Published in final edited form as:
Qual Life Res. 2013 September ; 22(7): 1859–1869. doi:10.1007/s11136-012-0323-3.

Psychometric characteristics of daily diaries for the Patient-Reported Outcomes Measurement Information System (PROMIS®): a preliminary investigation

Stefan Schneider,
Department of Psychiatry and Behavioral Science, Stony Brook University, Putnam Hall, South Campus, Stony Brook, NY 11794-8790, USA

Seung W. Choi,
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Doerte U. Junghaenel,
Department of Psychiatry and Behavioral Science, Stony Brook University, Putnam Hall, South Campus, Stony Brook, NY 11794-8790, USA

Joseph E. Schwartz, and
Department of Psychiatry and Behavioral Science, Stony Brook University, Putnam Hall, South Campus, Stony Brook, NY 11794-8790, USA

Arthur A. Stone
Department of Psychiatry and Behavioral Science, Stony Brook University, Putnam Hall, South Campus, Stony Brook, NY 11794-8790, USA

Stefan Schneider: Stefan.Schneider@StonyBrook.edu

Abstract

Purpose—The Patient-Reported Outcomes (PRO) Measurement Information System (PROMIS®) has developed assessment tools for numerous PROs, most using a 7-day recall format. We examined whether modifying the recall period for use in daily diary research would affect the psychometric characteristics of several PROMIS measures.

Methods—Daily versions of short-forms for three PROMIS domains (pain interference, fatigue, depression) were administered to a general population sample ($n = 100$) for 28 days. Analyses used multilevel item-response theory (IRT) models. We examined differential item functioning (DIF) across recall periods by comparing the IRT parameters from the daily data with the PROMIS 7-day recall IRT parameters. Additionally, we examined whether the IRT parameters for day-to-day within-person changes are invariant to those for between-person (cross-sectional) differences in PROs.

Results—Dimensionality analyses of the daily data suggested a single dimension for each PRO domain, consistent with PROMIS instruments. One-third of the daily items showed uniform DIF when compared with PROMIS 7-day recall, but the impact of DIF on the scale level was minor. IRT parameters for within-person changes differed from between-person parameters for 3 depression items, which were more sensitive for measuring change than between-person

Correspondence to: Stefan Schneider, Stefan.Schneider@StonyBrook.edu.

differences, but not for pain interference and fatigue items. Notably, mean scores from daily diaries were significantly lower than the PROMIS 7-day recall norms.

Conclusions—The results provide initial evidence supporting the adaptation of PROMIS measures for daily diary research. However, scores from daily diaries cannot be directly interpreted on PROMIS norms established for 7-day recall.

Keywords

PROMIS®; Patient-reported outcomes; Electronic diaries; Item response theory

Introduction

The Patient-Reported Outcomes (PRO) Measurement Information System (PROMIS) initiative has developed measures for a variety of quality of life domains, including physical functioning, fatigue, and emotional distress [1]. A goal of PROMIS has been to offer common metrics for the measurement of PROs to maximize comparability across studies and illnesses. Item banks for PRO domains were developed using state-of-the-art qualitative methods [2], calibrated using item response theory (IRT) to derive scales that are maximally reliable along the full spectrum of the latent trait, and scaled to a normative sample representing the general US population [3]. The utility of PROMIS as a foundation health measure for monitoring quality of life in the United States population is evaluated as part of the Healthy People 2020 initiative [4].

The majority of PROMIS measures ask patients to report about the “past 7 days”. This reporting period is desirable for many clinical settings as it allows capture of one’s clinically relevant experiences over a sufficiently long time interval with a single assessment [1, 2]. However, PRO measures that rely on extended recall periods are not capable of (1) characterizing the dynamic ebb and flow of everyday experiences or (2) detecting transient changes that occur on a day-to-day-basis, and (3) are likely subject to some degree of recall bias (including peak and recency effects). Daily diaries provide high-resolution information and can be uniquely useful for capturing the frequency and duration of acute symptom exacerbations [5] and for examining symptom trajectories shortly before and after medical treatments [6–8]. Furthermore, daily diaries limit the amount of retrospection in patients’ self-reports, which can increase the accuracy of data [9–11].

For many PROMIS domains, a daily diary format could be a valuable addition to the existing tools. However, it is not clear whether PROMIS measures can be directly translated into a daily format without affecting their psychometric properties. Evidence from the cognitive literature suggests that different recall periods can elicit disparate emotional and cognitive processes [12] and can change a respondent’s interpretation of the actual item content [13]. Thus, it is possible that ostensibly equivalent items do not measure the same construct in the same way when changing the recall period from 7 days to 1 day, a problem that may be thought of as lack of measurement invariance or differential item functioning (DIF) across recall periods.

A second important consideration is that PROMIS measures have been established based on cross-sectional data and calibrated for the measurement of *between-person* differences. On the other hand, diary studies commonly employ intensive longitudinal designs and focus on *within-person* changes in PROs. A fundamental—but often untested—assumption of many longitudinal analyses is that the psychometric properties of a measure generalize from the level of between-person differences to the level of within-person change [14, 15]. If different measurement models underlie within- and between-person sources or variation, this is known as a lack of “cross-level” invariance, that is, DIF across between- and within-person levels of measurement [16, 17].

In this report, we examine whether PROMIS measures can be modified for use in daily diary studies without affecting their psychometric properties. Short-forms of three PROMIS domains (pain interference, fatigue, and depression) were modified from a 7-day into a 1-day format and administered to a general population sample for 28 days. We evaluate measurement invariance across recall periods by comparing the psychometric characteristics of the daily items with the national item parameters from the PROMIS Version 1 item bank, and we evaluate cross-level measurement invariance by comparing psychometric characteristics of between- and within-person levels of diary data.

