Quality of Life Research (2020) 29:745–753
https://doi.org/10.1007/s11136-019-02344-0

Does recall period matter? Comparing PROMIS® physical function with no recall, 24-hr recall, and 7-day recall

David M. Condon[1] · Robert Chapman[1] · Sara Shaunfield[1] · Michael A. Kallen[1] · Jennifer L. Beaumont[1,2] · Daniel Eek[3] · Debanjali Mitra[4] · Katy L. Benjamin[5] · Kelly McQuarrie[6] · Jamae Liu[7] · James W. Shaw[8] · Allison Martin Nguyen[9] · Karen Keating[10] · David Cella[1]

Accepted: 21 October 2019 / Published online: 7 November 2019

Abstract

Purpose To evaluate the influence of recall periods on the assessment of physical function, we compared, in cancer and general population samples, the standard administration of PROMIS Physical Function items without a recall period to administrations with 24-hour and 7-day recall periods.

Methods We administered 31 items from the PROMIS Physical Function v2.0 item bank to 2400 respondents ($n=1001$ with cancer; $n=1399$ from the general population). Respondents were randomly assigned to one of three recall conditions (no recall, 24-hours, or 7-days) and one of two "reminder" conditions (with recall periods presented only at the start of the survey or with every item). We assessed items for potential differential item functioning (DIF) by recall time period. We then tested recall and reminder effects with analysis of variance controlling for demographics, English fluency, and co-morbidities.

Results Based on conservative pre-set criteria, no items were flagged for recall time period-related DIF. Using analysis of variance, each condition was compared to the standard PROMIS administration for Physical Function (no recall period). There was no evidence of significant differences among groups in the cancer sample. In the general population sample, only the 24-hour recall condition with reminders was significantly different from the "no recall" PROMIS standard. At the item level, for both samples, the number of items with non-trivial effect size differences across conditions was minimal.

Conclusions Compared to no recall, the use of a recall period has little to no effect upon PROMIS physical function responses or scores. We recommend that PROMIS Physical Function be administered with the standard PROMIS "no recall" period.

Keywords Patient reported outcomes · PROMIS · Recall period · Physical function

Over the last decade, considerable progress has been made towards the standardization of methods for assessing patient reported outcomes (PROs) [1–4]. The Patient Reported Outcome Measurement Information System® (PROMIS®) has contributed to this progress through consistent implementation of the PROMIS methodology [5–7] for the development of item banks and short forms in more than 100 domains of physical, mental, and social health (www.HealthMeasures.net). Many recent studies have demonstrated

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-019-02344-0) contains supplementary material, which is available to authorized users.

✉ David Cella
d-cella@northwestern.edu

1 Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

2 Terasaki Research Institute, Los Angeles, CA, USA

3 AstraZeneca, Gothenburg, Sweden

4 Pfizer, Inc., New York, NY, USA

5 Health Economics and Outcomes Research, AbbvVie Inc., North Chicago, IL, USA

6 Janssen Global Services, Malvern, PA, USA

7 Health Economics and Outcomes Research, Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA

8 Worldwide Health Economics and Outcomes Research, Bristol-Myers Squibb, Lawrenceville, NJ, USA

9 Merck & Co., Inc., Kenilworth, NJ, USA

10 Bayer HealthCare Pharmaceuticals, Inc., West Haven, CT, USA

the validity of these measures for use in a diverse range of contexts and disease populations [8–11]. The PROMIS framework is widely cited and used when assessing common symptoms and functional domains of health-related quality of life [12–19].

Qualitative item review is particularly critical to ensure that the items capture relevant patient concerns in each domain, and that they are unambiguous and intelligible to people with a range of literacy [20, 21]. Qualitative item review also helps to ensure consistency of style, response options, and recall periods. For recall periods specifically, PROMIS investigators sought to identify the option(s) that would reduce the potential for bias in responding by drawing upon research from several disciplines, including memory encoding and recall [22–25], and judgment and decision-making [26–28]. For most PROMIS domains, the qualitative item review process led to the selection of a 7-day recall period as a general convention. The physical function domain is one of a few exceptions [29, 30]. The PROMIS Physical Function domain does not specify a recall period because of a prevailing preference in this particular domain to focus on self-evaluations of current capability rather than specific recollections over a defined time period (e.g., over the last 24 hours or 7 days). The latter approach—asking about functioning over a specified time period rather than perception of current capability—introduces uncertainty about the extent of functioning when respondents have not engaged in the activities described within the specified time period (e.g., getting in and out of a car, climbing stairs, exercising) [6, 20]. Thus, the PROMIS Physical Function items are phrased in the present tense to assess patients' *assessment of their current capability* to carry out various physical activities. It is expected that those items that reflect an activity that patients have recently performed would naturally be responded to based upon that recent experience. In cases where a physical function item's exemplar activity has not been performed recently, patients estimate their capability based on recent experience with similar tasks and/or reasoned estimation of their current physical capability [6].

