The Performance and Association Between Patient-reported and Performance-based Measures of Physical Functioning in Research on Individuals with Arthritis

Laura C. Pinheiro, Leigh F. Callahan, Rebecca J. Cleveland, Lloyd J. Edwards, and Bryce B. Reeve

ABSTRACT. ***Objective.*** To evaluate the association between patient-reported outcome (PRO) and performance-based (PB) measures of physical functioning (PF) among individuals with self-identified arthritis to inform decisions of which to use when evaluating the effectiveness of a physical activity intervention.

Methods. Secondary data analysis of a nonrandomized 2-arm pre-post community trial of 462 individuals who self-identified as having arthritis and participated in the Walk with Ease (WWE) intervention. Two PRO and 8 PB assessments were collected at baseline (preintervention) and at 6-week followup. We calculated correlations between PB and PRO measures, assessed how measures identified changes in PF from baseline to followup, and compared PRO and PB measures to arthritis symptoms of pain, stiffness, and fatigue.

Results. Strength of correlations between PB and PRO measures varied depending on the PB measure, ranging from 0.21 to 0.54. PRO and PB measures identified PF improvements from baseline to followup, but none showed significant differences between the 2 WWE modalities (instructor-led or self-directed groups). Correlations with arthritis symptoms were stronger for PRO (0.30–0.46) than PB measures (0.03–0.31).

Conclusion. PRO measures may provide us with insights into aspects of PF that are not identified by PB measures alone. Use of PRO measures allows patients to communicate their perceptions of PF, which may provide a more accurate representation of overall PF. Our study does not suggest abandoning the use of PB measures to characterize PF in patients with self-identified arthritis, but recommends that PRO measures may serve as complementary or surrogate endpoints for some studies.

(J Rheumatol First Release December 1 2015; doi:10.3899/jrheum.150432)

Key Indexing Terms:
ARTHRITIS
PATIENT-REPORTED OUTCOMES
PERFORMANCE-BASED MEASURES
PHYSICAL FUNCTIONING

Researchers conducting evaluations of behavioral interventions such as exercise often include physical functioning (PF) as an outcome to determine effectiveness. PF is a unique endpoint because it can be assessed by a variety of methods, including observable performance-based (PB) tasks and patient-reported outcome (PRO) assessments using measures such as the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS).

Each assessment method has strengths and limitations. PB measures of PF are directly observable and use well-accepted metrics such as "time to complete" a task measured by a stopwatch; however, PB measures require an observer to be present. Thus, PB assessments are often scheduled in clinic. Additionally, for multisite studies, there may be observer-to-observer measurement error when setting up the task or using the timer. PRO measures of PF have the advantage of a consistent measure (e.g., everyone uses the PROMIS PF short form) and the measure can be completed in clinic, at assessment sites, or from home as often as justifiably needed.

From the Department of Health Policy and Management, and Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill; Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill; School of Medicine, Thurston Arthritis Research Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

Supported by a cooperative agreement between the US Centers for Disease Control and Prevention and the Association of American Medical Colleges (MM-0975-07/07).

L.C. Pinheiro, MPH, Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill; L.F. Callahan, PhD, School of Medicine, Thurston Arthritis Research Center, University of North Carolina at Chapel Hill; R.J. Cleveland, PhD, School of Medicine, Thurston Arthritis Research Center, University of North Carolina at Chapel Hill; L.J. Edwards, PhD, Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill; B.B. Reeve, PhD, Health Policy and Management, Gillings School of Global Public Health, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill.

Address correspondence to Dr. L.C. Pinheiro, Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Campus Box 7411, 1101-D McGavran-Greenberg, Chapel Hill, North Carolina 27599-7411, USA. E-mail: lpinheir@live.unc.edu

Accepted for publication September 30, 2015.

Also, PRO allows the patient to evaluate their own performance and limitations, and improving a PRO score may be more meaningful than improving a PB score. However, PRO measures are considered subjective and may be more likely than PB measures to be biased based on a respondent's characteristics such as sex, age, or race/ethnicity[1]. Sources of bias may relate to over- or under-estimation of PF or differences in ways groups of people interpret particular items[2]. Some PRO questions such as "going up or down stairs" may also present problems if the patient has not had the opportunity to do the task.