Methods

Measures

Details on the development and calibration of the PROMIS item banks are available for pain interference [18], fatigue [19], and depression [20]. PROMIS affords measurement via computerized adaptive testing (CAT) or by selecting any subset of items from the larger bank for administration as static short-form [1]. For this study, we created daily versions of PROMIS short-forms consisting of 6 (pain interference), 7 (fatigue) and 8 (depression) items (see Table 3 for item contents). Items were selected to represent the range of item content and difficulty of the larger banks, and were largely consistent with the PROMIS Version 1 short-forms (one pain interference and two fatigue items addressing events that may not occur on a daily basis were substituted by other calibrated items from the banks). The reporting period of each item was changed from “In the past 7 days…” to “In the last day…”. Response options (fatigue and depression: never, rarely, sometimes, often, always; pain interference: not at all, a little bit, somewhat, quite a bit, very much) were left unchanged.

Participants

One hundred participants from Eastern and Central US time zones were recruited for this study. They were selected so that their demographic composition (age, sex, race, and ethnicity) approximated the 2009 US Census. Eligibility criteria for participation were (1) age ≥21 years, (2) ability to make ratings each night for 28 consecutive days, (3) high-speed Internet access at home, (4) English fluency, (5) no visual impairment, (6) no night shift job.

Procedure

The study was approved by the Stony Brook Institutional Review Board. Recruitment was conducted using an Internet panel of 1.7 million respondents who regularly participate in online surveys (www.surveyspot.com). Panelists were invited to contact our research staff to be screened for eligibility. Eligible and interested participants were scheduled for a registration phone call to provide electronic consent and learn how to complete the daily ratings. Data were collected on the Internet via PROMIS Assessment Center^{SM} (http://www.assessmentcenter.net/). Participants were instructed to complete the daily ratings over 28 days prior to going to bed and before midnight. At the end of each week, they also completed the 7-day recall PROMIS measures for pain interference, fatigue, and depression, administered via CAT. Compliance was monitored daily, and participants were contacted if they missed an assessment. Participants received $150 for study completion.

Statistical methods

The PROMIS measures have been calibrated using Samejima’s [21] graded response model (GRM) [22]. For the present analyses, we used a multilevel extension of the GRM to account for the nesting of 28 days in 100 subjects and to discriminate measurement models for within- and between-person data. The GRM was estimated using multilevel factor analysis (FA) for ordinal response variables, as implemented with M*plus*, Version 6.11 [23]. The equivalence between the conditional probability formulation underlying IRT models and the latent response variable formulation underlying ordinal FA models is well-established [24].

An ordinal FA model is based on two components: a threshold model and a FA model. The threshold model relates a continuous latent response variable y^* to its observed categorical counterparts y for each item via threshold parameters τ. Given the categories $c = 0, 1, \ldots, C - 1$, the observed ordinal outcome $y = c$ if $\tau_c < y^* \leq \tau_{c+1}$, where $\tau_0 = -\infty$ and $\tau_C = +\infty$. A standard FA model then estimates factors representing the latent PRO measures from the continuous variables y^*. In multilevel FA, separate factors are extracted for within- and between-person levels of measurement. The value of y^* for item h of individual j on day i can be expressed as

$$y^*_{hij} = \mu_h + \left[\sum_{m=1}^{\mathrm{M_B}} \lambda^{\mathrm{B}}_{mh}\theta^{\mathrm{B}}_{mj} + \varepsilon^{\mathrm{B}}_{hj}\right] + \left[\sum_{m=1}^{\mathrm{M_W}} \lambda^{\mathrm{W}}_{mh}\theta^{\mathrm{W}}_{mij} + \varepsilon^{\mathrm{W}}_{hij}\right],$$

where $\mathrm{M_B}$ indicates the number of between-person factors θ^{B} with corresponding loadings λ^{B}, $\mathrm{M_W}$ indicates the number of within-person factors θ^{W} with loadings λ^{W}, μ is an item intercept, and the ε's are item specific errors [for details, see 25].

Two forms of "cross-level" measurement non-equivalence can be distinguished based on this model. First, if the number of between- and within-person factors ($\mathrm{M_B}$ and $\mathrm{M_w}$) is not the same, then the items address different constructs across measurement levels, indicating a lack of *dimensional invariance* [26]. Second, if the loadings λ^{B} and λ^{W} differ, then the constructs have different interpretations on the between- and within-person levels, denoting "cross-level" DIF [17, 25].

Similarly, different forms of measurement non-equivalence across recall periods can be examined by comparing the between-person parameters from the daily data with the parameters for 7-day recall established in PROMIS [1]: first, the number of factors may differ across recall periods, indicating a lack of *dimensional invariance*. Second, factor loadings may differ across recall periods, which would denote *nonuniform* DIF across recall periods [27]. Finally, item category thresholds (i.e., "difficulty" levels) may differ across daily and 7-day recall formats, indicating *uniform* DIF across recall periods [27].

Dimensional invariance across recall periods and measurement levels—The PROMIS measures have been shown to represent one-dimensional constructs [1, 18–20, 22]. Thus, the presence of more than one between- or within-person factor for the daily data would indicate a lack of dimensional invariance across recall periods or measurement levels. We used exploratory multilevel FA models for ordinal data to examine the between- and within-person dimensionality of the daily data for each domain. To evaluate how many factors were needed on each level, we inspected the factor eigenvalues and improvements in model fit when varying the number of factors on both levels [28]. Model fit was evaluated with the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), root mean square error of approximation (RMSEA), and standardized Root Mean Residual (SRMR). In prior work on PROMIS measures, acceptable levels of fit were suggested as CFI > 0.95, TLI > 0.95, and RMSEA < 0.06 [22].

