reporting any form of improvement were considered improved (GRoC categories 1-3, much improved, improved, or slightly improved) and patients reporting to be unchanged and patients reported to be worse were considered not improved (GRoC categories 4-7, unchanged, slightly worse, worse, or much worse). The AUC is considered to reflect the ability of the instrument to discriminate between patients who reported to be improved and patients who reported to be not improved. An AUC of >0.70 indicates adequate ability to distinguish patients who have or have not changed[45]. As a third approach, the correlation between the change in T-scores with the change in the scores on the legacy instruments was assessed, and the Effect Size (ES) and the Standardized Response Mean (SRM) of the PROMIS-PI item bank was compared to the ES and SRM of the legacy instruments. ES is calculated by dividing the mean change score by the SD at baseline. The SRM is calculated by dividing the mean change score by the SD of that change score. We expected the PROMIS-PI item bank to measure change comparable to the legacy instruments, or higher due to the

absence of floor and ceiling effects[17,30,32]. Responsiveness measures are reported for the T-scores calculated from the full item bank as well as the T-scores derived from the subset of items making up the standard 8-item short form and the simulated CAT.

The following hypotheses were tested:

- We expected a correlation of at least -0.50 between the change in the PROMIS-PI T-score and the GRoC score. This correlation was expected to be negative because improvement on the PROMIS-PI item bank is represented by a lower score while improvement on the GRoC is represented by higher score.
- We expected a correlation of at least 0.50 between the change in T-score of the PROMIS-PI item bank and the legacy instruments measuring functional disability (RDQ, NDI, DASH, LEFS and HIT-6). These correlations were expected to be positive for NDI, the RDQ, the HIT-6 and the DASH, in which the disability increases with higher scores. Correlations were expected to be negative for the LEFS, in which disability decreases with higher scores.
- We expected an AUC in excess of 0.70 for the PROMIS-PI item bank.
- We expected the ES and the SRM of the PROMIS-PI measures to be larger, the same, or at the most 0.05 smaller than the ES and the SRM of the legacy instruments.

Responsiveness was considered sufficient if at least 75% of the results were in accordance with the predefined hypotheses.

Minimal Important Change (MIC) was defined as a threshold for a minimal within-person change over time above which patients perceive themselves importantly changed. Assuming that all patients have their individual threshold of what they consider a minimal important change, the MIC can be conceptualized as the mean of these individual thresholds[46]. The MIC can be used as a threshold to determine the number of patients who have improved[46]. We estimated MIC based on data of the full item bank, using predictive modeling. With predictive modeling, the MIC is defined as the change

score where the post-test probability of belonging to the improved group equals the pre-test probability (Likelihood ratio = 1)[44]. The predictive modeling MIC estimates the mean of the hypothetical individual MIC values if the proportion of improved patients is 50%[43]. To detect minimal important change (in this case minimal important improvement), all patients reporting any form of improvement were considered improved and patients reporting to be unchanged and patients reported to be worse were considered not improved. MIC values were adjusted for the proportion of improved patients, i.e. adjusted predictive modelling[43], and bootstrapping was used to obtain confidence intervals.

Results

Sample characteristics

From October 2013 to February 2014 2610 patients were invited for our study of whom 2171 consented to participate. Of these patients 1677 (77%) completed the questionnaires at baseline, and 960 patients (57% of baseline responders) completed the questionnaires after the follow-up period of three months.

Demographic characteristics of the included sample are presented in Table 1. The average age of the whole sample was 47 years and most patients were female (59%), most patients were treated for spinal pain (75%), predominantly low back pain with or without sciatica (51%) and neck pain (16%), only a small number of patients reported complaints of shorter duration than three months (19%), and more than half of the patients had complaints longer than one year (61%). The RDQ was completed by 493 patients, the NDI by 167 patients, the LEFS by 98 patients, the DASH by 51 patients, and the HIT-6 by 35 patients.

Comparing groups of patient who did or did not answer the follow-up questionnaire showed no significant differences as far as baseline T-scores and baseline legacy scores were concerned. Non-responders differed statistically significant from responders in terms of age (non-responders were on average 4 years younger), and gender (non-responders were more likely male (45% versus 39%)).