Empirical evidence regarding the use of various recall periods has been mixed. Generally, the length of recall period is inversely related to the accuracy of recall [31–34]. Shorter recall periods can lead to the under-reporting of symptoms in some conditions, while longer recall periods can lead to over-reporting [35]. Still others have found no significant effects based on the length of recall and have recommended that recall periods be selected as needed to meet the needs of the administering clinicians/researchers [36, 37], including the possibility of using multiple "ecological momentary assessments" to acquire more robust data regarding respondents' experiences over time [38].

In the current study, the influence of recall period on self-report was evaluated by administering 31 PROMIS Physical Function items to a large online sample. Specifically, these items were administered to samples from two distinct populations using three different recall conditions and two different administration conditions. The primary aim was to evaluate whether—and to what extent—the use of different recall periods and reminder options might lead to significantly different means among the items as a set and individually. The three recall options were: (1) no recall period (i.e., the current PROMIS approach); (2) 24-hour recall; and (3) 7-day recall. A second independent variable in this experiment was the use of reminders regarding the recall period: One mention of time frame at the beginning of the assessment (i.e., no reminders) versus a reminder with every item. The impetus and design of this study is a consequence of guidance and feedback received from the Food and Drug Administration (FDA) Center of Drug Evaluation and Research (CDER) Qualification Review Team (QRT), during the development of a Drug Development Tool (DDT) Clinical Outcome Assessment (COA) of PROMIS Physical Function in oncology.

Methods

The study protocol was reviewed by the Institutional Review Board Office of Northwestern University (IRB ID STU00205190) and exempted from full review.

Participants

Participants included 2400 English-speaking individuals who were recruited online between May 16 and May 25, 2017 by *Opinions For Good* (Op4G), a market research firm that maintains relationships with a large panel of survey respondents. Of these, 1001 respondents (40% female) were invited to participate because they were currently undergoing treatment for a cancer diagnosis (the "Cancer" sample). The remaining 1399 respondents (50.5% female) were recruited as part of a representative sample of the U.S. population with respect to age, gender, race/ethnicity, and education (the "General Population" sample). All participants gave their consent to participate by clicking "I agree" on a customized informed consent page and were required to actively agree to participate by opting in. For the largest sub-group (no recall, no reminder, general population; $N = 598$), representative proportions were achieved for gender, age, and education, though the joint representativeness for these demographic characteristics was not fully achieved in all race/ethnicity groups. Demographic characteristics of the cancer and general population samples are in Table 1.

Table 1 Descriptive statistics for the general population sample ($n = 1399$) and the cancer sample ($n = 1001$)

Recall condition	No recall		24-h recall				7-day recall			
Reminder condition	No reminder		Reminder		No reminder		Reminder		No reminder	
Sample	GP	Ca	GP	Ca	GP	Ca	GP	Ca	GP	Ca
Sample size	598	201	201	200	200	200	200	200	200	200
Gender (% female)	24%	59%	66%	63%	70%	61%	75%	57%	71%	60%
Age (years)										
Mean	43.0	40.0	46.9	38.8	50.1	38.6	48.7	39.5	46.5	38.7
Median	42	37	46	36	51	37	51	39	46	36
SD	13.5	14.2	14.0	13.0	14.3	12.84	13.7	12.7	13.6	13.1
Race										
White/Caucasian	56%	74%	80%	80%	79%	79%	82%	78%	83%	80%
Black/African-American	19%	10%	10%	6%	9%	10%	8%	10%	8%	6%
Asian/Pacific Islander	7%	8%	4%	6%	5%	6%	4%	5%	2%	5%
Hispanic/Latino	18%	7%	5%	8%	5%	5%	6%	7%	4%	8%
Other	1%	0%	0%	0%	2%	0%	0%	1%	2%	1%
Education										
8th grade or less	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9th to 11th grade	2%	1%	2%	0%	3%	1%	2%	0%	2%	1%
GED/HS grade	22%	17%	28%	15%	24%	16%	24%	16%	23%	17%
1–3 years of college	34%	26%	32%	28%	37%	32%	36%	30%	33%	30%
4 year college graduate	28%	36%	26%	32%	26%	34%	23%	32%	28%	36%
Graduate degree	15%	20%	13%	24%	9%	17%	15%	21%	14%	15%

GP general population sample, *Ca* cancer sample

Materials

31 items from the PROMIS Physical Function item bank were administered online, either in the standard PROMIS format without a specified recall period or with a recall period of either 7 days or 24 hours. These 31 items included the 10 items in the PROMIS Short Form v2.0—Physical Function 10b [39, 40], the 16 items previously validated for use with a diverse U.S. population-based cohort of cancer patients [9], and 10 additional items that were suggested for inclusion by the U.S. Food and Drug Administration in a collaborative project. Note that there are five overlapping items in the 10b- and 16-item short forms, bringing the total PF item count to 31. In addition to the Physical Function items, participants were asked to complete the 10-item PROMIS Scale v1.2—Global Health [41], the Functional Assessment of Cancer Therapy—General [42], and to provide information regarding demographics, prior diagnoses, and co-morbidities.