DIF across recall periods and measurement levels—To examine evidence for DIF, one must specify a baseline model. Following prior recommendations, we used a free-baseline model, where the parameters of only one reference item are constrained for model identification [29, 30]. For the present analyses, we constrained the parameters of the reference item to the national item parameters established for PROMIS (http://www.nihpromis.org). The reference item was empirically selected using an iterative purification procedure, which compared the χ^2 fit between a model in which the parameters of *all* items were constrained to the PROMIS parameters and a series of models in which the parameters *for one focal* item at a time were freely estimated [29]. For each domain, the item with the largest *p* value from the χ^2 difference tests was selected as reference item [29].

To test DIF across recall periods, the parameters for all items (except for the reference item) were then freely estimated and compared with the corresponding PROMIS population parameters using Wald χ^2 tests. Specifically, uniform DIF across recall periods was examined by comparing the item thresholds (difficulty levels) with the PROMIS thresholds, and nonuniform DIF across recall periods was evaluated by comparing the between-person factor loadings with corresponding PROMIS loadings. To test for cross-level DIF, we compared the between- and within-person factor loadings for each item, again using Wald χ^2 tests. Because a series of DIF-tests were conducted, we controlled for multiple comparisons using the Benjamini-Hochberg correction [31].

Impact of DIF—In addition to testing the statistical significance of DIF, we also evaluated the magnitude and impact of DIF by comparing the characteristics of the daily measures from models that ignored DIF versus those accounting for DIF. First, we plotted the test characteristic curves (TCCs) to examine the impact of DIF on the expected scale scores (the sum of the expected item scores). Second, we examined the magnitude of between- and within-person correlations among the different PRO-domains before and after DIF-calibration. Third, we examined the "structural model parameters" (i.e., population mean, between- and within-person variances) of the daily PROs before and after DIF-calibration. In each case, the effect of DIF was evaluated by comparing results from models that held all item parameters fixed at the PROMIS population parameters versus those in which the parameters for all items except for the reference item were freely estimated.

Results

Sample characteristics and missing data

Table 1 shows the demographic characteristics of the study participants (*N* = 100). The sample was close to the PROMIS national norms in terms of their levels of pain interference, fatigue, and depression based on the PROMIS 7-day recall format, as evidenced in the averages of the 4 weekly administered PROMIS CATs: the sample means (SDs) in standard units were 0.13 (0.96) for pain interference, −0.12 (1.03) for fatigue, and −0.02 (0.90) for depression (where PROMIS measures are scaled with mean = 0 and SD = 1 in the general population).

Participants completed the web-based assessments on an average of 26.4 (SD = 3.64) of the 28 days. Out of 2,800 daily diaries across participants, 2,641 (94.3%) were completed. The rates of partially completed assessments (e.g., items skipped) were 0.8 % for pain interference, 1.1 % for fatigue, and 0.9 % for depression items.

Intraclass correlations for individual items

A basic requirement of multilevel FA is the presence of both between- and within-person variation. To test this, we obtained the intraclass correlation coefficient from a univariate multilevel ordinal regression model, separately for each item [25]. Intraclass correlations approaching 0 (no between-person variance) or 1 (no within-person variance) would suggest that multilevel FAs are unnecessary. The intraclass correlations ranged from 0.72 to 0.76 for the pain interference items, from 0.64 to 0.73 for the fatigue items, and from 0.64 to 0.77 for the depression items (*p*s < 0.001), indicating that multilevel FA was indicated.

Evaluation of dimensionality

Next, we examined the dimensionality of daily items. For each domain, the eigenvalues suggested a one-factor solution on both the between- and within-person level; with eigenvalue ratios between the first and subsequent factors >10 in all instances (Table 2). A one-factor (between and within) solution showed appropriate model fit with CFI and TLI values approaching 1.0, and RMSEA and SRMRs < 0.05. Thus, the data suggested a single factor on the between- and within-person levels for each domain, supporting dimensional invariance in relation to the 7-day recall PROMIS measures.

DIF across recall periods and measurement levels

Table 3 shows the estimated IRT parameters for the daily items with corresponding national PROMIS item parameters. Also shown are the Wald χ^2 test results for DIF (uniform and non-uniform) across recall periods and for cross-level DIF.

DIF across recall periods—None of the items showed significant non-uniform DIF across 1-day and 7-day recall periods. One pain interference item (“How much did pain interfere with your enjoyment of recreational activities”) showed significant ($p < .05$) uniform DIF across recall periods; the daily format had higher difficulty parameters (thresholds) compared with the PROMIS 7-day recall format. Two fatigue items (“How often did you feel tired”, “How often did you run out of energy”) evidenced highly significant (*p*s < .001) uniform DIF, with both items showing higher difficulty parameters on the daily format than the PROMIS 7-day recall format. In addition, three depression items yielded significant (*p*s < .05) uniform DIF, with two items (“I felt sad”, “I felt like a failure”) showing higher difficulty parameters, and one item (“I felt hopeless”) showing lower difficulty parameters on the daily format than the PROMIS 7-day recall format.

DIF across measurement levels—No significant DIF across between- and within-person measurement levels was evident for pain interference and fatigue items. However, three depression items evidenced cross-level DIF (“I felt sad”, “I felt depressed”, “I felt unhappy”): in each case, the factor loadings were significantly ($p < .001$) higher on the within-person level than on the between-person level, indicating that these items discriminated day-to-day changes in depression *within* given people significantly more strongly than they discriminated differences in depression *between* people.

