Estimating minimally important differences for the PROMIS® Pain Interference Scales: results from three randomized clinical trials

Chen X. Chen, PhD, MB, MS,
Indiana University School of Nursing

Kurt Kroenke, MD,
Indiana University School of Medicine, Regenstrief Institute, Inc., VA Health Services Research and Development Center for Health Information and Communication

Timothy E. Stump, MA,
Indiana University School of Medicine, Department of Biostatistics

Jacob Kean, PhD,
University of Utah School of Medicine Department of Population Health Sciences, Salt Lake VA Health Care System Decision-Enhancement and Analytic Sciences Center

Janet S. Carpenter, PhD, RN, FAAN,
Indiana University School of Nursing

Erin E. Krebs, MD, MPH,
Minneapolis VA Center for Chronic Disease Outcomes Research, University of Minnesota Department of Medicine

Matthew J. Bair, MD, MS,
VA Health Services Research and Development Center for Health Information and Communication, Indiana University School of Medicine, Regenstrief Institute, Inc

Teresa M. Damush, PhD, and
Indiana University School of Medicine, Regenstrief Institute, Inc., VA Health Services Research and Development Center for Health Information and Communication; Precision Monitoring for Quality Improvement (PRIS-M QUERI Center)

Patrick O Monahan, PhD
Indiana University School of Medicine, Department of Biostatistics

Corresponding Author: Chen X. Chen, 600 Barnhill Drive E415, Indianapolis, IN 46202, cxchen@iu.edu, Tel: (317) 274-7441.

1. Introduction

The most accurate and reliable method of measuring pain is self-report, making validation of patient-reported pain outcome measures critical to both research and clinical care.[8] Although many pain measures have been validated, actual adoption of a measure in research and clinical care depends heavily on practical aspects such as brevity, public domain accessibility, and appropriateness for the settings.[8; 16] To fulfill these pragmatic criteria, the National Institutes of Health developed the Patient-Reported Outcomes Measurement Information System (PROMIS)®, which includes several pain measures.

The PROMIS-Pain Interference (PROMIS-PI) scales measure the extent to which pain interferes with physical, mental, and social activities.[1] These scales were developed based on item response theory,[1] which allows for both computerized adaptive testing (CAT) and fixed-length scales with small numbers of informative items to minimize response burden. In addition to the CAT version, 4 fixed-length PROMIS-PI scales are available for adults: one with 4 items, two with 6 items, and one with 8 items.

Although the PROMIS-PI scales have demonstrated reliability and validity across diverse populations,[1; 3] enhanced interpretability of the PROMIS-PI scales is needed to support their usefulness in clinical trials and patient care. One important aspect to improve interpretability is minimally important difference (MID), defined as "the smallest difference in score in the domain of interest that patients perceived as important, either beneficial or harmful, and that would lead the clinician to consider a change in the patient's management."[12 (p. 377)] Estimates of MIDs can help researchers, clinicians, and policy makers better interpret the magnitude of treatment effects and can provide a metric to calculate statistical power.[18]

Three studies have estimated the MID for adult PROMIS-PI scales[7,2; 26] and contributed to the limited knowledge on PROMIS-PI interpretability, but gaps still exist. First, MIDs are often context-specific and can vary by populations.[19] It is essential to obtain MIDs from various samples to evaluate convergence. The previous studies estimated MIDs with either one disease population[7; 26] or with data pooled from two disease populations (i.e., low back pain and depression) rather than analyzed separately.[2] Second, triangulation of methods for MID estimation is still needed, as each method has strengths and weaknesses.[9; 18] Yost et al.[26] used method triangulation but focused on PROMIS-Cancer scales with 10 pain interference items (i.e., different from non-cancer scales). Third, MID estimates from randomized clinical trials (RCTs) are still needed, as MIDs derived from RCTs may differ from those estimated through observational studies.[18] No previous studies used data from RCTs. Fourth, it is unknown whether MIDs derived from fixed-length scales that vary in number of items are similar, because the prior studies focused on only one version.

The study purpose was to estimate MIDs for the 4 fixed-length PROMIS-PI scales. Fixed-length scales rather than CAT administration were chosen because in many clinical and research settings fixed-length scales are more feasible to administer, which is why they have been offered as a viable option by PROMIS developers. We contribute to the literature by separately analyzing 3 clinical samples from RCTs, administering 4 fixed-length scales, and triangulating methods for MID estimation.

2. Methods

2.1. Design and Participants

In this psychometric study, data were analyzed from three RCTs conducted between 2012 and 2017 with 759 patients. Sample 1 consisted of 261 primary care patients participating in an RCT to compare the effectiveness of pharmacological versus behavioral approaches for chronic low back pain (NCT01236521). Sample 2 consisted of 240 primary care patients participating in a pragmatic RCT comparing opioid therapy versus non-opioid medication therapy for chronic back pain or hip or knee osteoarthritis pain (NCT01583985). Sample 3 consisted of 258 stroke survivors participating in an RCT evaluating the efficacy of a stroke-self-management program (NCT01507688).