Procedure

In order to test the effect of recall periods, participants were randomly assigned to one of three groups: 799 participants (201 and 598 from the cancer and general population samples, respectively) were administered the items without any reference to a recall period; 801 participants (400 and 401 from the cancer and general population samples, respectively) were administered the items with a 24-hour recall period (i.e., "Over the last 24 hours, …"); and 800 participants (400 and 400 from the cancer and general population samples, respectively) were administered the items with a 7-day recall period (i.e., "Over the last 7 days, …"). To test the effect of reminders of the recall period, participants in the 24-hour and 7-day recall groups were assigned, via a second randomization, into one of two reminder conditions: recall period presented only once at the beginning of the survey; or recall period presented with each of the 31 items. This recruitment and randomization scheme ensured enrollment of at least 200 people in each group.

Analyses

For our unidimensionality analyses, we sought to confirm that a single-factor structure underlies the PROMIS Physical Function-Oncology 31-item set. Using our full combined sample ($N = 2400$), we first examined item-total score correlations (i.e., corrected for item overlap) to identify any correlations < 0.40 as indicating possible low item-construct association. We reviewed all inter-item correlations for negligible correlations (e.g., < 0.10), suggesting unrelated item pairs, and extremely high correlations (e.g., > 0.90), suggesting

potential item content redundancy. We also obtained an estimate of internal consistency reliability (Cronbach's alpha). Next, we conducted a single-factor confirmatory factor analysis (CFA) to determine if all factor loadings were ≥ 0.50, all residual correlations were ≤ 0.20, and the overall model's fit statistics indicated good fit (e.g., root mean square error of approximation (RMSEA) ≤ 0.10, Comparative Fit Index (CFI) ≥ 0.95, Tucker-Lewis Index (TLI) ≥ 0.95, standardized root mean square residual (SRMR) ≤ 0.08), thereby confirming a unidimensional model [43–48]. Then, for bifactor modeling, we began by conducting an exploratory factor analysis (EFA), with plans to extract two to three factors. We obtained percent of variance accounted for by eigenvalues 1–3; we also calculated the eigenvalue 1-to-2 ratio, with values ≥ 4.0 suggestive of a single, dominant first factor. Subsequently, we conducted a confirmatory bifactor analysis (CBFA), using our EFA findings to establish two to three evidence-based specific factors. We calculated omega, McDonald's $omega_{hierarchial}$ (omega-H), and explained common variance (ECV). Omega-H values ≥ 0.70 are considered suggestive of sufficient unidimensionality, while ECV values > 0.50 are interpreted as a majority percentage of "common variance" having been explained by a single, general factor [49–52]. For Stage 1 DIF analyses, we implemented a hybrid logistic ordinal regression (LOR) and IRT approach to DIF detection. This involved the use of an IRT-derived ability (or trait) estimate for LOR modeling rather than a traditionally modeled summed-score ability (trait) term. In this stage we sought to identify items flagged for DIF by our investigated DIF factors: (1) cancer vs. general population (both "no recall" only); (2) recall time period. These analyses used (1) "general population-no recall" as the reference group and "cancer-no recall" as the focal group, and (2) "no recall" as the reference group and both "24-hour recall" and "7-day recall" as focal groups. For the cancer vs. general population (no recall only) DIF factor, tested item content and time frame context were identical across tested groups. This is analogous to gender DIF factor testing, where items are fixed and tested groups vary (female vs. male), with the null hypothesis being items do not perform differently per gender status. For the recall time period DIF factor, tested item content was again identical across tested groups, while time frame context was allowed to vary. This is analogous to language DIF factor testing, where common-content items are presented in distinct languages (e.g., English and Spanish), tested groups have common characteristics except for their language status (English-speaking vs. Spanish-speaking), and the null hypothesis is that items do not perform differently per language-presentation status. We used a McFadden pseudo-R^2 change criterion of ≥ 0.02 to flag items for DIF and utilized the lordif R package, version 0.3–3, for conducting the DIF analyses [53]. Our DIF analyses evaluated uniform DIF (Model 1 vs Model 2), non-uniform DIF (Model 2 vs Model 3), and overall or total DIF (Model 1 vs Model 3). In lordif-based Stage 1 DIF analyses, the initial run employs the full set of tested items as anchors. In subsequent iterations, if item performance differences are found, such "flagged" items are removed as anchors, creating an empirically purified anchor set. Iterations continue until no additional item performance differences are identified and a final DIF-free item anchor set is established.