Impact of DIF

Figure 1 shows the test characteristic curves (TCCs) for between- and within-person levels of analysis. The X-axes indicate the latent PRO scores (thetas) in standard units, centered around the PROMIS general population mean. The Y-axes indicate the expected summed score at each level of theta. The magnitude of DIF at the scale level is reflected in the discrepancy of TCCs from models ignoring DIF versus accounting for DIF. On the between-person level, accounting for DIF shifts the TCCs slightly toward higher theta levels for each PRO domain. However, the divergence in TCCs is small in magnitude, with maximal expected summed score differences of 1.8 points (pain interference; on a 25-point scale), 0.6 points (Fatigue, 29-point scale), and 1.1 points (Depression, 33-point scale). On the within-person level, accounting for DIF results in a notably steeper TCC for depression (maximal difference of 4.9 points), which mirrors the statistically significant evidence for "cross-level" DIF for depression.

Table 4 shows the between- and within-person correlations *among* the three domains. The correlations were moderate to high on the between-person level (*r*s of 0.66–0.88), and moderate (*r*s of 0.36–0.43) on the within-person level when ignoring DIF. DIF-calibration had very little impact on these correlations: the biggest difference was found for the within-person correlation of pain interference and depression, where *r* changed from 0.36 to 0.30 (Cohen's *q* = 0.07 for the difference in correlations, where 0.10 indicates a small effect size).

Table 5 shows the means and variance components of the latent daily PROs estimated from models ignoring DIF and accounting for DIF. The means and between-person variances did not significantly differ between the two models for any domain (*p*s > .05). For depression, the within-person variance component was significantly (*p* < .05) reduced after DIF-calibration (n.s. for pain interference and fatigue). For each domain, the within-person variance accounted for approximately 20–30 % of the total variance; for depression, this percentage was reduced from 24 to 16 % after DIF-calibration.

Finally, it is of interest to compare the means and between-person SDs in Table 5 against the national norms for the PROMIS 7-day recall measures (cf. mean = 0 and SD = 1.0). The between-person variances of the daily measures are close to 1.0 in magnitude, consistent with the 7-day recall norms. Notably, the population means estimated from the daily diaries are significantly (*p*s < .05) less than zero in all instances, indicating that the daily versions for all 3 domains yielded latent PRO scores that were lower on average than the PROMIS 7-day recall norms. In models ignoring DIF, the means for daily measures are 0.32 (pain interference), 0.82 (fatigue), 0.51 (depression) *SD*s lower than the PROMIS norms (*p*s < .01); accounting for DIF reduces the differences in means to 0.28 (pain interference), 0.66 (fatigue), and 0.45 (depression) SDs, but the effect remains significant for all domains (*p*s < .05).

Discussion

The purpose of this pilot-study was to examine whether some of the existing PROMIS measures could be modified for use in daily diary studies without affecting the psychometric properties of the instruments. The immediate concern was that altering the recall-period from 7 to 1 day could affect the measurement models underlying the latent PROs. We found that the daily versions formed a single factor for each PRO domain, consistent with the PROMIS 7-day recall measures. There was no evidence for significant non-uniform DIF across recall periods, but 6 out of 18 tested daily items were flagged with uniform DIF. With one exception, the daily versions had higher difficulty parameters, suggesting that, at a given level of the latent PRO, people had a tendency to report *less* problems in daily diaries compared to 7-day recall. However, the impact of DIF at the overall scale level was minor: DIF had little effect on the between-person test-characteristic curves and correlations among the PRO domains. These results provide preliminary evidence that 1-day and 7-day recall versions of PROMIS measures capture comparable constructs with compatible metrics, such that direct comparisons of PRO scores derived from different recall versions may be largely unbiased.

This does not mean, however, that the different recall versions could be used interchangeably. Notably, the population averages for the daily PROs were considerably (between 0.30 and 0.80 SDs) lower than the PROMIS 7-day recall norms for each of the three domains. An important implication is that norm-based interpretations are no longer valid when items are modified from the current 7-day recall format into a daily reporting format, despite comparable measurement models. Prior research has similarly documented that shorter recall-periods yield lower mean symptom ratings than longer recall periods, suggesting that people may evaluate the intensity and frequency of symptoms differently, or that they may be less influenced by peak symptoms in daily reporting relative to 7-day recall [9, 11, 32, 33]. However, these findings were limited to individual items and therefore could not rule out that the discrepancies were just an artifact of DIF. In the present study, even though DIF accounted for a small portion of the effect, the daily means remained noticeably lower than the 7-day recall norms after DIF-calibration. Thus, the effect may be less a function of measurement properties, but can be attributed to real differences in the way people perceive or report PROs depending upon the length of recall. To level these differences and to map daily scores onto the existing PROMIS 7-day recall norms, a linear transformation may be sufficient. However, the exact nature of this transformation for each domain should be determined with larger samples, using linking strategies that have been proposed to create “cross-walks” between PROMIS norms and alternative measures [34–36].

A second goal of this study was to examine whether the PROMIS IRT-parameters for between-person (cross-sectional) differences are preserved on the level of within-person (day-to-day) change. Using a multilevel extension of the GRM, we found no significant “cross-level DIF” for pain interference and fatigue. This suggests that daily changes in scale scores have the same interpretation and measurement precision as between-person differences for these PROs [17, 25]. For depression, however, 43 % (3 out of 7) of the items were flagged with significant cross-level DIF. Specifically, items that asked about feeling “sad”, “depressed”, and “unhappy” had higher loadings on the within- than the between-person level, indicating that they are more sensitive (i.e., informative) indicators of day-to-day changes than of between-subject differences in depression. Conceptually, it is interesting that these three items tap into depressive mood, whereas the remaining items target cognitive aspects of depression [20]. One common interpretation of DIF is the presence of a dormant secondary dimension [37]. Thus, even though the results suggested a single within-person factor for depression, day-to-day changes in the three DIF items may not exclusively be influenced by a person’s depression levels, but also by a secondary factor reflecting transient changes in mood. Further research is warranted to replicate this finding and to examine its psychological mechanisms and implications for clinical research, for example, through cognitive interviewing techniques.