Given the advantages and limitations of both types of assessments, researchers may feel inclined to include both in a research study; however, selecting both poses additional costs to collect data and further challenges to determine which method (PRO or PB) should be used as the primary endpoint to determine treatment effectiveness. If choosing 1 assessment method, which one? The goal of our study was to inform this decision through an analysis of secondary data generated from the evaluation of a walking physical activity intervention in adults with arthritis.

Arthritis, the most common cause of disability in the United States, especially among older adults, is an exemplar for addressing these issues[3]. The prevalence of arthritis has grown with the increase in obesity and it is projected that by 2030, 67 million adults will be affected by arthritis[3]. In addition to several uncomfortable symptoms, functional disability is a serious consequence of arthritis, which many clinical tests are unable to identify[4].

The parent study evaluated a 6-week, community-based program Walk With Ease (WWE), which was developed to help individuals with arthritis to reduce symptoms[5,6]. WWE aimed to educate those affected by arthritis about the benefits of physical activity, increase awareness of symptom management, and offer a convenient, low-cost, moderate-intensity fitness regimen[5,7]. The parent study collected PRO and PB assessments of PF, along with other arthritis-relevant symptoms, at baseline and 6-week followup (end of study). Published results used PB and Health Assessment Questionnaire (HAQ) measures and found that WWE improved PF over time[6].

With these data, our study addressed the following research questions that will inform future decisions on the use of specific PB and PRO-based measures of PF in research studies:

(1) Research Question 1: Do PB and PRO measures identify the same concept of PF? This will be addressed through looking at associations between PB and PRO measures.

(2) Research Question 2: Do PB and PRO measures provide similar results of the evaluation of the effectiveness of the WWE intervention over time and between modalities? This will be evaluated using standardized effect size estimates of change over time and differences-in-differences estimates between modalities for PB and PRO measures individually.

(3) Research Question 3: Which method for measuring PF is more associated with arthritis-relevant symptoms of pain, stiffness, and fatigue? This will be evaluated through looking at associations of PB and PRO measures with self-reported symptom measures.

MATERIALS AND METHODS

Data. Data came from a nonrandomized 2-arm pre-post community trial of individuals with self-reported arthritis[6]. The study enrolled nearly 500 participants and was conducted at 33 sites. Participants were aged 18 years and older, English-speaking, cognitively able, and did not have serious medical conditions beyond arthritis[6]. Participants self-selected to be in the instructor-led or self-directed group. Baseline PB and PRO-based assessments were collected on the same day and followup assessments were collected 6 weeks later[6]. Institutional Review Board permission was obtained from the University of North Carolina.

PRO measures. The PROMIS PF measure was collected using Computerized Adaptive Testing (CAT) technology available in the Assessment Center and used a current recall period[8]. Participants completed about 5 questions per CAT[9]. Questions selected in the CAT were based on maximum posterior-weighted information criteria and the CAT stopped when the maximum standard error was 0.3[10]. CAT tailored assessments based on an individual's response to each question so administered items maximized the ability of the PROMIS CAT to measure a person's PF with the minimal number of questions[8]. Completed items were scored based on item response theory-calibrated variables to derive a PROMIS PF T score metric using the expected *a posteriori* estimator[11]. T scores have a mean of 50 (SD 10) in the US general population with higher scores reflecting better PF. An example of a PROMIS item is: "Are you able to walk a block on flat ground?" Response options include "1. Unable to do," "2. With much difficulty," "3. With some difficulty," "4. With a little difficulty," and "5. Without any difficulty.[9]"

The HAQ measure of PF used a 7-day recall period and was collected using paper-based surveys at community sites[6]. The HAQ includes 20 items that sum together (0 to 60), with higher scores representing poorer PF[12]. Of the 20 items, 11 cover upper body mobility (e.g., shampoo hair, open a new milk carton) and 9 cover lower body mobility (e.g., climb up 5 steps, walk outdoors on flat ground)[12]. An example HAQ item is: "What is your ability to carry out daily activities?" Response options are "0. Without any difficulty," "1. With a little difficulty," "2. With some difficulty," "3. With much difficulty," and "4. Unable to do.[12]"

A visual analog scale (VAS) was used to measure 3 symptoms reported by patients with arthritis: pain, stiffness, and fatigue[13,14]. VAS uses a 100-mm line and participants are asked to mark a spot on the line reflecting pain experienced in the last 7 days[13,14]. The line ranges from "no pain" (furthest point left) to "pain as bad as it could be" (furthest point right)[13,14]. The same type of scale was used for stiffness and fatigue symptoms with higher scores indicating greater stiffness or fatigue[6,13,14].