For Stage 2 DIF score impact studies, we planned to analyze the potential score impact of using a common set of item parameters for scoring vs. group-specific item parameters for any items flagged for DIF. We planned to conduct unadjusted (using a common set of item parameters) vs. DIF-adjusted (using a common subset of item parameters plus group-specific item parameters for flagged items) score difference analyses to obtain the following scoring impact evidence: (a) Pearson correlation (unadjusted vs. adjusted scores); (b) mean difference (unadjusted minus adjusted score); (c) standard deviation (SD) of the score differences; (d) root mean squared difference (RMSD) of the score differences; and (e) percentage of individual case score differences greater than their associated unadjusted score standard error (SE). We prepared to utilize the statistical program IBM SPSS, version 25.0.0.1, for conducting these Stage 2 DIF analyses [54]. All score estimates and score-related statistics, unless specifically noted otherwise, would be reported in or based on the theta metric (mean = 0; SD = 1).

Group differences in raw item-level and IRT-scored scale-level (Theta) scores across the recall periods and the reminder conditions were evaluated using fixed main effects analyses of variance (ANOVA). In follow-up multiple regression analyses, adjusting for age, gender, education, race/ethnicity, English fluency, and co-morbidities in both the general population and cancer samples. In addition, in the cancer sample, we adjusted scores for time since diagnosis, primary cancer site, and type of treatment. Model estimated mean differences were derived for each group; effect size differences were evaluated in T-score units (mean of 50 and standard deviation of 10), consistent with the PROMIS scoring metric [5, 6]. While the DIF analyses investigate the possibility that individual items may perform differently with one recall period versus another, the analysis of group differences in theta/T-scores for the different recall periods shows whether there is any consequence associated with DIF at the scale level.

All PROMIS Physical Function measures are optimally scored using IRT-based EAP (*Expected* A Priori) scoring methods, which rely on item-level calibrations (i.e., discrimination and threshold parameters) to convert a raw sum-score to a weighted distribution-based score centered on the U.S. general population (T-score) [5, 6]. Score conversions were completed using an electronic summed-score-to-IRT-score conversion table. Differences were considered trivial

if below 2 T-score units (i.e., effect size 0.2) for the group differences and less than 0.2 SD (effect size) units for the item-level differences. [55] Higher scores on the PROMIS Physical Function metric indicate better physical functioning. No a priori hypotheses were made regarding the expectation of significant effects in either direction, though we did expect, given the large sample size ($n = 2400$) and multiple comparisons (31 items across 10 possible conditions), that some item-level differences would be observed by chance alone. With 200 patients per group, setting alpha = .05, this study had 85% power to detect a group difference of 0.3 SD units (3 T-score units). We did not adjust alpha for multiple comparisons.

Results

From our unidimensionality analyses, we confirmed a single-factor structure underlies the PROMIS Physical Function-Oncology 31-item set. With our full combined sample ($N = 2400$), we examined item-total score correlations; no correlations were < 0.40, thus, there was no indication of possible low item-construct association. In our review of inter-item correlations we found no negligible correlations (< 0.10) and no extreme high correlations (> 0.90); therefore, all items appeared sufficiently inter-related, and no items appeared redundant (minimum and maximum inter-item correlations were 0.34 and 0.88, respectively). Our estimate of Cronbach's alpha (internal consistency reliability) was 0.98. In our single-factor CFA, all factor loadings were ≥ 0.50, all residual correlations were ≤ 0.20, and the overall model's fit statistics indicated good fit, confirming a unidimensional model (i.e., RMSEA = 0.107, CFI 0.963, TLI = 0.961, and SRMR = 0.056). For our bifactor modeling, we first extracted two factors in our EFA (a potential third factor had no item loadings ≥ 0.30). We obtained the percent of variance accounted for by eigenvalues 1 (76.3%), 2 (5.0%), and 3 (2.2%); we also calculated the eigenvalue 1-to-2 ratio (15.1). EFA findings were all suggestive of an essentially unidimensional factor structure. In our confirmatory bifactor analysis (CBFA), using two EFA evidence-based specific factors, omega was 0.99, omega-H was 0.95, and ECV was 0.96. Our omega-H value was ≥ 0.70, suggestive of sufficient unidimensionality, and our ECV value was > 0.50, representing a majority percentage of "common variance" as explained by a single, general factor. Thus, our CBFA general factor accounted for 95% of PROMIS Physical Function-Oncology total score variance (omega-H). Model fit statistics from the bifactor model indicated good fit (i.e., RMSEA = 0.078, CFI = 0.982, TLI = 0.979, SRMR = 0.025).

Initial lordif analyses used the full 31-item set of tested PROMIS Physical Function-Oncology items as anchors. In all subsequent iterations within each of the lordif analyses conducted, no item performance differences were identified. Thus, in Stage 1 of the planned DIF analyses, no items were flagged for the DIF factor cancer vs. general population (cancer: $n = 1001$; general population: $n = 1399$), and no items were flagged for the DIF factor recall time period (no recall period: $n = 799$; 7-day recall period: $n = 800$; 24-hour recall period: $n = 801$). As a result, for all analyses, the full 31-item set served as the purified or DIF-free anchor set. Therefore, our empirical conclusion is that the use of a no recall vs. 24-hour recall and vs. 7-day recall period did not create item differences either within or across cancer and general population samples big enough to detect via our effect size-based analyses or important enough to impact scores. No Stage 2 DIF score impact studies were required. We therefore used the established (existing) PROMIS PF item parameters for all subsequent PROMIS PF scoring.