Change in PROM scores over time

Most patients reported improvement on the GRoC between baseline and 3 months follow-up. In Table 2 the changes in scores (T0-T1) on the PROMIS-PI item bank and on the legacy instruments are presented, stratified by the GRoC scores. When using the full item bank patients who reported to be slightly improved, improved, or much improved, changed on average 2.1, 6.0, and 15.1 T-score points, respectively.

Responsiveness

All responsiveness results are presented in Tables 3 and 4. The correlation with the GRoC was more than -0.50, as hypothesized, for the full bank, the short form, and for the simulated CAT (-0.63, -0.60, and -0.57 respectively). The AUC was above 0.70 for the full bank, the short form and for the simulated CAT (0.77, 0.75, and 0.74 respectively). Correlations with legacy instruments were above the hypothesized 0.50 (ranging from 0.58 to 0.68), except for the LEFS (-0.45, -0.50 and –0.38 for the full bank, the 8-item SF and for the simulated CAT respectively). The AUC was above the hypothesized 0.70 for all PROMIS-PI measures. The ES and SRM of the PROMIS-PI item bank was higher than the ES and SRM of all the legacy instruments, except for the SRM of the DASH, which was slightly higher than the PROMIS-PI (0.76 as compared to 0.69-0.72 for PROMIS-PI). This difference was more than the hypothesized 0.05. Combining all these findings, 94% of our results were in accordance with the predefined hypotheses, strongly supporting responsiveness of the PROMIS-PI item bank (Table 5).

Minimal Important Change

The MIC for the PROMIS-PI was estimated to be 3.2 T-score points, with a confidence interval of 2.6-3.7.

Discussion

We studied responsiveness of the PROMIS-PI item bank in a population of patients with musculoskeletal complaints treated by musculoskeletal physicians. Predefined hypotheses about the relation between PROMIS-PI change scores with the change scores of several legacy instruments were tested. Furthermore, we reported the responsiveness for the full item bank as well as the responsiveness of the subset of items making up the standard 8-item short form and a simulated CAT, and we estimated the minimal important change of the PROMIS-PI. Almost all previously defined hypotheses (94%) were met, which strongly supports the responsiveness of the PROMIS-PI item bank in patients with various musculoskeletal complaints. Using adjusted predictive modeling we estimated a MIC of 3.2 for the full PROMIS-PI item bank (CI 2.6-3.7). Since short forms and CAT are based on the full item bank, we consider this MIC value also applicable to short forms and CAT derived from this item bank.

All previous studies except for one[25] reported positively on the responsiveness of the PROMIS-PI item bank. The study with a negative outcome was carried out in a population with musculoskeletal complaints treated with telecare management, overall showing small effect sizes(15). It cannot be ruled out, however, that the telecare was ineffective, rather than that the PROMIS-PI was not responsive. Since responsiveness is concerned with measuring change over time, it is necessary to study responsiveness in a population that shows change over time. To estimate MIC values, it is also necessary to have a proportion of unchanged patients. In that respect, our study population was very well suited to evaluate both responsiveness and to estimate MIC values, as some patients reporting various levels of improvement, and some patients reporting no change or deterioration. Comparing previous studies, the techniques used to evaluate responsiveness differed strongly. Many studies used effect sizes as a measure of responsiveness, a method that has limitations. An instrument should not only measure change in the purported construct, but it should measure the right amount of change, i.e. it should not under- or overestimate the real change in the construct that has occurred[30]. Therefore, change scores from new measures should be compared to change scores of

existing instruments in which the responsiveness was properly evaluated. However, even with the different techniques used, almost all studies reported positive findings, and we would suggest that there is strong evidence for the responsiveness of the PROMIS-PI item bank.