2.2. Procedures

The study was approved by the Indiana University Institutional Review Board. For each RCT, informed consent was obtained from all participants. The participants completed the questionnaires at baseline and follow-up. Follow-up assessments were conducted 6 months after baseline for Sample 1 and 3 months after baseline for Samples 2 and 3. Demographic and clinical characteristics were captured at baseline through self-report. Patient-reported outcome data were collected from interviews administered by trained research personnel.

2.3. Measurement

2.3.1. PROMIS-Pain Interference (PI) fixed-length scales—Participants completed the following 4 fixed-length PROMIS-PI scales: the 6-item original short form (6b), and the 4-, 6-, and 8-item scales (4a, 6a, 8a) that are part of the PROMIS adult profile instruments (a collection of short forms containing a fixed number of items from key PROMIS domains).[4] Response formats for all scales were a 5-point ordinal rating scale of "Not at all," "A little bit," "Somewhat," "Quite a bit," and "Very much." Raw score totals on each scale were converted to an item response theory-based T-score using the PROMIS scoring manual.[4] (More information can be found at www.healthmeasures.net.) T-scores allow for comparing scores between PROMIS-PI scales with different lengths and comparing scores to the population norm. In this paper, all PROMIS-PI scores were reported in the T-scores metric. A T-score of 50 is the average for the US general population with a standard deviation (SD) of 10.[1] A higher T-score represents higher pain interference. The reliability and validity of the PROMIS-PI scales have been well supported.[3] For the current samples, Cronbach's alphas for PROMIS-PI raw scores at baseline ranged from 0.88 to 0.97.

2.3.2. The Brief Pain Inventory Interference (BPI-I) Scale—The BPI is among the most extensively used pain scales in clinical research.[5] The 7-item BPI interference (BPI-I) scale measures pain interference on mood, physical activity, work, social activity, relations with others, sleep, and enjoyment of life. This scale is conceptually comparable to the PROMIS-PI measures. Each BPI-I item is scored from 0="Does not interfere" to 10="Completely interferes," and the scale score is the mean of the 7 items.[5] Scores range from 0 to 10 with higher scores indicating greater pain interference. The reliability and validity of the BPI are well-established.[5] For the current samples, the Cronbach's alphas for BPI-I ranged from 0.85 to 0.94.

2.3.3. Disability Days—A single item used in several previous studies[15; 21] assessed the number of patient-reported disability days due to pain. "During the past 4 weeks, how many days did you cut down on the things you usually do for one-half day or more because of problems with pain?" Disability days were coded into 4 ordinal categories: <7 days, 7–14 days, 15–21 days, and 22–28 days. Each ordinal category is considered a clinically-distinct group.[21]

2.3.4. Cross-sectional Global Ratings of Pain—The cross-sectional global rating of pain assesses patient pain on average in the past 7 days. Following the approach developed by Yost et al.,[26] a 5-point ordinal scale ranging from 0 = "no pain" to 4 = "very severe pain" was used. Each ordinal category is considered a clinically-distinct group.[26]

2.3.5. Retrospective Global Ratings Change (RGRC)—The RGRC assesses the overall clinical response as judged by the participant.[8] At follow-up, participants rated their pain change compared to their baseline pain. Response options ranged from −3 = "very much worse," to +3 = "very much better," with 0 representing no change (7 options in total). The RGRC is widely used as an outcome measure in chronic pain clinical trials[8] and is commonly used to establish MIDs for patient-reported pain scales.[9; 26]

2.4. Data Analysis

Data for each participating RCT were analyzed separately rather than pooled because the 3 RCTs involved different clinical populations, interventions, and follow-up time frames. Data analyses were carried out using SAS software (version 9.4, SAS Institute, Cary, NC, 2002–2015).

We estimated the MIDs for the PROMIS-PI scales by triangulating distribution- and anchor-based methods as suggested in the literature.[12; 25; 26] Distribution-based methods are based on the statistical distribution of the measures, while anchor-based methods are based on external criteria (anchors) that are clinically meaningful.[18]

2.4.1. Distribution-based methods—Two established distribution-based methods were used: (1) effect size and (2) standardized error of measurement (SEM). For effect size, we calculated 0.2 SD, 0.35 SD, and 0.5 SD of baseline PROMIS-PI scores. Because 0.2 SD approximates a small effect size,[6] score differences less than 0.2 SD are likely to have less than a minimally important difference.[11] Because 0.5 SD approximates a medium effect size, score differences significantly above 0.5 SD are likely to have more than a minimally important difference. A score difference between those boundaries (e.g., 0.35 SD) can be a good approximation of a MID.[11]

Standardized error of measurement (SEM) was calculated using baseline PROMIS-PI scores.[23; 24] The SEM indicates the precision of the outcome measure and can be interpreted as the smallest difference likely to reflect a true difference or change rather than a measurement error.[18] In the item response theory framework, each participant has a standard error associated with that individual's T-score. The SEM for each sample was obtained by averaging the individuals' standard errors across the sample.[10; 13] Specifically, the square root of the mean of variance (i.e., standard error squared) for each T score across persons in

the sample was computed to derive the sample SEMs. The literature suggests that 1 SEM corresponds closely with anchor-based MIDs for health-related quality of life measures.[23; 24] Depending on the context, 2 SEMs can also be an appropriate approach to estimate the MID. [22] One SEM and 2 SEMs respectively correspond to 68% and 95% confidence interval bands around individual scores.[17; 22] To reconcile different recommendations, we decided that the final MIDs should be neither notably lower than one SEM nor notably higher than two SEMs.