While our DIF analyses assessed IRT model differences between recall conditions, observed group differences for the recall period conditions were evaluated using analysis of variance (ANOVA). Unadjusted means and standard deviations for IRT-scored scale-level scores for each condition are reported in Table 2. The mean T-score difference among those in the cancer sample and those in the general population sample ($m = 11.1$ T-score points), reflect a substantial impact of cancer upon physical functioning. We observed the full range of T-scores (11.6–62.8) in both the cancer and general population samples. These minimum and maximum scores represent the floor (1% of general population sample; 1% for cancer sample) and ceiling (26% of general population sample; 3% of cancer sample), on the 31-item PF measure.

Analyses of variance (ANOVA) comparing the main effects of the recall and reminder conditions were conducted separately in the cancer and general population samples. A significant difference among groups was found in the general population sample ($F(4, 1374) = 2.67$; $p = .03$) but not the cancer sample ($F(4, 960) = 2.35$; $p = .052$). In the general population, slightly higher (better) physical function scores

Table 2 ANOVA-based T-score means and standard deviations by sample, recall period, and reminder condition

Recall period and reminder administered	General population sample		Cancer sample	
	Mean	SD	Mean	SD
No recall (PROMIS standard)	49.5	10.6	37.6	7.2
24-h recall with reminders	51.4	10.1	39.9	8.4
24-h recall without reminders	49.7	10.3	39.3	8.0
7-day recall with reminders	50.4	10.3	38.2	7.7
7-day recall without reminders	48.4	11.1	39.1	8.5

were observed when using a recall period. Table 3 shows the estimated mean differences from multiple regression models adjusting for age, gender, education, race/ethnicity, English fluency, presence of co-morbidities, and—for the cancer sample—the time since diagnosis, primary site of cancer diagnosis, and the cancer treatment type. Based on the evidence for statistical significance in the general population ANOVA, pairwise tests of significance were conducted between estimated adjusted means for the PROMIS standard condition (no recall) and each of the other recall and reminder conditions. Only the 24-hour recall condition with reminders at every item was significantly different from the PROMIS standard "no recall" condition (Table 3; $p < .01$).

Results for the item-level analyses are provided in the Supplementary Materials (Table S1). In the general population sample, the overall number of non-trivial differences at the item level was small (6 of 124 differences with $d > |0.2|$; 1 with $d > |0.3|$). At the total score level, the 24-hour recall condition with reminders was most distinct from the PROMIS no recall standard. However, 5 of the 6 differences with $d > |0.2|$ among individual items were actually in the 7-day recall condition without reminders. In all five cases, the 7-day recall condition with reminders had lower (worse) average responses than the no recall condition. We note with caution (due to the post hoc exploratory nature of these analyses) that these 5 items are targeted to moderate-to-strenuous physical activities (e.g., "doing 2 hours of physical labor", "running, lifting heavy objects"). In the cancer sample, the overall number of non-trivial ($d \geq |0.2|$) differences at the item level was also relatively small (19 of 124 differences with $d > |0.2|$). All but one of these differences were in the 24-hour recall conditions and reflected higher (better) physical function than responses in the no recall condition.

In our cognitive interviews, most patients considered the absence of a directed recall period as appropriate when responding to the 31-item custom PROMIS Physical Function questionnaire. Specifically, patients were asked the following question regarding recall period appropriateness: "In general, when thinking about your ability to carry out the physical activities we discussed today, is it better to consider a specific timeframe or respond according to your current capability?" In response, the majority of participants (74.2%) reported it is better to consider or respond to one's current capability (without a directed recall period) as opposed to a specific timeframe. Patients further explained that responding to one's current capability is ideal for considerations of physical function. The minority of patients ($n = 8$, 25.8%) who reported it is better to consider a specific timeframe were asked what timeframe they recommend for responding to questions of physical function. Patients' recommendations included a range of different timeframes such as, time since diagnosis, overall experience, during treatment, since completing treatment, and past 1–2 months. Of the alternative recall period suggestions made by patients, none are appropriate for assessing measurable improvements in physical function in a clinical trial setting.

Moreover, based on participant responses to questions regarding recall period, a majority ($n = 18$; 58%) considered their current capability when responding to questions of physical function. In addition to responding according to their current capability, participants reported considering physical capability since diagnosis or treatment (16.13%) or over the past weeks or months (12.9%) when responding. Of the 17 patients who were asked whether it was difficult to respond to the questionnaire without a directed recall period, 16 (94.1%) reported having no problems.