Similarly, estimates of the MIC will be influenced by the population studied and the techniques used. The only study that reported a very low MIC value was conducted on a cohort of patients with arthritis, rheumatism and aging, without specific treatment[4]. The correlation with the anchor in this study was very low (0.13-0.29), questioning the validity of the reported MIC value of 1.9. Other studies reporting on MIC values were conducted on patients with low back pain, osteoarthritis or stroke (MIC of patients with pain: 2-3, MIC of non-pain patients: 3.5-4.5)[13], cancer (MIC: 4.0-6.0)[49], low back pain or depression (MIC: 3.5-5.5)[2], and patients undergoing knee arthroscopy (MIC: 3.2)[26] or surgery for carpal tunnel syndrome (MIC: 8.9, 9.7, and 4.1)[6]. Most studies used the mean change method, or the ROC method with either global ratings of change or known MIC values of legacy instruments as an anchor. Estimating MIC values with predicting modeling is more precise than when using the ROC method[44]. Another advantage of predictive modeling compared to the ROC method is that MIC values can be corrected for bias if the proportion of improved patients does not equal 50%[43]. This correction cannot be done with the ROC method. Although the variation in the MIC values reported can be explained by differences in population studies and methods used, the values estimated in our study are comparable to most of those previously reported[2,6,13,26,49].

In our study, the responsiveness results were similar for the full bank, the derived short form, and simulated CAT, which may indicate that the PROMIS-PI short form and CAT were equally responsive as the full item bank, although they contain much less items. Correlation with the legacy instruments were very comparable between the scores obtained from the full item bank and scores obtained only using the items from the 8-item short form and the simulated CAT, which is in line with previous reports about the correlation of PROMIS short-forms with full item bank scores[10]. It must be noted that our population consisted predominantly of patients with spinal complaints, and a much smaller proportion of patients presented with headache or with complaints of the upper or lower extremity.

Therefore the comparisons with the DASH, the LEFS and the HIT-6 may be less reliable than the comparisons with the RDQ and the ODI.

Our study confirmed previous reports about the responsiveness of the PI item bank.

In our study, we evaluated responsiveness using predefined hypotheses about the relationship with the change scores of a number of legacy instruments, and (adjusted) predictive modeling to estimate MIC values. The PROMIS-PI item bank showed strong evidence supporting responsiveness. Other studies showed similar correlations with legacy instruments in cross-sectional studies, supporting the generic applicability of the item bank[5,22,28,35,42]. A previous study also showed that scores can be compared across different populations (limited differential item functioning was found)[15]. The PI item bank, therefore, may replace a number of disease-specific instruments, which would greatly simplify routine monitoring of patients with different musculoskeletal complaints.

Strengths and weaknesses

A strength of our study is the large sample of patients with MSK complaints completing the full PROMIS-PI item bank together with GRoC and legacy instruments both before and after treatment. Furthermore we tested predefined hypotheses to assess responsiveness, as recommended by COSMIN[32]. A weakness of our study was that the short form and CAT were not independently administered but derived from the answers to the full item bank. Therefore the results presented for the short form and CAT will be approximations of the responsiveness for these ways of administrating the item bank. Another weakness could be the relatively low percentage of responders at follow-up. Only 57% of baseline responders completed the questionnaires at three months follow-up. Non-responders were slightly younger and more often of male gender[38]. The T-scores on the PROMIS-PI item bank did not differ between responders and non-responders, nor did the scores on the legacy instruments. As the goal of our study was to test responsiveness rather than to measure the effect of treatment, selective loss to follow-up is not likely to cause bias. In responsiveness studies the change scores on measurement instruments for a similar construct are

compared with each other or between a group of improved and not improved patients. Therefore, representativeness of the population is a less important issue. Another weakness is the relatively small proportion of patients with headache or with upper or lower extremity complaints, but we did not aim to estimate MIC values for subgroups of patients.

Significant differences in the age and sex are similar to previous studies in the same population, recruiting slightly more older patients and more patients of the female sex[38-40]. We think that is unlikely that this may have biased our study results. Although it may be interesting to evaluate whether the responsiveness is different for different subpopulations, this is not one of the aims of the present study.

Effect sizes in our study are high, which may reflect the effectiveness of the treatment. However, our study is not intended to evaluate the effect of the treatment, but solely to evaluate responsiveness. Recruiting patients at a first consultation likely selects patients at a point in time where their fluctuating pain is high, and the high effect sizes may (partially) be explained by regression to the mean.