2.4.2. Anchor-based methods—Anchor-based methods map PROMIS-PI score differences onto differences in clinically meaningful anchors. The clinical anchors share a conceptual similarity to PROMIS-PI. One factor in evaluating an anchor was the correlation between the score on the anchor measures and the PROMIS-PI score. Pearson correlations 0.3 indicated that the anchor might be a stronger measure for estimating a MID. In contrast, MID estimates derived from anchors that had lower correlations with the PROMIS-PI should be interpreted more cautiously.[18; 26] We performed both cross-sectional anchor-based analysis and longitudinal anchor-based analysis as described below.[11; 26]

<u>2.4.2.1. Cross-sectional anchor-based analyses:</u> Cross-sectional analyses address minimally important between-individual differences. In these analyses, PROMIS-PI scores within each time point were mapped onto clinically meaningful anchors.

We used BPI-I as the cross-sectional anchor given its conceptual similarity to the PROMIS-PI and the known information about its MIDs. In our study, correlations between PROMIS-PI and BPI-I scores ranged from 0.63 to 0.85. One point on the BPI-I scale represents a MID,[18] so we sought to estimate the PROMIS-PI score that corresponded to 1 point change on the BPI-I scale. Using linear regression, we regressed the PROMIS-PI scores on the BPI interference scores. The linearity assumption was confirmed by inspecting scatter plots.

In addition, we conducted supplementary analyses using two less established anchors: the Cross-sectional Global Ratings of Pain (correlations with PROMIS-PI: 0.37 – 0.82.) and Disability Days due to pain (correlations with PROMIS-PI: 0.46 to 0.61). Participants were divided into 5 distinct categories based on global ratings response categories: no pain, mild pain, moderate pain, severe pain, and very severe pain. Then, participants were divided into 4 distinct categories based on disability days response categories: <7 days, 7–14 days, 15–21 days, and 22–28 days.[21] Because these two anchors are infrequently cited in the MID literature, it is less clear if the calculated score differences represent minimally important differences.

<u>2.4.2.2. Longitudinal anchor-based analyses:</u> While cross-sectional analyses address minimally important between-individual differences, longitudinal anchor-based analyses address minimally important changes based on within-individual change scores. In these analyses, changes in the PROMIS-PI scores (from baseline to follow-up) were mapped onto global pain changes, which were determined both retrospectively and prospectively.[26]

The RGRC score collected at follow-up was used as the retrospective anchor. Correlations between PROMIS-PI change scores and the RGRC ranged from 0.23 to 0.49. Participants

were divided into 7 distinct categories based on RGRC: "much better," "moderately better," "a little better," "no change," "a little worse," "moderately worse," and "much worse." PROMIS-PI change scores corresponding to one category shift (e.g., between "no change" and "a little worse," or between "a little worse" and "moderately worse") were used as MID estimates.[26]

The prospective change in global rating of pain was used as the prospective anchor. To calculate an individual's prospective change in global rating, we subtracted the individual's follow-up global rating of pain from his or her baseline global rating of pain. In our study, the correlations between the PROMIS-PI change scores and prospective change in global rating of pain scores ranged from 0.26 to 0.64. Since the cross-sectional global rating of pain is on a 5-point scale ranging from 0 ("No pain") to 4 ("Very severe pain"), change scores had a possible range of −4 to +4, where negative numbers indicated worsening pain and positive numbers improved pain. For example, a patient with severe pain at baseline and mild pain at follow-up had a +2 change (3 *minus* 1), whereas a patient who had moderate pain at baseline and severe pain at follow-up had a −1 change (2 *minus* 3). We considered a 1-point change in the negative (worsened) or positive (improved) direction as clinically meaningful.[26] Mean changes in the PROMIS-PI T-score corresponding to a 1-point change were considered as estimates of the MID.[26]

2.4.3. Methods reconciliation—Since MIDs estimated by different methods can differ,[9] the final recommended MIDs were derived from considering various distribution- and anchor-based methods. We attempted to identify a range of MID estimates as opposed to a fixed value.[8] Among the various anchors, we considered the proximity of the anchor to the PROMIS-PI scales and the anchors' level of acceptance, as discussed in the literature. We prioritized two types of anchors: (1) those that share conceptual similarity with PROMIS-PI scales and are considered a legacy pain interference scale (i.e., BPI-I); and (2) those that are widely accepted in the literature (i.e., retrospective global ratings of change, prospective change in global ratings of pain). Distribution-based estimates were used to set approximate bounds of MID estimates. The final MIDs should not be notably lower than a 0.2 effect size and one SEM to ensure that the MIDs exceed both a trivial difference and the measurement error. At the same time, the final MID estimates should not be notably higher than a 0.5 effect size or two SEMs to ensure that the difference is minimally important as opposed to moderately or substantially important. All PROMIS-PI MIDs were reported in T-score metric.

3. Results

3.1. Sample Characteristics

A total of 759 participants completed the baseline and follow-up assessments. Table 1 summarizes the sample characteristics. For all 3 samples, participants were mostly male, non-Hispanic, white, married, had some college education, and had just enough income to make ends meet. At baseline, Samples 1 and 2 had worse pain than the US population norm (i.e., mean PROMIS-PI scores were one SD above the population norm of 50), while Sample

3 had a pain interference level close to the US population norm (i.e., mean PROMIS-PI scores were within 1/3 SD above the US population norm).