This study has several limitations. The modest sample size may have limited the statistical power to detect small but systematic deviations from measurement invariance, although this was substantially offset by the increased power provided by 28 days of repeated measurements [38]. Second, we cannot be certain whether the current study sample was randomly equivalent to the PROMIS norming samples (even though the sample was selected to be representative of the general population) and the analyses treated the PROMIS parameters as known without error. This may have increased the likelihood of false positives in DIF testing. Third, it is unclear if the results generalize to other PROMIS domains, across people with various acute or chronic medical conditions, and across demographic subgroups (e.g., gender, age). Furthermore, data were collected from online survey respondents and the results may not generalize to other diary administration methods (e.g., ambulatory devices and interactive voice response systems). Fourth, we evaluated only a subset of items from the PROMIS banks, administered as static short forms. Examining a daily format for larger item pools would be desirable, especially given that calibrated item banks are the basis for CAT administration of PROMIS measures [1, 22].

Daily diaries have great potential for PRO research and clinical practice. They permit examination of changes in PROs in people's everyday context and are less susceptible to bias from retrospection [10]. Diary measures are increasingly used to capture PROs, but their psychometric properties are often not evaluated [39]. The results of this study provide initial evidence that it may be viable to adapt PROMIS measures for daily diary research without significantly affecting the underlying measurement models. Moreover, our findings suggesting that between-person differences and within-person changes in PROs can for the most part be interpreted on the same metric are encouraging. However, scores from daily versions cannot be directly interpreted on PROMIS 7-day recall norms. Further research using larger samples is warranted to document the generalizability of these findings to other settings and PROs and to allow direct mapping of daily PRO scores onto the corresponding PROMIS norms.

Acknowledgments

This research was supported by a grant from the National Institutes of Health (1 U01AR057948-01). We would like to thank Gim Yen Toh and Laura Wolff for their assistance with data collection.

References

1. 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. Journal of Clinical Epidemiology. 2010; 63(11): 1179–1194. [PubMed: 20685078]
2. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates—The PROMIS qualitative item review. Medical Care. 2007; 45(5):S12–S21. [PubMed: 17443114]
3. 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. Journal of Clinical Epidemiology. 2010; 63(11):1169–1178. [PubMed: 20688473]
4. Institute of Medicine. Leading health indicators for healthy people 2020: Letter Report. Washington, DC: National Academies Press; 2011.
5. Leidy NK, Wilcox TK, Jones PW, Murray L, Winnette R, Howard K, et al. Development of the EXAcerbations of Chronic Obstructive Pulmonary Disease Tool (EXACT): A Patient-Reported Outcome (PRO) Measure. Value in Health. 2010; 13(8):965–975. [PubMed: 20659270]
6. Jim HS, Small B, Faul LA, Franzen J, Apte S, Jacobsen PB. Fatigue, depression, sleep, and activity during chemotherapy: daily and intraday variation and relationships among symptom changes. Annals of Behavioral Medicine. 2011; 42(3):321–333. [PubMed: 21785899]
7. Chapman CR, Donaldson GW, Davis JJ, Bradshaw DH. Improving individual measurement of postoperative pain: The Pain Trajectory. Journal of Pain. 2011; 12(2):257–262. [PubMed: 21237721]
8. Begg A, Drummond G, Tiplady B. Assessment of postsurgical recovery after discharge using a pen computer diary. Anaesthesia. 2003; 58(11):1101–1105. [PubMed: 14616597]
9. Broderick JE, Schneider S, Schwartz JE, Stone AA. Interference with activities due to pain and fatigue: Accuracy of ratings across different reporting periods. Quality of Life Research. 2010; 19(8):1163–1170. [PubMed: 20535565]
10. 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(3):274–281. [PubMed: 19070550]
11. 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(1):146–157. [PubMed: 18455312]
12. Schwarz, N. Retrospective and concurrent self-reports: The rationale for real-time data capture. In: Stone, AA.; Shiffman, SS.; Atienza, A.; Nebeling, L., editors. The science of real-time data capture: Self-reports in health research. New York: Oxford University Press; 2007. p. 11-26.

13. Winkielman P, Knauper B, Schwarz N. Looking back at anger: Reference periods change the interpretation of emotion frequency questions. Journal of Personality and Social Psychology. 1998; 75(3):719–728. [PubMed: 9781408]

14. Roesch SC, Aldridge AA, Stocking SN, Villodas F, Leung Q, Bartley CE, et al. Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. Multivariate Behav Res. 2010; 45(5):767–789. [PubMed: 21399732]

15. Mehta PD, Neale MC. People are variables too: Multilevel structural equations modeling. Psychological Methods. 2005; 10(3):259–284. [PubMed: 16221028]

16. Zyphur MJ, Kaplan SA, Christian MS. Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. Group Dynamics-Theory Research and Practice. 2008; 12(2):127–140.

17. Skrondal, A.; Rabe Hesketh, S. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Boca Raton, FL: Chapman & Hall; 2004.

18. 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(1):173–182. [PubMed: 20554116]

19. 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. Archives of Physical Medicine and Rehabilitation. 2011; 92(10 Suppl):S20–S27. [PubMed: 21958919]

20. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, et al. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS (R)): Depression, Anxiety, and Anger. Assessment. 2011; 18(3):263–283. [PubMed: 21697139]

21. Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometrika. 1969; 34:100–114.

22. 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). Medical Care. 2007; 45(5):S22–S31. [PubMed: 17443115]

23. Muthén, LK.; Muthén, BO. Mplus user's guide. 6th ed. Los Angeles, CA: Muthén & Muthén; 1998–2010.

24. McDonald RP. Linear Versus Non-Linear Models in Item Response Theory. Applied Psychological Measurement. 1982; 6(4):379–396.