PB measures. Eight PB measures administered by a trained assessor were intended to assess several PF components. The assessor was blinded to intervention assignment for each patient. The 8 tests included timed chair stands, timed left and right turns, left and right single-leg standing assessments, 4 walking speed tests over a 20-foot stretch, and finally the 2-min step test[15,16,17,18]. Timed chair stands, turn tests, and single-leg stands were measured in seconds[15,16]. Walking speed was measured in meters per second, and the 2-min step test was measured in number of steps in 120 s[15,16,17]. Three timed chair stands assessed lower extremity strength[6]. The 360° turn tests and single-leg stands assessed balance[15,16,18]. The normal (average of 2 tests) and fast walking (average of 2 tests) scores measured one's functional mobility, and lastly, the step test measured an individual's aerobic endurance[15,16]. For most PB measures, higher scores indicated better PF, but for chair stands and right and left turns, higher scores indicated worse PF.

Traditionally, the 8 PB measures were used together to reliably assess an individual's PF[15,16,17,18].

Missing data. For both PRO and PB measures, most missing data were from followup assessments. At baseline, 4% were missing HAQ or PROMIS scores, and at followup, 14% were missing a HAQ score and 29% missing a PROMIS score. At baseline, 2–8% of PB measures had missing data, and at followup, 33–37% of PB measures had missing data. Using chi-square tests to evaluate associations between missing data and covariates listed in Table 1, we could not determine a missing-data pattern by demographics or baseline PF so we assumed it was missing at random and used complete case analysis.

Sample characteristics. Self-reported demographic characteristics included age, sex, marital status, race/ethnicity, level of education, and body mass index (BMI). In analyses, age and BMI were continuous. Education was grouped as less than high school, high school graduate, and more than high school. Marital status was dichotomized as married or not. Race/ethnicity was grouped as non-Hispanic white, African American, and other (Hispanic, Asian, multi-race/unknown race).

Statistical analysis. Unadjusted comparisons of demographic characteristics between instructor-led and self-directed walking groups were conducted using chi-square and Student t tests.

Research Question 1: Pearson correlations were calculated between PRO and PB measures at baseline and followup for the entire cohort and stratified by WWE modality. Correlation strength was defined as weak (0.0 to < 0.01), modest (0.1 to < 0.3), moderate (0.3 to < 0.5), strong (0.5 to < 0.8), and very strong (0.8–1.0)[19].

Research Question 2: To compare measures with different metrics, PB and PRO measures were individually standardized to Z-scores by subtracting each measure's mean and dividing by the SD. Standardized effect sizes from baseline to 6-week followup were calculated for PB and PRO measures and stratified by WWE modality. Within-modality effect size (ES) was calculated as the difference between average baseline and followup scores dividing by the baseline's SD[6]. We used Cohen classification of ES magnitude: < 0.32 was considered "small," 0.33–0.55 "medium," and 0.56–1.2 "large"[20]. A difference-in-difference model using standardized scores evaluated differences between PB and PRO measures of PF in identifying change in PF between WWE modalities. Sensitivity analyses adjusting for demographic covariates were conducted, but results are not shown.

Research Question 3: Pearson correlations between PB and PRO measures with VAS symptoms (pain, fatigue, and stiffness) were calculated. We expected there would be moderate associations between PF and these symptoms.

Analyses were performed in Stata (version 13.1) with 2-sided statistical tests and a significance level of 5%.

RESULTS

Participant characteristics. There were 462 adults who self-identified as having arthritis and who participated in our study. Respectively, the instructor-led and self-directed groups included 192 and 270 adults. Demographic characteristics are shown in Table 1. Marital status, BMI, and race were similarly distributed between WWE modalities. In both groups, there were higher proportions of women (85% and 90%). Level of education was significantly different between groups, with the instructor-led being less educated than the self-directed group. Further, the instructor-led group tended to be older, with mean age of 70.6 years compared with the self-directed group mean age of 64.4 years.

Table 1. Cohort characteristics at baseline. Chi-square tests and Student t tests for differences between instructor-led and self-directed groups. Some percentages do not add up to 100 because of rounding.