Conclusion

The change scores of the PROMIS-PI item bank correlated well with Global Ratings of Change and with the change scores of a number of legacy instruments, except for the LEFS. Effect Sizes, Standardized Response Means, and Area Under the Curve of the item bank were mostly slightly higher than those of the legacy instruments. Based on a priori hypotheses, the PROMIS-PI item bank showed sufficient responsiveness in patients with musculoskeletal complaints. MIC was estimated to be 3.2 T-score points. Our study supports the generic use of the PROMIS-PI in patients with a variety of musculoskeletal complaints.

Acknowledgements

We would like to thank all members of the Dutch Association for Musculoskeletal Medicine (Nederlandse Vereniging voor Artsen Musculoskeletale Geneeskunde, NVAMG) who cooperated in this study.

Declaration of Competing Interest

Caroline B. Terwee is head of the Dutch-Flemish PROMIS National Center. Caroline B. Terwee and Henrica C.W. de Vet are members of the PROMIS Health Organization, of which Caroline B. Terwee is board member. Caroline B. Terwee previously received grants for work on the translation and validation of the PROMIS item banks. Wouter Schuller is one of the owners of Prismascan, a Dutch company that developed an application using PROMIS CAT for routine outcome measurements. Berend Terluin, Daphne C. Rohrich, and Raymond W. Ostelo declare that they have no conflicts of interest.

References

1. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, Cella D, Rothrock N, Keefe F, Callahan L, Lai JS. Development of a PROMIS item bank to measure pain interference. Pain 150:173-82, 2010
2. Amtmann D, Kim J, Chung H, Askew RL, Park R, Cook KF. Minimally important differences for Patient Reported Outcomes Measurement Information System pain interference for individuals with back pain. J Pain Res 9:251-5, 2016
3. Askew RL, Cook KF, Revicki DA, Cella D, Amtmann D. Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. J Clin Epidemiol 73:103-11, 2016
4. Beaumont JL, Davis ES, Fries JF, Curtis JR, Cella D, Yun H. Meaningful change thresholds for Patient-Reported Outcomes Measurement Information System (PROMIS) fatigue and pain interference scores in patients with rheumatoid arthritis. J Rheumatol 48:1239-1242, 2021
5. Bernstein DN, Greenstein AS, D'Amore T, Mesfin A. Do PROMIS Physical Function, Pain Interference, and Depression Correlate to the Oswestry Disability Index and Neck Disability Index in Spine Trauma Patients? Spine (Phila Pa 1976) 45:764-9, 2020
6. Bernstein DN, Houck JR, Mahmood B, Hammert WC. Minimal Clinically Important Differences for PROMIS Physical Function, Upper Extremity, and Pain Interference in Carpal Tunnel Release Using Region- and Condition-Specific PROM Tools. J Hand Surg Am 44:635-40, 2019
7. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 65:71-6, 1996
8. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine (Phila Pa 1976) 20:1017-28, 1995
9. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. Phys Ther 79:371-83, 1999
10. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, Yount S, Cook KF, Gershon RC, Amtmann D, DeWalt DA, Pilkonis PA, Stone AA, Weinfurt K, Reeve BB. PROMIS((R)) Adult Health Profiles: Efficient Short-Form Measures of Seven Health Domains. Value Health 22:537-44, 2019
11. 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 63:1179-94, 2010
12. Chen CX, Kroenke K, Stump T, Kean J, Krebs EE, Bair MJ, Damush T, Monahan PO. Comparative Responsiveness of the PROMIS Pain Interference Short Forms With Legacy Pain Measures: Results From Three Randomized Clinical Trials. J Pain 20:664-75, 2019
13. Chen CX, Kroenke K, Stump TE, Kean J, Carpenter JS, Krebs EE, Bair MJ, Damush TM, Monahan PO. Estimating minimally important differences for the PROMIS pain interference scales: results from 3 randomized clinical trials. Pain 159:775-82, 2018
14. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, Cook KF, Revicki D, van LJ, Boers M, Dekker J, Terwee CB. Calibration and Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients with Chronic Pain. PLoS One 10:e0134094, 2015
15. Crins MHP, Terwee CB, Ogreden O, Schuller W, Dekker P, Flens G, Rohrich DC, Roorda LD. Differential item functioning of the PROMIS physical function, pain interference, and pain behavior item banks across patients with different musculoskeletal disorders and persons from the general population. Qual Life Res 28:1231-1243, 2019
16. Crins MHP, Terwee CB, Westhovens R, van Schaardenburg D, Smits N, Joly J, Verschueren P, Van der Elst K, Dekker J, Boers M, Roorda LD. First Validation of the Full PROMIS Pain Interference