25. Grilli L, Rampichini C. Multilevel factor models for ordinal variables. Structural Equation Modeling-a Multidisciplinary Journal. 2007; 14(1):1–25.

26. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Medical Care. 2006; 44(11):S78–S94. [PubMed: 17060839]

27. Mellenberg GJ. Contingency table models for assessing item bias. Journal of Educational Statistics. 1982; 7:105–108.

28. Muthén, BO.; Asparouhov, T. Beyond multilevel regression modeling: multilevel analysis in a general latent variable framework. In: Hox, J.; Roberts, JK., editors. The Handbook of Advanced Multilevel Analysis. New York: Taylor and Francis; 2009. p. 15-40.

29. Woods CM. Empirical Selection of Anchors for Tests of Differential Item Functioning. Applied Psychological Measurement. 2009; 33(1):42–57.

30. Stark S, Chernyshenko ES, Drasgow F. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. Journal of Applied Psychology. 2006; 91(6):1292–1306. [PubMed: 17100485]

31. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological. 1995; 57(1):289–300.

32. Stone AA, Schwartz JE, Broderick JE, Shiffman SS. 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(10):1340–1346. [PubMed: 16143666]

33. Schneider S, Stone AA, Schwartz JE, Broderick JE. Peak and End Effects in Patients' Daily Recall of Pain and Fatigue: A Within-Subjects Analysis. Journal of Pain. 2011; 12(2):228–235. [PubMed: 20817615]

34. Noonan VK, Cook KF, Bamer AM, Choi SW, Kim J, Amtmann D. Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form. Quality of Life Research. 2011

35. Thissen D, Varni JW, Stucky BD, Liu Y, Irwin DE, Dewalt DA. Using the PedsQL 3.0 asthma module to obtain scores comparable with those of the PROMIS pediatric asthma impact scale (PAIS). Quality of Life Research. 2011; 20(9):1497–1505. [PubMed: 21384264]

36. Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, et al. Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. Quality of Life Research. 2011; 20(9):1349–1357. [PubMed: 21409516]

37. Ackerman TA. A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective. Journal of Educational Measurement. 1992; 29(1):67–91.

38. Muthén BO, Curran PJ. General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. Psychological Methods. 1997; 2(4):371–402.

39. Cranford JA, Shrout PE, Iida M, Rafaeli E, Yip T, Bolger N. A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? Personality and Social Psychology Bulletin. 2006; 32(7):917–929. [PubMed: 16738025]

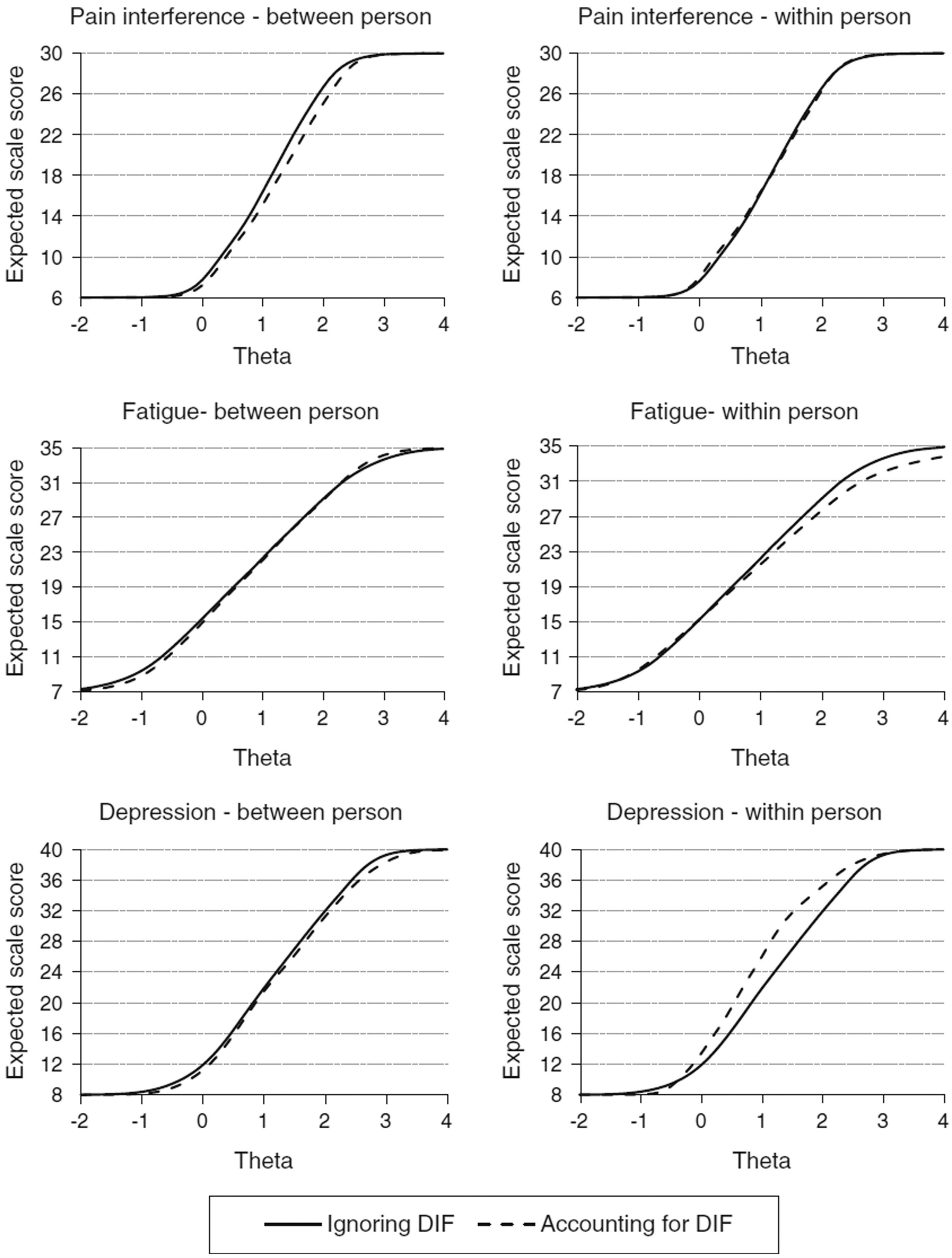

Figure 1.