Characteristics	All Participants		Instructor-led Group		Self-directed Group		p
Baseline, mean (SD)							
PROMIS	42.8 (6.5)		42.4 (6.8)		43.2 (6.2)		0.1997
HAQ	14.5 (13.7)		16.2 (15.3)		13.2 (12.4)		0.0213*
Fatigue	36.8 (28.4)		37.7 (29.6)		36.3 (27.5)		0.5972
Stiffness	37.6 (25.4)		37.6 (27.4)		37.6 (23.9)		0.995
Pain	41.9 (26.8)		41.0 (27.8)		42.7 (26.1)		0.4857
Age, yrs, mean (SD)	66.9 (11.5)		70.6 (9.9)		64.4 (11.8)		< 0.0001*
BMI, kg/m^2, mean (SD)	29.7 (6.8)		29.3 (6.4)		30.1 (7.0)		0.1128
Characteristics	n = 462	%	n = 192	%	n = 270	%	p
Sex							0.1715
Male	56	12	28	15	28	10	
Female	406	88	164	85	242	90	
Education							0.0032*
< High school	21	5	14	7	7	3	
High school	105	23	53	28	52	19	
> High school	336	73	125	65	211	78	
Marital status							0.0986
Married	240	52	91	47	149	55	
Not married	222	48	101	53	121	45	
Race/ethnicity							0.9876
White	325	70	135	70	190	71	
African American	117	25	49	26	68	25	
Other	20	4	8	4	12	4	

* Statistical significance at 0.05. PROMIS: Patient-Reported Outcomes Measurement Information System; HAQ: Health Assessment Questionnaire; BMI: body mass index.

PROMIS and HAQ correlations. Baseline PROMIS and HAQ scores ranged from 26–63 and 0–56.25, respectively. Followup PROMIS and HAQ scores ranged from 25–61.75 and 0–58.75, respectively. Correlations between PROMIS and HAQ scores were negative because higher PROMIS scores indicate better PF, while larger HAQ scores indicate worse PF. At baseline, correlations between PRO measures of PF were strong at –0.68 for the overall cohort (instructor-led group: –0.64, self-directed group: –0.72). At followup, correlations were slightly stronger at –0.72 overall (instructor-led group: –0.71, self-directed group: –0.73).

Research Question 1: Do PB and PRO measures identify the same concept of PF? Because higher PROMIS scores indicate better PF and higher HAQ scores indicate worse PF, expected correlations between PRO and PB measures were positive for PROMIS and negative for HAQ for most PB outcomes. However, for chair stands and right/left turns, expected correlations between PB measures were negative with PROMIS and positive with HAQ.

As noted in Table 2, strength of correlations between PB and PRO measures varied depending on PB. Modest associations of PRO measures were observed for single-leg stances. Moderate correlations of PRO measures were observed for steps, chair stands, and right/left turns. Strong associations were observed for normal/fast walk. All correlations between PB and PRO measures were statistically significant ($p < 0.05$).

Research Question 2: Do PB and PRO provide similar conclusions of the evaluation of the effectiveness of the WWE intervention over time and between modalities? Unadjusted standardized ES for PB measures (except number of steps) showed small improvements from baseline to followup for the self-directed group and moderate improvements for the instructor-led group (Table 3). PRO measures of PF were consistent with small ES of 0.22–0.20 for PROMIS and 0.18–0.20 for the HAQ. Unadjusted standardized ES presented are similar in magnitude and direction to adjusted ES reported in the parent study[6].

Standardized differences-in-differences results are shown (Table 4). Neither PB nor PRO measures showed statistically significant differences in PF changes between WWE modalities over time (Table 4). Although both types of measures identify improvements, there were no significant differences between modalities. This conclusion was consistent across PB and PRO measures when demographic characteristics were included.

Research Question 3: Which method for measuring PF is more associated with VAS symptoms of pain, stiffness, and fatigue? As expected, correlations between PROMIS and VAS measures were negative and correlations between HAQ and VAS measures were positive (Table 5). Both PRO measures had stronger correlations with VAS measures than any PB measure (Table 5). However, correlations between PRO and VAS measures were moderate, ranging from absolute values of 0.30 to 0.46.

PB measures generally had poor correlations with VAS measures and some correlations were not statistically significant (Table 5). PB and VAS correlations ranged from absolute values of 0.03 to 0.31, with chair stands having the strongest correlations.

DISCUSSION

Within the context of evaluating the effectiveness of a walking intervention program for individuals with arthritis, our study examined 2 types of measures (PB and PRO) of PF. The overall goal is to inform investigators wishing to include similar endpoints in future studies. Our study examines how 8 types of PB and 2 PRO measures of PF are related and how they perform when measuring changes in PF over time and between 2 intervention modalities.