and Pain Behavior Item Banks in Patients With Rheumatoid Arthritis. Arthritis Care Res (Hoboken) 72:1550-9, 2020
17. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine. In. Cambridge: Cambridge University Press; 2011
18. Griffiths P, Terluin B, Trigg A, Schuller W, Bjorner JB. A confirmatory factor analysis approach was found to accurately estimate the reliability of transition ratings. J Clin Epidemiol 141:36-45, 2022
19. Hoogeboom TJ, de Bie RA, den Broeder AA, van den Ende CH. The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. BMC Musculoskelet Disord 13:117, 2012
20. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med 29:602-8, 1996
21. Hung M, Saltzman CL, Greene T, Voss MW, Bounsanga J, Gu Y, Wang AA, Hutchinson D, Tyser AR. The responsiveness of the PROMIS instruments and the qDASH in an upper extremity population. J Patient Rep Outcomes 1:12, 2017
22. Hung M, Saltzman CL, Voss MW, Bounsanga J, Kendall R, Spiker R, Lawrence B, Brodke D. Responsiveness of the Patient-Reported Outcomes Measurement Information System (PROMIS), Neck Disability Index (NDI) and Oswestry Disability Index (ODI) instruments in patients with spinal disorders. Spine J 19:34-40, 2019
23. Jorritsma W, de Vries GE, Geertzen JH, Dijkstra PU, Reneman MF. Neck Pain and Disability Scale and the Neck Disability Index: reproducibility of the Dutch Language Versions. Eur Spine J 19:1695-701, 2010
24. Jorritsma W, Dijkstra PU, de Vries GE, Geertzen JH, Reneman MF. Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J 21:2550-7, 2012
25. Kean J, Monahan PO, Kroenke K, Wu J, Yu Z, Stump TE, Krebs EE. Comparative Responsiveness of the PROMIS Pain Interference Short Forms, Brief Pain Inventory, PEG, and SF-36 Bodily Pain Subscale. Med Care 54:414-21, 2015
26. Kenney RJ, Houck J, Giordano BD, Baumhauer JF, Herbert M, Maloney MD. Do Patient Reported Outcome Measurement Information System (PROMIS) Scales Demonstrate Responsiveness as Well as Disease-Specific Scales in Patients Undergoing Knee Arthroscopy? Am J Sports Med 47:1396-403, 2019
27. Kosinski M, Bayliss MS, Bjorner JB, Ware JE, Jr., Garber WH, Batenhorst A, Cady R, Dahlof CG, Dowson A, Tepper S. A six-item short-form survey for measuring headache impact: the HIT-6. Qual Life Res 12:963-74, 2003
28. Lu Y, Beletsky A, Nwachukwu BU, Patel BH, Okoroha KR, Verma N, Cole B, Forsythe B. Performance of PROMIS Physical Function, Pain Interference, and Depression Computer Adaptive Tests Instruments in Patients Undergoing Meniscal Surgery. Arthrosc Sports Med Rehabil 2:e451-e9, 2020
29. Martin M, Blaisdell B, Kwong JW, Bjorner JB. The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. J Clin Epidemiol 57:1271-8, 2004
30. Mokkink LB, Terwee CB, de Vet HCW. Key concepts in clinical epidemiology: Responsiveness, the longitudinal aspect of validity. J Clin Epidemiol 140:159-162, 2021.
31. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol 6:2, 2006
32. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 19:539-49, 2010
33. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of

measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 63:737-45, 2010
34. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J 24:69-71, 2012
35. Nixon DC, McCormick JJ, Johnson JE, Klein SE. PROMIS Pain Interference and Physical Function Scores Correlate With the Foot and Ankle Ability Measure (FAAM) in Patients With Hallux Valgus. Clin Orthop Relat Res 475:2775-80, 2017
36. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 8:141-4, 1983
37. Schuller W, Ostelo R, Rohrich DC, Apeldoorn AT, de Vet HCW. Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients. BMC Musculoskelet Disord 18:512, 2017
38. Schuller W, Ostelo RW, Rohrich DC, Heymans MW, de Vet HCW. Pain trajectories and possible predictors of a favourable course of low back pain in patients consulting musculoskeletal physicians in The Netherlands. Chiropr Man Therap 29:38, 2021
39. Schuller W, Terwee CB, Klausch T, Roorda LD, Rohrich DC, Ostelo RW, Terluin B, de Vet HCW. Psychometric properties of the Dutch-Flemish Patient-Reported Outcomes Measurement Information System Pain Behavior item bank in patients with musculoskeletal complaints. J Pain 20:1328-37, 2019
40. Schuller W, Terwee CB, Klausch T, Roorda LD, Rohrich DC, Ostelo RW, Terluin B, de Vet HCW. Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients With Musculoskeletal Complaints. Spine (Phila Pa 1976) 44:411-9, 2019
41. Sharma M, Ugiliweneza B, Beswick J, Boakye M. Concurrent Validity and Comparative Responsiveness of PROMIS-SF Versus Legacy Measures in the Cervical and Lumbar Spine Population: Longitudinal Analysis from Baseline to Postsurgery. World Neurosurg 115:e664-e75, 2018
42. Strong B, Maloney M, Baumhauer J, Schaffer J, Houck JR, Hung M, Bounsanga J, Voss MW, Gu Y, Voloshin I. Psychometric evaluation of the Patient-Reported Outcomes Measurement Information System (PROMIS) Physical Function and Pain Interference Computer Adaptive Test for subacromial impingement syndrome. J Shoulder Elbow Surg 28:324-9, 2019
43. Terluin B, Eekhout I, Terwee CB. The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. J Clin Epidemiol 83:90-100, 2017
44. Terluin B, Eekhout I, Terwee CB, de Vet HC. Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. J Clin Epidemiol 68:1388-96, 2015
45. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34-42, 2007
46. Terwee CB, Peipert JD, Chapman R, Lai JS, Terluin B, Cella D, Griffith P, Mokkink LB. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 30:2729-2754, 2021.
47. Veehof MM, Sleegers EJ, van Veldhoven NH, Schuurman AH, van Meeteren NL. Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). J Hand Ther 15:347-54, 2002
48. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 14:409-15, 1991
49. 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 64:507-16, 2011
50. Yount SE, Atwood C, Donohue J, Hays RD, Irwin D, Leidy NK, Liu H, Spritzer KL, DeWalt DA. Responsiveness of PROMIS(R) to change in chronic obstructive pulmonary disease. J Patient Rep Outcomes 3:65, 2019

Table 1;

Title: Demographic data and treatment details for the whole sample, responders, and non-responders at follow-up.

Legend: Baseline data of entire population and of responders versus non-responders at follow-up. Differences responders versus non-responders at follow-up were analyzed with regression analyses for continuous and dichotomous variables, and Chi2 for nominal variables.

Table 2;

Title: Change scores and standard deviation (SD) of all instruments across GRoC response categories.

Legend: GRoC, Global Ratings Of Change; PROMIS-PI, PROMIS Pain Interference; FB, Full Bank; SF8, fixe 8-item Short Form; CAT, Computer Adaptive Test; RDQ, Roland Disability Questionnaire; NDI, Neck Disability Index; LEFS, lower Extremity Function Scale; HIT-6 Headache Impact Test; DASH, Disability of Arm, Shoulder and Hand. No change scores are presented for GRoC categories with less than 2 patients (nn)

Table 3;

Title: Mean baseline scores (SD), mean follow-up scores, mean change (SD), effect size and standardized response mean of all instruments.