Test characteristic curves based on models ignoring differential item functioning (*solid lines*) and accounting for differential item functioning (*dashed lines*), for between-person and within-person levels of analysis.

Table 1

Demographic characteristics of study participants (*N* = 100)

	Percent
Age (mean = 43.6, SD = 14.8)	
21–24	10
25–34	24
35–44	17
45–54	25
55–64	15
65+	9
Gender	
Female	52
Education	
Less than high school	2
High school graduate	16
Some college	45
College graduate	27
Masters/doctoral	10
Race	
White	71
African American	15
Native American	2
Asian	6
Other/multiple	6
Ethnicity	
Hispanic	14
Marital status	
Never married/living together	36
Married	45
Separated/widowed	4
Divorced	15
Family income	
$0–19,999	6
$20,000–49,999	51
$50,000–74,999	22
$75,000 and higher	21

Table 2

Multilevel exploratory factor analysis model results

	Pain interference (6 items)	Fatigue (7 items)	Depression (8 items)
Eigenvalues			
Between person level			
1st/2nd/3rd factor	5.81/0.11/0.05	6.45/0.27/0.11	7.30/0.39/0.13
Within person level			
1st/2nd/3rd factor	4.95/0.28/0.23	5.30/0.38/0.31	5.76/0.49/0.38
Fit indices for 1-factor solution on both levels			
χ^2 (*df*)	30.70 (18)	32.49 (28)	69.13 (40)
CFI	0.999	1.000	0.997
TLI	0.998	0.999	0.995
RMSEA (90% CL)	0.016 (0.005/0.026)	0.008 (0.000/0.018)	0.017 (0.010/0.023)
SRMR for between	0.006	0.024	0.039
SRMR for within	0.010	0.014	0.027

CFI Comparative Fit Index, *TLI* Tucker-Lewis Index, *RMSEA* root mean square error of approximation, *SRMR* standardized root mean residual

Table 3

Item parameters and tests for differential item functioning (DIF) of daily diary form

		Item parameters (standard errors)						Chi-square for DIF		
		α_{within}	$\alpha_{between}$	β1	β2	β3	β4	Cross-level[a]	Non-uniform (recall)[b]	Uniform (recall)[c]
Pain interference										
How much did pain interfere with your enjoyment of life?	Daily	6.42 (0.76)	6.28 (0.69)	0.14 (0.05)	0.91 (0.08)	1.51 (0.13)	2.12 (0.19)	0.03	3.53	3.31
	PROMIS		**4.98**	**0.13**	**0.88**	**1.38**	**1.91**			
How much did pain interfere with your ability to concentrate?	Daily/PROMIS	3.75 (0.00)	3.75 (0.00)	0.40 (0.00)	1.11 (0.00)	1.69 (0.00)	2.34 (0.00)	Reference item		
How much did pain interfere with your daily activities?	Daily	7.31 (0.65)	6.86 (0.68)	0.12 (0.05)	0.86 (0.05)	1.54 (0.11)	2.17 (0.17)	0.36	0.23	7.62
	PROMIS		**6.53**	**0.16**	**0.90**	**1.44**	**2.01**			
How much did pain interfere with your enjoyment of recreational activities?	Daily	5.86 (0.75)	6.15 (0.66)	0.23 (0.04)	0.91 (0.09)	1.43 (0.14)	1.92 (0.26)	0.09	2.29	16.75*
	PROMIS		**5.15**	**0.13**	**0.79**	**1.26**	**1.85**			
How much did pain interfere with doing your tasks away from home (e.g., getting groceries, running errands)?	Daily	5.95 (0.54)	5.64 (0.64)	0.30 (0.05)	1.00 (0.08)	1.58 (0.12)	2.18 (0.17)	0.20	1.60	13.01
	PROMIS		**4.83**	**0.43**	**1.00**	**1.46**	**2.04**			
How much did pain feel like a burden to you?	Daily	4.64 (0.49)	5.46 (0.63)	0.08 (0.05)	0.77 (0.07)	1.27 (0.10)	1.79 (0.18)	1.31	3.53	2.91
	PROMIS		**4.28**	**0.11**	**0.76**	**1.19**	**1.72**			
Fatigue										
How often did you feel tired?	Daily	4.01 (0.32)	3.45 (0.28)	−1.34 (0.07)	−0.27 (0.06)	1.03 (0.08)	2.27(0.17)	5.90	0.52	44.45***
	PROMIS		**3.25**	**−1.62**	**−0.48**	**0.73**	**1.79**			
How often did you experience extreme exhaustion?	Daily	2.19 (0.58)	3.02 (0.55)	−0.10 (0.08)	0.65 (0.09)	1.51 (0.16)	2.54 (0.26)	3.91	0.42	12.42
	PROMIS		**2.66**	**−0.11**	**0.84**	**1.72**	**2.83**			
How often did you run out of energy?	Daily	3.92 (0.43)	4.18 (0.35)	−0.72 (0.06)	0.19 (0.06)	1.20 (0.09)	2.21 (0.17)	0.64	5.18	28.04***
	PROMIS		**3.38**	**−1.01**	**0.04**	**1.08**	**2.16**			
How often did your fatigue limit you at work (include work at home)?	Daily	2.97 (0.30)	3.55 (0.41)	−0.44 (0.06)	0.42 (0.07)	1.35 (0.14)	2.31 (0.24)	3.98	1.21	6.64
	PROMIS		**3.09**	**−0.54**	**0.33**	**1.35**	**2.32**			
How often were you too tired to think clearly?	Daily	1.73 (0.41)	2.78 (0.43)	−0.17 (0.08)	0.64 (0.10)	1.66 (0.18)	2.59 (0.29)	6.42	0.19	12.27
	PROMIS		**2.97**	**−0.11**	**0.82**	**1.82**	**3.05**			
How often did you feel tired even when you hadn’t done anything?	Daily/PROMIS	2.84 (0.00)	2.84 (0.00)	−0.66 (0.00)	0.24 (0.00)	1.30 (0.00)	2.38 (0.00)	Reference item		