The first question we address is the extent to which PB and PRO measures in our study identify the same concept of PF. Fair to moderate correlations (0.21 to 0.49) were observed with higher associations between PB measures of normal and fast walking with both PRO measures. Lack of stronger associations between PB and PRO measures are not surprising because PB measures specific body parts or particular skills while PRO measures include questions combining several body parts and skills to provide a comprehensive representation of PF. For instance, timed chair stands hone in on lower extremity strength, which depicts 1 aspect of PF. PRO measures relate PF to activities of daily living, which attempts to convey a holistic view of PF.

Table 2. Pearson correlations between PROMIS and HAQ scores and PB measures. All are statistically significant at 0.05.

Variables	F-Walk	N-Walk	L-Turn	R-Turn	Chair	SLL	SLR	Steps
Baseline								
HAQ	–0.54	–0.52	0.46	0.46	0.47	–0.28	–0.24	–0.37
PROMIS	0.5	0.45	–0.41	–0.41	–0.41	0.25	0.26	0.39
Followup								
HAQ	–0.46	–0.45	0.46	0.47	0.49	–0.24	–0.21	–0.33
PROMIS	0.5	0.48	–0.45	–0.47	–0.47	0.25	0.31	0.41

PROMIS: Patient-Reported Outcomes Measurement Information System; HAQ: Health Assessment Questionnaire; PB: performance-based; F-walk: fast walk; N-walk: normal walk; L-turn: average time to turn left; R-turn: average time to turn right; Chair: time taken to complete 3 chair stands; SLL: single-leg stance left; SLR: single-leg stance right; Steps: number of steps taken in 2 min.

Table 3. Unadjusted standardized effect sizes from baseline to 6-week followup by intervention.

Variables	Unadjusted Standardized Effect Size*		
	Total Cohort	Self-directed	Instructor-led
Fast walk	0.24	0.19	0.35
Normal walk	0.29	0.24	0.4
Left turn	0.34	0.27	0.49
Right turn	0.33	0.28	0.47
Chair stands	0.33	0.32	0.37
Left leg stand	0.17	0.12	0.3
Right leg stand	0.23	0.19	0.38
No. steps	–0.05	–0.08	–0.02
PROMIS	0.2	0.22	0.2
HAQ	0.18	0.2	0.18

* Effect size is calculated as the difference between baseline and followup scores divided by the SD of the baseline score. Positive effect sizes denote improvements in physical functioning whereas negative effect sizes denote decrements. PROMIS: Patient-Reported Outcomes Measurement Information System; HAQ: Health Assessment Questionnaire.

The second question we examined was the ability of PB and PRO measures to detect changes over time and between intervention modalities. The WWE program was expected to improve PF from baseline to 6-week followup. Previous studies found that moderate-intensity exercises resulted in notable improvements in the strength, balance, and functional status of patients with arthritis[5]. In our study, we found that 6 of 8 PB measures showed small to moderate improvements (0.12–0.49) with slightly higher effect sizes for the instructor-led arm, while both PRO measures found small improvements (0.18–0.22) in both arms over time. Thus, most PB measures and both PRO measures correctly detected improvements in PF consistent with prior literature. To our knowledge, no prior studies have been conducted to determine differences between WWE modalities tested in the parent study. Neither PB nor PRO found statistically significant differences between modalities. Together, our comparisons show either PRO measure or the set of PB measures could be used to determine the WWE program effectiveness with respect to the study design.

The third question examined the association between PB and PRO measures of PF with VAS symptoms of pain, fatigue, and stiffness, which the WWE program aimed to reduce. Consistently, we found that PRO measures had stronger correlations than PB measures with pain (PRO 0.38–0.46, PB 0.03–0.31), fatigue (PRO 0.33–0.38, PB 0.05–0.20), and stiffness (PRO 0.30–0.41, PB 0.05–0.30). What may partly drive higher associations with PRO measures is that pain, fatigue, and stiffness were measured by self-report and we do not have clinical measures of each; however, it is accepted that the gold standard for measuring these symptoms is by self-report. Thus, findings support stronger evidence for construct validity (i.e., convergent validity) of PRO-based measures based on their association with clinically important arthritis symptoms.