3.2. Distribution-based Estimates

We considered a 0.35 effect size and one SEM together for distribution-based MID estimates. As shown in Table 2, the distribution-based MID estimates for Samples 1 and 2 (pain samples) were between 1.5 and 2.5 points. For Sample 3 (non-pain sample), the distribution-based MID estimates were between 3 and 4 points.

3.3. Anchor-based Estimates

3.3.1. Cross-sectional anchor-based estimates—As shown in Table 2, for Samples 1 and 2, each 1-point difference on the baseline BPI-I score (i.e., MID) corresponded to about a 2-point difference in the PROMIS-PI score. For Sample 3, each 1-point difference on the BPI-I corresponded to about a 3-point difference in the PROMIS-PI scores.

3.3.2. Longitudinal anchor-based estimates (i.e., minimally important change)
—Longitudinal anchor-based estimates are also listed in Table 2. For Samples 1 and 2, the minimally important change estimates ranged from 2.0 to 3.5 points. For Sample 3, the minimally important change estimates ranged from 3.0 to 4.5 points.

3.3.3. Secondary anchor-based estimates—When Cross-sectional Global Ratings of Pain and Disability Day were used as anchors, the PROMIS-PI difference scores were 3.0 to 4.0 points for Samples 1 and 2, and 5.0 to 7.0 points for Sample 3. (See supplemental material.) These differences were notably beyond the 0.5 effect size.

3.4. Summary of MID Estimates across Methods, Samples, and Four Fixed-Length Scales

The distribution and anchor-based MID estimates are plotted in Figure 1. The final MID recommendation was derived from combining the distribution- and anchor-based methods. The MID estimates were rounded to the nearest half-integer to inform the recommended MIDs ranges.[26] Taken together, the MID estimates for the two pain-specific samples (Samples 1 and 2) ranged from 2.0 to 3.0 points (i.e., in the vicinity of 2.5 points). The MID estimates for the non-pain specific sample (Sample 3) ranged from 3.5 to 4.5 points (i.e., in the vicinity of 4 points).

3.5. MID Estimates across Four Fixed-Length PROMIS-PI Scales

Across the 4 fixed-length PROMIS-PI scales, the MID estimates were largely comparable (See Figure 2). We observed no particular pattern for the MIDs estimates with SEM estimates being the only exception: as expected, the longer the scale, the smaller the SEM estimate. In other words, the longest PROMIS-PI scale (i.e., 8-item) consistently had the smallest SEMs, while the shortest PROMIS-PI scale (i.e., 4-item scale) consistently had the highest SEMs. However, the difference between the SEM-based estimates for the scores were within 0.6 points. Because the MID estimates were quite similar regardless of the scale length, the MIDs reported in Table 2 and Figure 1 are the averages of the estimates across the 4 fixed-length scales.

4. Discussion

In this study, we triangulated multiple distribution- and anchor-based methods to establish MIDs for PROMIS-PI scales. We estimated MIDs with 3 different samples and compared MID estimates for the PROMIS-PI scales of different lengths. Based on our findings, the distribution-and anchor-based MID estimates showed convergence. For the pain samples, the MIDs ranged from 2.0 to 3.0 points, and for the non-pain sample, the MIDs ranged from 3.5 to 4.5 points. The MID estimates were comparable across the 4 fixed-length PROMIS-PI scales.

To our knowledge, this was the first study to estimate MIDs for PROMIS-PI scales with various clinical samples. By separately analyzing data from 3 RCTs with different clinical samples, we uncovered the sample-dependence issue related to PROMIS-PI MID estimates. As methodologists have suggested, MIDs are often context-specific and can vary by population.[19] For a measurement tool, there is not necessarily a single MID value that is appropriate across various applications. We found that the MID estimates were smaller for the pain samples than the non-pain sample (i.e., stroke survivors). These findings may be explained by several factors. First, the pain and non-pain samples differed in the baseline level of pain interference. On average, the patients in two pain samples had a pain interference level of 62, while the patients in the non-pain sample had a pain interference level of 53. The PROMIS-PI scales are most precise at the 60–65 level.[4] The smaller measurement error when the scores were close to 62 explains smaller SEM estimates for the two pain samples. Second, our non-pain sample was more heterogeneous than the pain samples in terms of pain severity and interference. For the two RCTs of chronic pain, having moderate to severe pain was part of the inclusion criteria, whereas for the RCT of stroke survivors, having pain was not part of the inclusion criteria. The heterogeneity was reflected by larger SDs for PROMIS-PI scores for the non-pain sample than the pain samples. Because SD is a multiplication term to calculate the effect-size based estimates, the larger SDs led to larger effect-sized based estimates for the non-pain sample. Another factor was that pain in the two pain trials was chronic and musculoskeletal whereas pain in the stroke sample may have varied more in type, duration, and location, again contributing to greater variability in MID estimates.