Legend: RDQ, Roland Disability Questionnaire; NDI, Neck Disability Index; LEFS, lower Extremity Function Scale; HIT-6 Headache Impact Test; DASH, Disability of Arm, Shoulder and Hand.

Table 4;

Title: Correlations between the PROMIS-PI item bank and legacy instruments.

Legend: AUC, Area Under the Curve; GRoC, Global Ratings Of Change; RDQ, Roland Disability Questionnaire; NDI, Neck Disability Index; LEFS, lower Extremity Function Scale; HIT-6 Headache Impact Test; DASH, Disability of Arm, Shoulder and Hand.

Table 5;

Title: Summary of responsiveness hypotheses.

Legend: Outcomes that are according to predefined hypotheses are marked with V, outcomes that are contrary to predefined hypotheses are marked X. For effect sizes and standardized response means the differences between PROMIS-PI and the legacy instruments are presente

Table 1; Demographic data and treatment details for the whole sample, responders, and non-responders at follow-up.

	Baseline responders	Follow-up	No follow-up	p-value[1]
Age mean (SD)	47.4 (13.8)	49.3 (13.3)	45.0 (14.1)	0.000
Sex (male)	41.5	38.6	45.3	0.006
Main complaint				0.015
Neck Pain	16.2	17.4	14.6	
Back Pain	8.0	6.8	9.6	
Low Back Pain without radiation	25.2	23.4	27.6	
Low Back Pain with radiation	25.4	28.3	21.5	
Upper extremity	6.8	5.8	8.3	
Lower extremity	9.7	10.2	9.0	
Headache	3.3	3.6	2.9	
Other	5.3	4.5	6.4	
Time since complaints started				0.239
< 1 month	6.1	5.8	6.4	
1-3 months	13.0	12.7	13.5	
3-6 months	11.4	11.1	12.0	
6-12 months	13.2	14.9	10.9	
1-2 years	14.8	13.6	16.4	
2-5 years	16.2	16.8	15.4	
> 5 years	25.3	25.2	25.4	
Current status				0.032
Single	12.6	12.8	12.4	
Married/ living together	77.5	78.9	75.6	
LAT	3.2	3.3	3.1	
With parents	3.9	2.6	5.7	
Other	2.7	2.4	3.1	
Activities				0.042
School/ study	4.0	3.1	5.2	
Working full-time (≥36 hours)	39.5	37.7	42.0	
Working part-time (<36 hours)	31.6	32.2	30.8	
Unpaid work	6.3	7.1	5.1	
Retired	10.9	12.0	9.3	
Unemployed	2.7	3.1	2.2	
Other	5.1	4.8	5.4	
Baseline scores PROMs (range)				
PROMIS PI full bank (T-score)	58.1 (6.9)	58.1 (6.6)	58.2 (6.8)	0.648
DASH (0-100)	31.6 (16.5)	31.3 (16.3)	31.9 (16.8)	0.877
HIT-6 (36-78)	60.2 (7.5)	60.4 (6.9)	59.9 (8.6)	0.841
LEFS (0.80)	55.0 (15.8)	55.7 (15.4)	53.8 16.6)	0.451
NDI (0-50)	13.1 (7.2)	13.7 (7.0)	12.0 (7.3)	0.069
RDQ (0-24)	8.9 (5.3)	9.0 (5.3)	8.7 (5,3)	0.380

Table 2; Change scores of all instruments across GRoC response categories.