Discussion

PROMIS provides a framework for measuring a range of common symptoms and functional abilities, including a large item bank for physical function. To help respondents answer questions that include physical capabilities that may not have been experienced in the recent past, PROMIS has opted to use a present tense, "no recall period" framing for each item. This study evaluated whether or not a recall period (7 days or

Table 3 Multiple regression model estimated means and differences based on sample and population

Recall period and reminder administered	General population sample			Cancer sample		
	Mean	Difference	SE	Mean	Difference	SE
No recall	44.8			38.9		
24-h recall with reminders	46.9	2.1	0.82	41.2	2.3	.78
24-h recall without reminders	45.0	0.2	0.82	40.5	1.6	.78
7-day recall with reminders	45.3	0.5	0.83	39.9	1.0	.78
7-day recall without reminders	44.1	−0.7	0.82	40.3	1.5	.78

Differences and standard errors of the differences shown for each condition are relative to the no recall condition in each sample. All estimates reflect adjustment (relative to the values shown in Table 2) for age, gender, education, race, English fluency, and the presence of co-morbidities. Estimates in the cancer sample are further adjusted for the time since diagnosis, primary cancer site, and the type of treatment

24 hours) makes any measurable difference in the way people respond to individual questions relative to one another, and whether it affects the score one would obtain on the PROMIS physical function metric. We found no important differences on physical function item responses, or physical function score, across the studied recall periods, suggesting that recall period has little to no effect.

Analyses of variance in the general population sample indicated a significant difference among groups but subsequent analyses suggested that the differences were small and difficult to interpret. Among the group-level mean scores for all 31 PROMIS Physical Function items, only the 24-hour recall condition without reminders was significantly different from the standard PROMIS no specified recall period. The item-level analyses indicated that the items with non-trivial differences had small effects, and that they indicated lower physical function when they used 7-day recall conditions with reminders for infrequent moderate-to-strenuous activity. This suggests evidence for a small effect among individuals who infrequently engage in physically demanding tasks. When consistently reminded to reference one's response to the past week, respondents reported slightly worse physical function. It is unclear whether this constitutes over-reporting of physical function difficulties based on experience with the exemplars in these more difficult items or the potential influence of consistent reminders.

In the cancer sample, the analysis of variance results were not statistically significant. The item-level analyses indicated relatively few non-trivial differences in means relative to the PROMIS standard protocol, and these differences were present under different conditions than those found in the general population. That is, the non-trivial differences in the cancer sample were generally present only in the 24-hour recall conditions (with and without reminders), and these were all in the direction of higher physical functioning. It seems that the use of brief recall periods may have prompted respondents in the cancer sample to evaluate their physical function more positively than the other conditions. While scores in the cancer sample were more than 1 SD lower than those in the general population sample on average, it seems likely that slightly higher scores in the 24-hour condition for the cancer sample are less generalizable (i.e., more specific to very recent experiences) than those in the other conditions.

Future studies could further inform these findings by targeting additional patient populations, examining effects of anti-cancer treatment, employing alternative methods of data collection, and/or incorporating a broader range of content relating to physical function. It would be particularly useful to evaluate whether the evidence for slightly elevated scores in the 24-hour conditions subgroups in the cancer sample would be maintained in a longitudinal sample with daily assessments over a 7-day period as this would provide evidence regarding the generalizability (or its absence) when using these shorter recall periods outside of longitudinal data collection.

Limitations of this work should also be noted. These analyses did not differentiate among the many forms of cancer, the severity of the cancer being treated, or features of the treatment regimen. It is possible that these and other characteristics may influence responses to different recall periods. Further, generalizability of these findings may be limited to the population represented by this relatively young online sample. Replication of this study in clinical and community-based samples would be reassuring.

The results presented herein support the use of the standard PROMIS "no recall period" approach to measuring physical function. In both cancer and general population samples, there were no score differences when a 7-day recall was compared to no recall for the overall score. This was true even when respondents were reminded of the recall period with each item. When compared to 24-hour recall, there was a small but statistically significant difference such that the "every item reminder" group indicated better physical function by 2.1 T-score units. Indeed, the 2.1 point difference exceeded our a priori 2 point difference; however, we note the 0.21 effect size of that difference is quite small. Other than this, there was little to no evidence of meaningful differences between the current PROMIS standard protocol for Physical Function (no recall period) and alternative recall periods. In cases where differences were suggested, it seems that the absence of a recall period tends to elicit slightly lower (worse) physical function scores. We could find no evidence, in the cancer sample or the general population, to suggest that there are substantial differences between the standard PROMIS "no recall" condition and the use of a 7-day recall. We believe that an important finding of this study is that the magnitude of observed differences between recall conditions was small to negligible across all conditions. The "no recall" condition is the standard context in PROMIS Physical Function assessment; however, there is insufficient evidence to suggest that use of a 24-hour or 7-day recall period substantially alters the assessment. This was especially true in the cancer sample.