This study was also the first in which MIDs were estimated among all fixed-length adult PROMIS-PI scales. Four short forms are available for adult PROMIS-PI scales.[4] Researchers and clinicians face the decisions of choosing between these scales with different lengths. We found that the 4 fixed-length scales had highly comparable MID estimates. The SEM-based estimates were the only exception: the longer the scale, the smaller the SEM estimate, which makes sense because reliability and precision generally increase as scale length increases. Among the 4 fixed-length scales, the 8-item version was most precise, and therefore it had the smallest SEM estimates. For a given study, when the precision of a measure is a priority, the 8-item scale is ideal. However, the differences in SEM-based estimates were small. Also, other types of MID estimates were almost identical across the 4 fixed-length scales. Therefore, use of the shorter versions would be reasonable based on other factors such as respondent burden.

Our MID estimates for the two pain samples are close to the 2 point MID estimated by Deyo and colleagues[7] in which the 4-item PROMIS-PI scale was evaluated among older adults with chronic musculoskeletal pain. According to Deyo et al.,[7] a 2-point difference can be considered to be an MID, as it corresponded to a difference between "slightly better" and "much better" (or "slightly worse" and "much worse"). Our study addressed a limitation of Deyo et al. by incorporating legacy measures and triangulating multiple methods. Our MID estimates in the two pain samples are also consistent with MIDs in other studies for the PROMIS physical function scales (i.e., 2 points)[14] and PROMIS pediatrics pain interference form (i.e., 2.0 to 3.0 points).[20]

However, our MID estimates for the two pain samples are smaller than those reported in two earlier studies. In a study of the PROMIS cancer scales by Yost et al.,[26] 4.0 to 6.0 points were reported as the MIDs for the 10-item Cancer Pain Interference Scale. In a study by Amtmann and colleagues combining patients with low back pain and patients with depression,[2] 3.5 to 5.5 points were reported as the MID range for the PROMIS-PI CAT version. Differences in MID estimates between our study and these pioneer studies can be explained by two factors. The first factor has to do with the sample differences. Compared to our pain samples, the samples in Yost et al. and Amtmann et al. were more heterogeneous in terms of pain interference (i.e., higher SDs), and they had lower pain interference scores on average. Their MID estimates were closer to the MID estimated in our non-pain sample (i.e., 3.5 to 4.5). The second factor has to do with methodological differences. For distribution-based estimates, Yost et al. and Amtmann et al. used 0.8 SD as the upper bound of their MID estimates, whereas we consider that MIDs should not be notably higher than 0.5 SD (a moderate effect size).[9] In addition, Yost et al. and Amtmann et al. used anchors that were slightly different from ours. They both used 0.5 SD of the BPI-I as an MID anchor, whereas we followed the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) recommendation[9] by using a 1-point BPI-I as the MID anchor. Some anchor categories they used were also wider than the categories we used. For example, Yost et al. combined "a little better" and "moderately better" into one category, whereas we treated those as two clinically distinct categories for the MID estimation. Some of the MID estimates by Yost et al. and Amtmann et al. may be moderately important, or in other words, more than minimally important.

Our study has several strengths. First, the MIDs were estimated separately in 3 RCTs with various samples. Each individual RCT was large enough (sample sizes: 200–250) for psychometric analysis on its own. Analyzing data separately for each RCT allowed us to identify whether there were sample-dependence issues related to MID estimates. Second, to estimate MIDs, we triangulated multiple methods including distribution-based methods, anchor-based methods, analysis of cross-sectional data, and analysis of longitudinal data. No single method for MIDs estimation is without limitations, so method triangulation made our conclusions more robust. Third, we estimated MIDs for 4 adult fixed-length PROMIS-PI scales, which can inform scale selection for researchers and clinicians.

We acknowledge there were several study limitations. First, the sample size was small in a few clinically distinct anchor categories, particularly the cross-sectional global rating of pain categories. We omitted subgroups with a sample size less than 10 in MID estimations, as

estimates based on fewer observations would be unstable.[26] Second, study participants were largely men (81% to 92%). Third, the follow-up times differed across the three studies (3 months or 6 months). Fourth, controversies related to certain anchors exist. Although the retrospective global rating of pain (RGRP) is widely used in the MID literature, it can be subject to recall and reconstruction bias. We supplemented the RGRP measure by prospectively measuring cross-sectional global ratings of pain so we could calculate the prospective change in global rating of pain. However, this prospective change has not been widely cited in the MID literature, and additional research is needed to determine the optimal methods to measure global change of pain.

This study has implications for future research. First, although we estimated MIDs in several clinical groups, it would be useful to expand the evaluation of the PROMIS-PI to additional subgroups, such as samples with specific pain conditions (e.g., headache, visceral pain), with different durations of pain, and with a broader distribution of pain interference. Second, further evaluation of the quality of anchors (i.e., RGRC, disability days, cross-sectional global rating)[6] is still needed. Some anchors may capture clinically important differences that are meaningful but beyond minimally important. Using those anchors may have a risk of over-estimating the MID.[18]

In conclusion, we established a range of MIDs for 4 fixed-length PROMIS-PI scales. Researchers and clinicians should consider the sample heterogeneity and level of pain interference when choosing a specific MID number. The estimated MIDs can be used to interpret research data and guide clinical decisions. The MID estimates can also inform power calculations for efficacy and effectiveness studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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. 2010; 150(1):173–182. [PubMed: 20554116]

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. 2016; 9:251–255. [PubMed: 27175093]

3. Askew RL, Cook KF, Revicki DA, Cella D, Amtmann D. Evidence from diverse clinical populations supports clinical validity of PROMIS pain interference and pain behavior. J Clin Epidemiol. 2016

4. Cella D, Gershon R, Bass M, Rothrock N. Pain Interference: A brief guide to the PROMIS Pain Interference instruments. 2015

5. Cleeland, C., Ryan, K. Pain assessment: global use of the Brief Pain Inventory. Vol. 23. Annals of the Academy of Medicine; Singapore: 1994. p. 129-138.

6. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.

7. Deyo RA, Ramsey K, Buckley DI, Michaels L, Kobus A, Eckstrom E, Forro V, Morris C. Performance of a Patient Reported Outcomes Measurement Information System (PROMIS) Short Form in Older Adults with Chronic Musculoskeletal Pain. Pain Med. 2015

8. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005; 113(1–2):9–19. [PubMed: 15621359]

9. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, Haythornthwaite JA, Jensen MP, Kerns RD, Ader DN, Brandenburg N, Burke LB, Cella D, Chandler J, Cowan P, Dimitrova R, Dionne R, Hertz S, Jadad AR, Katz NP, Kehlet H, Kramer LD, Manning DC, McCormick C, McDermott MP, McQuay HJ, Patel S, Porter L, Quessy S, Rappaport BA, Rauschkolb C, Revickl DA, Rothman M, Schmader KE, Stacey BR, Stauffer JW, Von Stein T, White RE, Witter J, Zavislc S. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. Journal of Pain. 2008; 9(2):105–121. [PubMed: 18055266]

10. Embretson SE. The new rules of measurement. Psychol Assessment. 1996; 8(4):341–349.

11. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004; 57(9):898–910. [PubMed: 15504633]

12. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clinic proceedings. 2002; 77(4):371–383. [PubMed: 11936935]

13. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000; 38(9 Suppl):II28–42. [PubMed: 10982088]

14. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and Minimally Important Difference for the Patient-Reported Outcomes Measurement and Information System (PROMIS(®)) 20-Item Physical Functioning Short-Form in a Prospective Observational Study of Rheumatoid Arthritis. Ann Rheum Dis. 2015; 74(1):104–107. [PubMed: 24095937]

15. Kroenke K, Johns SA, Theobald D, Wu J, Tu W. Somatic symptoms in cancer patients trajectory over 12 months and impact on functional status and disability. Support Care Cancer. 2013; 21(3): 765–773. [PubMed: 22941116]

16. Kroenke K, Monahan PO, Kean J. Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. J Clin Epidemiol. 2015; 68(9):1085–1092. [PubMed: 25962972]

17. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res. 2011; 11(2):163–169. [PubMed: 21476818]

18. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology. 2008; 61(2):102–109. [PubMed: 18177782]

19. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes. 2006; 4(1):1–5. [PubMed: 16393335]

20. Thissen D, Liu Y, Magnus B, Quinn H, Gipson DS, Dampier C, Huang IC, Hinds PS, Selewski DT, Reeve BB, Gross HE, DeWalt DA. Estimating minimally important difference (MID) in PROMIS

pediatric measures using the scale-judgment method. Qual Life Res. 2016; 25(1):13–23. [PubMed: 26118768]

21. Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. Pain. 1992; 50(2):133–149. [PubMed: 1408309]

22. Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? J Biopharm Stat. 2004; 14(1):97–110. [PubMed: 15027502]

23. Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care. 1999; 37(5):469–478. [PubMed: 10335749]

24. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999; 52(9):861–873. [PubMed: 10529027]

25. Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: The FACIT experience. Evaluation & the Health Professions. 2005; 28(2): 172–191. [PubMed: 15851772]