PROM				PROMIS-PI		RDQ		NDI		LEFS		HIT-6		DASH
	N	FB	SF8	CAT	N		N		N		N		N	
Range		T-score	T-score	T-score		0-24		0-50		0-80		36-78		0-100
GRoC category														
1 Much improved	269	15.1 (8.8)	14.4 (8.6)	15.4 (9.6)	144	7.1 (5.8)	45	8.0 (5.9)	29	-13.1 (21.6)	8	11.9 (8.0)	14	19.7 (17.8)
2 Improved	256	6.0 (6.8)	6.3 (6.7)	6.4 (7.6)	140	3.5 (4.7)	37	4.5 (4.5)	24	-13.0 (12.2)	10	5.2 (7.3)	14	17.4 (17.1)
3 Little improved	145	2.1 (4.8)	2.2 (5.5)	2.4 (5.7)	71	1.3 (3.9)	37	1.5 (5.0)	14	-1.1 (8.4)	5	2.0 (2.9)	5	10.8 (20.0)
4 Unchanged	210	1.1 (4.6)	1.4 (5.3)	1.4 (5.9)	107	0.2 (3.6)	28	0.3 (4.7)	24	-1.8 (7.2)	9	0.3 (1.9)	17	4.1 (12.2)
5 Little worse	18	-0.4 (4.3)	0.4 (5.6)	-0.0 (6.0)	8	1.0 (4.1)	7	-1.3 (2.5)	2	6.5 (12.0)	0	nn	0	nn
6 Worse	9	-0.5 (4.7)	0.3 (5.4)	-1.4 (4.8)	4	-2.5 (3.4)	1	nn	1	nn	0	nn	0	nn
7 Much worse	2	-4.3 (3.1)	-6.2 (6.2)	-6.0 (1.5)	2	1.0 (8.5)	0	nn	0	nn	0	nn	0	nn

Table 3; Mean baseline scores (SD), mean follow-up scores, mean change (SD), effect size and standardized response mean of all instruments.

PROM	N	Baseline	SD	Follow-up	SD	change	SD	ES	SRM
PROMIS-PI Full Bank	960	58.1	6.6	51.6	9.7	6.4	9.1	0.97	0.71
PROMIS-PI SF8	960	59.0	7.0	52.6	9.2	6.4	8.9	0.92	0.72
PROMIS-PI CAT	960	59.1	6.8	52.3	9.8	6.7	9.7	1.00	0.69
RDQ	496	9.0	5.3	5.7	5.4	3.3	5.4	0.62	0.61
NDI	167	13.7	7.0	10.0	7.5	3.7	5.9	0.53	0.63
LEFS	98	55.7	15.4	63.7	16.6	-8.0	15.2	0.52	0.52
HIT-6	35	60.4	6.9	56.3	8.8	4.1	7.5	0.59	0.55
DASH	52	31.4	16.3	17.1	15.9	13.8	18.0	0.85	0.76

Table 4; Correlations between the PROMIS-PI item bank and legacy instruments.

PROM	AUC	Correlation with GRoC	Correlation PROMIS with legacy instruments		
			PROMIS-PI Full bank	PROMIS-PI SF8a	PROMIS-PI CAT
PROMIS-PI Full bank	0.77	-0.63			
PROMIS-PI SF8a	0.75	-0.60			
PROMIS-PI CAT	0.74	-0.57			
RDQ	0.74	-0.48	0.68	0.65	0.60
NDI	0.74	-0.55	0.59	0.60	0.58
LEFS	0.77	0.53	-0.45	-0.50	-0.38
HIT-6	0.73	-0.68	0.59	0.58	0.61
DASH	0.71	-0.34	0.67	0.69	0.58

Table 5; Summary of responsiveness hypotheses.

	PROMIS-PI full bank		PROMIS-PI SF-8		PROMIS-PI CAT	
Correlation GRoC	-0.63	V	-0.60	V	-0.57	V
Area Under the Curve	0.77	V	0.75	V	0.74	V
Correlation legacy instruments						
RDQ	0.68	V	0.65	V	0.60	V
NDI	0.59	V	0.60	V	0.58	V
LEFS	-0.45	X	-0.50	V	-0.38	X
HIT-6	0.59	V	0.58	V	0.61	V
DASH	0.67	V	0.69	V	0.58	V
Comparison ES						
RDQ	0.35	V	0.30	V	0.38	V
NDI	0.44	V	0.39	V	0.47	V
LEFS	0.45	V	0.40	V	0.48	V
HIT-6	0.38	V	0.33	V	0.41	V
DASH	0.12	V	0.07	V	0.15	V
Comparison SRM						
RDQ	0.10	V	0.11	V	0.08	V
NDI	0.08	V	0.09	V	0.06	V
LEFS	0.19	V	0.20	V	0.17	V
HIT-6	0.16	V	0.17	V	0.14	V
DASH	-0.05	V	0.04	V	-0.07	X