Funding The funding was provided by National Institutes of Health (Grant No. U2CCA186878), AbbVie, Amgen, AstraZeneca, Bayer, Bristol-Myers Squibb, Genentech, Janssen, Merck, Novartis, and Pfizer.

References

1. Weeks, W. B., & Weinstein, J. N. (2016). Patient-reported data can help people make better health care choices. *New England Journal of Medicine*. Retrieved from https://catalyst.nejm.org/

patient-reported-data-can-help-people-make-better-health-care-choices/. Accessed 18 Dec 2017.
2. Butt, Z., & Reeve, B. (2012). *Enhancing the patient's voice: Standards in the design and selection of patient-reported outcomes measures (PROMs) for use in patient-centered outcomes research*. Patient-Centered Outcomes Research Institute.
3. Food and Drug Administration. (2009). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register, 74*(235), 65132–65133.
4. Brundage, M., Blazeby, J., Revicki, D., Bass, B., de Vet, H., Duffy, H., et al. (2013). Patient-reported outcomes in randomized clinical trials: Development of ISOQOL reporting standards. *Quality of Life Research, 22*(6), 1161–1175. https://doi.org/10.1007/s11136-012-0252-1.
5. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(Suppl 1), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55.
6. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). Initial adult health item banks and first wave testing of the patient-reported outcomes measurement information system (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.
7. Calvert, M., Kyte, D., Mercieca-Bebber, R., Slade, A., Chan, A.-W., King, M. T., et al. (2018). Guidelines for inclusion of patient-reported outcomes in clinical trial protocols. *JAMA, 319*(5), 483–494. https://doi.org/10.1001/jama.2017.21903.
8. Cook, K., Jensen, S. E., Schalet, B. D., Beaumont, J. L., Amtmann, D., Czajkowski, S., et al. (2016). PROMIS measures of pain, fatigue, negative affect, physical function, and social function demonstrated clinical validity across a range of chronic conditions. *Journal of Clinical Epidemiology, 73,* 89–102. https://doi.org/10.1016/j.jclinepi.2015.08.038.
9. Jensen, R. E., Potosky, A. L., Reeve, B. B., Hahn, E., Cella, D., Fries, J., et al. (2015). Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Quality of Life Research, 24*(10), 2333–2344. https://doi.org/10.1007/s11136-015-0992-9.
10. Lai, J.-S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). 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, 92*(10), S20–S27. https://doi.org/10.1016/j.apmr.2010.08.033.
11. Broderick, J. E., Schneider, S., Junghaenel, D. U., Schwartz, J. E., & Stone, A. A. (2013). Validity and reliability of patient-reported outcomes measurement information system (PROMIS) instruments in osteoarthritis. *Arthritis Care & Research, 65*(10), 1625–1633. https://doi.org/10.1002/acr.22025.
12. Rothrock, N., Hays, R., Spritzer, K., Yount, S., Riley, W., & Cella, D. (2010). Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the patient-reported outcomes measurement information system (PROMIS). *Journal of Clinical Epidemiology, 63*(11), 1195–1204. https://doi.org/10.1016/j.jclinepi.2010.04.012.
13. Reeve, B., Hays, R., Bjorner, J., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care, 45*(5), S22–S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04.
14. Schalet, B. D., Revicki, D. A., Cook, K. F., Krishnan, E., Fries, J. F., & Cella, D. (2015). Establishing a common metric for physical function: Linking the HAQ-DI and SF-36 PF subscale to PROMIS physical function. *Journal of General Internal Medicine, 30*(10), 1517–1523. https://doi.org/10.1007/s11606-015-3360-0.
15. Riley, W., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E., et al. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research, 19*(9), 1311–1321. https://doi.org/10.1007/s11136-010-9694-5.
16. Hays, R. D., Spritzer, K. L., Schalet, B. D., & Cella, D. (2018). PROMIS-29 v2.0 profile physical and mental health summary scores. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-1842-3.
17. Cella, D., Lai, J.-S., Jensen, S. E., Christodoulou, C., Junghaenel, D. U., Reeve, B. B., et al. (2016). PROMIS® fatigue item bank has clinical validity across diverse chronic conditions. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2015.08.037.
18. Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. *Journal of Clinical Epidemiology, 63*(11), 1169–1178. https://doi.org/10.1016/j.jclinepi.2009.11.021.
19. Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS anxiety. *Journal of Anxiety Disorders, 28,* 88–96. https://doi.org/10.1016/j.janxdis.2013.11.006.
20. DeWalt, D., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12–S21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2.
21. Garcia, S. F., Cella, D., Clauser, S. B., Flynn, K. E., Lad, T., Lai, J. S., et al. (2007). *Standardizing patient-reported outcomes assessment in cancer clinical trials: A patient-reported outcomes measurement information system initiative, 25*(32), 5106–5112. https://doi.org/10.1200/JCO.2007.12.2341.
22. Schwarz, N., & Sudman, S. (2012). *Autobiographical memory and the validity of retrospective reports*. New York: Springer.
23. Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236*(4798), 157–161. https://doi.org/10.1126/science.3563494.
24. Erskine, A., Morley, S., & Pain, S. P. (1990). Memory for pain: A review. *Pain, 41,* 255–265. https://doi.org/10.1016/0304-3959(90)90002-U.
25. Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*(6), 934–960. https://doi.org/10.1037/0033-2909.128.6.934.
26. Gorin, A. A., & Stone, A. A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. *Handbook of Health Psychology, 23,* 405–413.
27. Redelmeier, D. A., & Pain, D. K. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66,* 3–8. https://doi.org/10.1016/0304-3959(96)02994-6.
28. Menon, G., & Yorkston, E. A. (1999). The use of memory and contextual cues in the formation of behavioral frequency judgments. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report implications for research and practice* (pp. 63–79). Mahwah: Lawrence Erlbaum Associates Publishers.

29. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the patient-reported outcomes measurement information system (PROMIS). *Journal of Clinical Epidemiology, 61*(1), 17–33. https://doi.org/10.1016/j.jclinepi.2006.06.025.
30. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E., Jr. (2014). The PROMIS physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526. https://doi.org/10.1016/j.jclinepi.2013.10.024.
31. de Vries, S. T., Haaijer-Ruskamp, F. M., de Zeeuw, D., & Denig, P. (2014). The validity of a patient-reported adverse drug event questionnaire using different recall periods. *Quality of Life Research, 23*(9), 2439–2445. https://doi.org/10.1007/s11136-014-0715-7.
32. Broderick, J. E., Schwartz, J. E., Vikingstad, G., Pribbernow, M., Grossman, S., & Stone, A. A. (2008). The accuracy of pain and fatigue items across different reporting periods. *Pain, 139*(1), 146–157. https://doi.org/10.1016/j.pain.2008.03.024.
33. Broderick, J. E., Schneider, S., Schwartz, J. E., & Stone, A. A. (2010). Interference with activities due to pain and fatigue: Accuracy of ratings across different reporting periods. *Quality of Life Research, 19*(8), 1163–1170. https://doi.org/10.1007/s11136-010-9681-x.
34. Stull, D. E., Leidy, N. K., Parasuraman, B., & Chassany, O. (2009). Optimal recall periods for patient-reported outcomes: Challenges and potential solutions. *Current Medical Research and Opinion, 25*(4), 929–942. https://doi.org/10.1185/03007990902774765.
35. Norquist, J. M., Girman, C., Fehnel, S., DeMuro-Mercon, C., & Santanello, N. (2012). Choice of recall period for patient-reported outcome (PRO) measures: Criteria for consideration. *Quality of Life Research, 21*(6), 1013–1020. https://doi.org/10.1007/s11136-011-0003-8.
36. Lai, J.-S., Cook, K., Stone, A., Beaumont, J., & Cella, D. (2009). Classical test theory and item response theory/Rasch model to assess differences between patient-reported fatigue using 7-day and 4-week recall periods. *Journal of Clinical Epidemiology, 62*(9), 991–997. https://doi.org/10.1016/j.jclinepi.2008.10.007.
37. Batterham, P. J., Sunderland, M., Carragher, N., & Calear, A. L. (2017). Psychometric properties of 7- and 30-day versions of the PROMIS emotional distress item banks in an Australian adult sample. *Assessment*. https://doi.org/10.1177/1073191116685809.
38. Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & M. Connor (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). New York: Guilford Press.
39. Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six patient-reported outcomes measurement information system-cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64*(5), 507–516. https://doi.org/10.1016/j.jclinepi.2010.11.018.
40. Northwestern University. (2018). HealthMeasures: Transforming how health is measured. Retrieved from http://www.healthmeasures.net/. Accessed 11 Jan 2018.
41. Hays, R., Bjorner, J., Revicki, D., Spritzer, K., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *European Journal of Cancer, 18*(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9.
42. Yanez, B., Pearman, T., Lis, C. G., Beaumont, J. L., & Cella, D. (2012). The FACT-G7: A rapid version of the functional assessment of cancer therapy-general (FACT-G) for monitoring symptoms and concerns in oncology practice and research. *Annals of Oncology, 24*(4), 1073–1078. https://doi.org/10.1093/annonc/mds539.
43. Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238.
44. Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*(4), 447–460. https://doi.org/10.1007/s11136-009-9464-4.
45. Hatcher, L. (1994). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Cary: SAS Institute Inc.
46. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling a Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.
47. Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
48. McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Lawrence Erlbaum Associates Inc.
49. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(Suppl 1), 19–31. https://doi.org/10.1007/s11136-007-9183-7.
50. Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477.
51. Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437.
52. Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045.
53. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software, 39*(8), 1–30.
54. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp. Released 2017.
55. Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge: Cambridge University Press.