26. 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(5):507–516. [PubMed: 21447427]
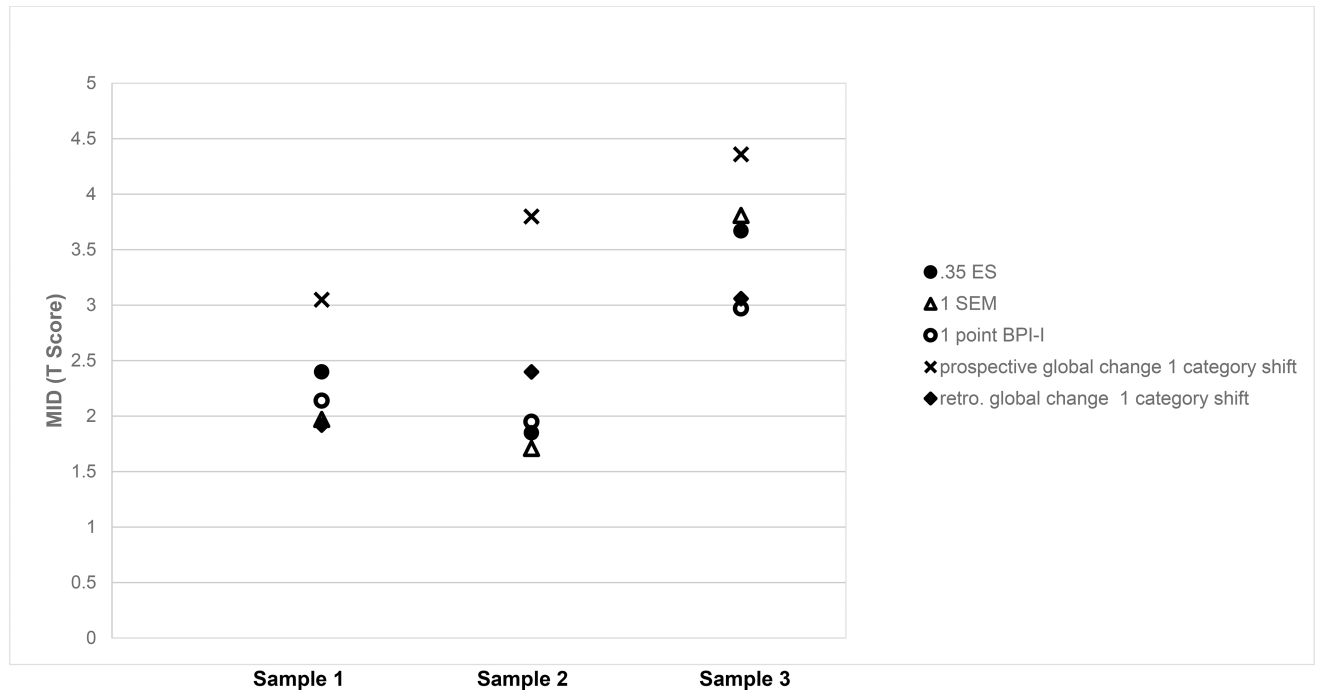
Figure 1.
Minimally Important Difference as Estimated Using Most Accepted Approaches
For cross-sectional anchor-based estimates, only estimates from baseline data are plotted.
The MID estimates in the figure were the average MIDs across the 4 PROMIS-PI measures,
as the 4 PROMIS-PI measures were comparable in MID estimates.
BPI-I: Brief Pain Inventory Interference Scale; ES: effect size (i.e., standard deviation of the
baseline scores); SEM: standard error of measurement; MID: minimally important
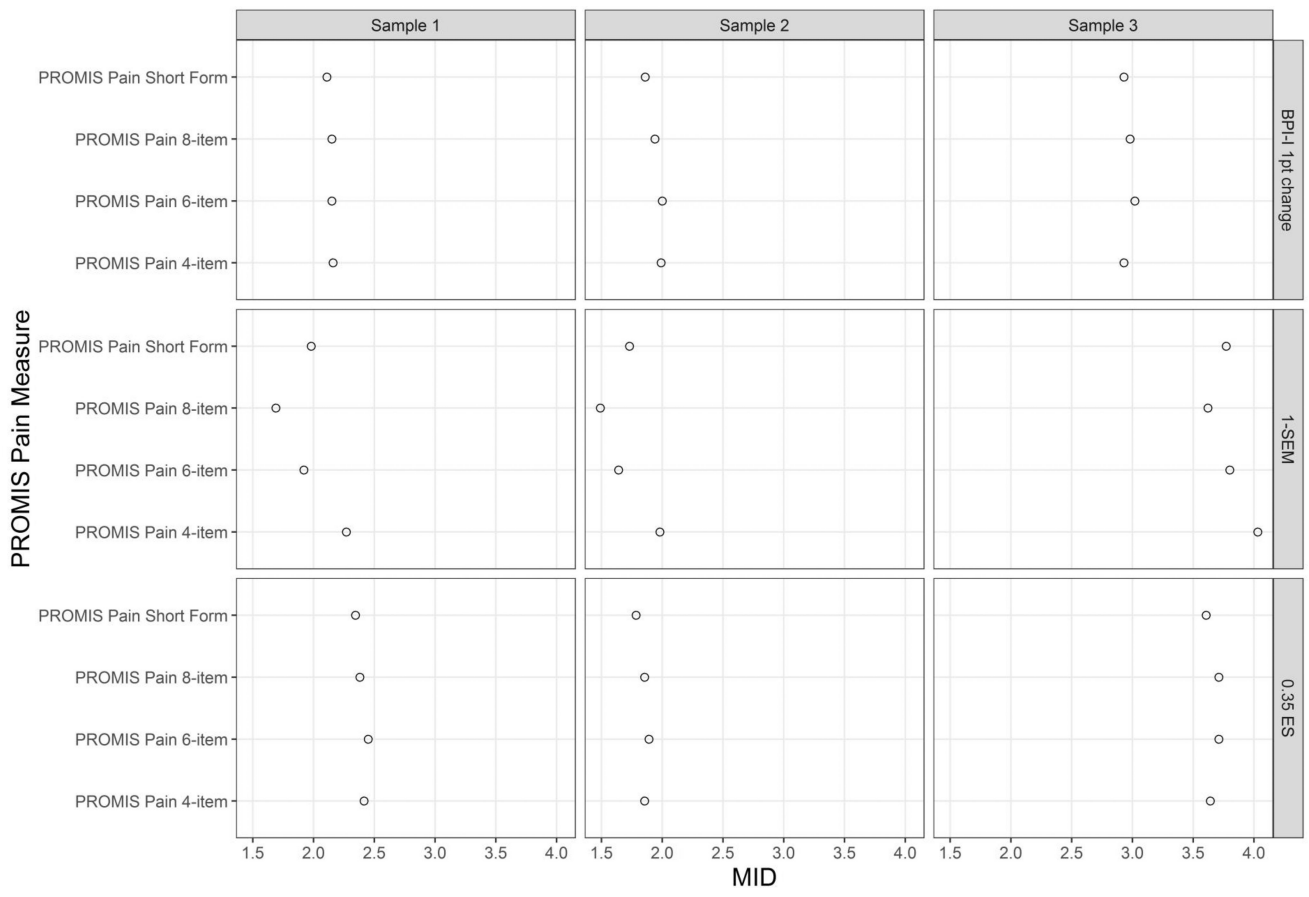difference; retro: retrospective