		Item parameters (standard errors)						Chi-square for DIF		
		α_{within}	$\alpha_{between}$	β1	β2	β3	β4	Cross-level[a]	Non-uniform (recall)[b]	Uniform (recall)[c]
How often did you have to push yourself to get things done because of your fatigue?	Daily	3.84 (0.30)	4.21 (0.32)	−0.58 (0.05)	0.23 (0.05)	1.16 (0.08)	2.04 (0.17)	2.57	3.04	5.08
	PROMIS		**4.77**	**−0.63**	**0.13**	**1.04**	**1.94**			
Depression										
I felt worthless	Daily	4.17 (0.50)	5.15 (0.47)	0.46 (0.05)	0.97 (0.05)	1.66 (0.06)	2.33 (0.12)	5.09	3.60	3.94
	PROMIS		**4.26**	**0.40**	**0.98**	**1.70**	**2.44**			
I felt that I had nothing to look forward to	Daily/PROMIS	3.93 (0.00)	3.93 (0.00)	0.30 (0.00)	0.91 (0.00)	1.59 (0.00)	2.41 (0.00)	Reference item		
I felt helpless	Daily	3.93 (0.50)	3.62 (0.38)	0.27 (0.06)	0.83 (0.06)	1.60 (0.09)	2.39 (0.16)	0.51	1.91	2.66
	PROMIS		**4.14**	**0.35**	**0.92**	**1.68**	**2.47**			
I felt sad.	Daily	8.99 (0.85)	4.05 (0.50)	−0.28 (0.09)	0.62 (0.09)	1.81 (0.20)	3.08 (0.32)	47.35 ***	2.43	16.96 *
	PROMIS		**3.27**	**−0.50**	**0.41**	**1.41**	**2.38**			
I felt like a failure	Daily	4.62 (0.56)	4.59 (0.43)	0.37 (0.05)	0.90 (0.06)	1.60 (0.08)	2.33 (0.12)	0.004	2.12	23.51 **
	PROMIS		**3.97**	**0.20**	**0.80**	**1.65**	**2.30**			
I felt depressed	Daily	7.65 (0.74)	3.92 (0.51)	0.01 (0.05)	0.61 (0.07)	1.43 (0.14)	2.16 (0.22)	36.13 ***	0.71	11.85
	PROMIS		**4.34**	**−0.12**	**0.60**	**1.43**	**2.27**			
I felt unhappy	Daily	8.65 (0.87)	4.13 (0.50)	−0.33 (0.09)	0.59 (0.08)	1.71 (0.17)	2.87 (0.27)	47.88 ***	1.69	15.40
	PROMIS		**3.48**	**−0.54**	**0.35**	**1.35**	**2.35**			
I felt hopeless	Daily	5.47 (0.57)	5.27 (0.49)	0.34 (0.04)	0.88 (0.04)	1.60 (0.07)	2.23 (0.10)	0.26	2.83	56.48 ***
	PROMIS		**4.45**	**0.56**	**1.07**	**1.78**	**2.53**			

Population parameters for 7-day recall PROMIS items are presented in bold; *DIF* differential item functioning, α_{within} = within-person discrimination parameter; $\alpha_{between}$ = between-person discrimination parameter; β1 to β4 = difficulty parameters;

[a] Compares α_{within} with $\alpha_{between}$ parameter (*df* = 1);

[b] compares $\alpha_{between}$ parameters across daily and PROMIS forms (*df* = 1);

[c] compares β parameters across daily and PROMIS forms (*df* = 4);

* $p < .05$;

** $p < .01$;

[***] $p < .001$.

Table 4

Between- and within-person correlations among latent daily diary PROs based on models ignoring DIF (above diagonal) and models accounting for DIF (below diagonal)

	Between-person correlations			**Within-person correlations**		
	Pain interference	**Fatigue**	**Depression**	**Pain interference**	**Fatigue**	**Depression**
Pain interference	-	0.88	0.66	-	0.43	0.36
Fatigue	*0.87*	-	0.71	*0.40*	-	0.38
Depression	*0.66*	*0.68*	-	*0.30*	*0.37*	-

Table 5

Means and variance components (standard errors in parentheses) of latent daily diary PROs based on models ignoring DIF and models accounting for DIF

		Mean	Between-person SD	Within-person SD
Pain interference	Ignoring DIF	−0.318 (0.11)	1.066 (0.08)	0.553 (0.04)
	Accounting for DIF	−0.275 (0.12)	1.047 (0.10)	0.575 (0.05)
Fatigue	Ignoring DIF	−0.823 (0.13)	1.276 (0.07)	0.770 (0.04)
	Accounting for DIF	−0.663 (0.13)	1.193 (0.10)	0.832 (0.08)
Depression	Ignoring DIF	−0.513 (0.11)	1.048 (0.07)	0.585 (0.03)
	Accounting for DIF	−0.450 (0.14)	1.068 (0.11)	0.459 (0.05)