Figure 2.
Minimally Important Difference Estimates across Four PROMIS-PI Measures
MID: minimally important difference.

Table 1

Characteristics of Three Samples

	Sample 1 (N₁=261)		Sample 2 (N₂=240)		Sample 3 (N₃=258)	
Clinical Population	Chronic low back pain		Chronic musculoskeletal pain		Stroke survivors	
Recruitment Setting	Primary care		Primary care		Neurology	
Age, mean (SD)	57.9	(9.5)	58.3	(13.7)	61.7	(10.8)
Male, n (%)	241	(92.3)	208	(86.7)	209	(81.0)
Race, n (%)						
White	191	(73.2)	207	(86.2)	166	(64.3)
Black	54	(20.7)	18	(7.5)	78	(30.2)
Other	16	(6.1)	15	(6.3)	14	(5.4)
Education, n (%)						
Less than high school	17	(6.5)	6	(2.5)	31	(12.2)
High school	82	(31.4)	71	(29.6)	85	(33.3)
Technical school or some college	122	(46.8)	103	(42.9)	80	(31.4)
College degree or greater	40	(15.3)	60	(25.0)	59	(23.1)
Marital status, n (%)						
Married	139	(53.3)	135	(56.5)	135	(52.5)
Divorced	77	(29.5)	60	(25.1)	68	(26.5)
Other	45	(17.2)	44	(18.4)	54	(21.0)
PROMIS-PI T-scores, mean (SD)						
Pain 4-item	62.3	(6.9)	61.7	(5.3)	53.2	(10.4)
Pain 6-item	62.3	(7.0)	61.6	(5.4)	53.1	(10.6)
Pain 8-item	62.0	(6.8)	61.5	(5.3)	53.1	(10.6)
Pain short-form	62.1	(6.7)	60.9	(5.1)	53.2	(10.3)
Brief pain inventory interference (possible range: 0–10), mean (SD)	6.4	(2.1)	5.5	(1.9)	2.7	(3.0)
Cross-sectional Global Ratings of Pain (possible range: 0–4), mean (SD)	2.4	(0.7)	2.2	(0.6)	1.4	(1.2)
Disability days in the past 4 weeks, mean (SD)	16.4	(8.6)	10.3	(9.0)	5.1	(7.7)

Table 2

Minimally Important Difference (MID) for PROMIS Pain Interference T-Scores across Three Clinical Trials Using Different Methods

MID Estimation Method	Sample 1 (N₁=261, Chronic Pain)		Sample 2 (N₂=240, Chronic Pain)		Sample 3 (N₃=258, Stroke Survivors)		Average Across Samples
	Baseline	6 months	Baseline	3 months	Baseline	3 months	
Distribution-based analysis							
Effect size							
• 0.2 SD	1.37		1.06		2.10		1.51
• 0.35 SD	2.40		1.85		3.67		**2.64**
• 0.5 SD	3.43		2.64		5.24		3.77
1-SEM	1.97		1.71		3.81		**2.50**
Cross-sectional anchor							
BPI-I change (1 point)	2.14	2.40	1.95	2.48	2.97	2.81	**2.46**
Longitudinal anchors Prospective global change							
• One category improved		2.31		3.83		3.11	3.08
• One-category worsened		−3.78		−3.82		−5.61	−4.40
• Average one category shift		3.05		3.83		4.36	**3.74**
Retrospective global change							
• Moderate to much better		4.04		3.40		0.97	2.80
• A little to moderate better		1.02		3.24		2.96	2.41
• Same to a little better		2.14		0.66		−2.38	0.14
• Average one-category positive change		2.40		2.43		1.97	**2.27**
• Same to a little worse		−1.63		−1.11		−6.69	−3.14
• A little to moderate worse		−1.51		−3.63		−1.60	−2.45
• Moderate to much worse		−1.19		--		--	−1.19
• Average category negative change		−1.44		−2.37		−4.15	**−2.65**
• Average one-category shift		1.92		2.40		3.06	**2.46**

"--" indicates sample size for the subgroup <10

SD: Standard Deviation; SEM: Standard Error of Measurement; BPI-I: The Brief Pain Inventory Interference Scale

The MID estimates in the table are the average MID estimates across the 4 versions. The 4 versions were comparable in MID estimates. These was no particular pattern across the 4 versions, with SEMs being the only exception. The longer the PROMIS-PI version, the smaller the SEM. The differences between the largest and smallest SEMs were within 0.6 points.