Our findings appear to be consistent with published literature, confirming PRO are a viable way to measure PF. A study in patients with multiple sclerosis suggested PRO and PB measures access independent constructs of PF because of poor correlations between types of measures[21]. The study explained that PB measures focus on specific movements and do not allow us to identify overall quality of life, while PRO reflect PF beyond symptom effect[21]. Another study with patients with osteoarthritis following joint replacement recognized PRO better represent patient satisfaction because they relay the patient's own perception of PF[22]. Finally, a study of patients with osteoporosis compared different PRO and PB measures, found moderate correlations, and concluded PRO instruments identified changes in daily activities of PF "quite well"[23].

Limitations. There are limitations to our analyses. First, participants self-identified as having arthritis and we did not have clinical confirmation of diagnosis, which is a limitation if they are not similar to individuals with clinically diagnosed

Table 4. Standardized mean (standard error) differences from baseline to followup between the instructor-led and self-directed groups. There were no statistically significant differences between the self-directed and instructor-led groups.

Measures	Baseline		6-week Followup		Diff-in-diff
	Self-directed	Instructor-led	Self-directed	Instructor-led	
Fast walk	0.153 (0.06)	–0.212 (0.07)	0.082 (0.08)	–0.091 (0.08)	0.194 (0.15)
Normal walk	0.156 (0.08)	–0.216 (0.08)	0.092 (0.08)	–0.102 (0.08)	0.178 (0.16)
Left turn	–0.219 (0.06)	0.304 (0.07)	–0.195 (0.07)	0.219 (0.09)	–0.111 (0.14)
Right turn	–0.216 (0.06)	0.299 (0.07)	–0.179 (0.07)	0.202 (0.09)	–0.135 (0.14)
Chair stands	–0.125 (0.06)	0.169 (0.07)	–0.135 (0.07)	0.145 (0.09)	–0.014 (0.15)
Left leg stand	0.161 (0.07)	–0.231 (0.07)	0.112 (0.08)	–0.127 (0.08)	0.152 (0.15)
Right leg stand	0.166 (0.07)	–0.247 (0.06)	0.129 (0.08)	–0.143 (0.08)	0.141 (0.15)
No. steps	0.014 (0.06)	–0.019 (0.07)	–0.013 (0.08)	0.014 (0.08)	0.061 (0.15)
PROMIS	0.053 (0.06)	–0.069 (0.08)	0.061 (0.07)	–0.068 (0.09)	–0.006 (0.15)
HAQ	–0.091 (0.06)	0.127 (0.08)	–0.094 (0.06)	0.121 (0.08)	–0.002 (0.14)

Diff-in-diff: differences-in-differences; PROMIS: Patient-Reported Outcomes Measurement Information System; HAQ: Health Assessment Questionnaire.

Table 5. Pearson correlations between PROMIS and HAQ scores, PB measures, and VAS symptom measures. Pain, fatigue, and stiffness are the VAS measures.

Measures	Baseline			6-week Followup		
	Pain	Fatigue	Stiffness	Pain	Fatigue	Stiffness
Fast walk	**–0.24**	**–0.19**	**–0.18**	**–0.2**	**–0.16**	**–0.17**
Normal walk	**–0.2**	**–0.17**	**–0.19**	**–0.21**	**–0.17**	**–0.2**
Left turn	**0.21**	**0.19**	**0.22**	**0.2**	0.1	**0.13**
Right turn	**0.21**	**0.17**	**0.21**	**0.27**	**0.2**	**0.18**
Chair stands	**0.25**	**0.19**	**0.3**	**0.31**	**0.18**	**0.25**
Left leg stand	–0.06	–0.07	–0.07	**–0.13**	–0.1	**–0.15**
Right leg stand	–0.03	–0.05	–0.05	–0.07	–0.08	–0.08
No. steps	**–0.21**	**–0.2**	**–0.24**	**–0.26**	**–0.18**	**–0.18**
PROMIS	**–0.38**	**–0.36**	**–0.37**	**–0.43**	**–0.33**	**–0.3**
HAQ	**0.46**	**0.38**	**0.41**	**0.42**	**0.38**	**0.38**

Statistically significant data at the 0.05 level are in bold face. PROMIS: Patient-Reported Outcomes Measurement Information System; HAQ: Health Assessment Questionnaire; PB: performance-based; VAS: visual analog scale.

arthritis. In addition, participants self-selected to be in the instructor-led or self-directed group, which could lead to selection bias because treatment is not random. Another concern is discrepancy in sample size between men and women; however, this did not vary by WWE modality. There may be measurement error in the way PB and PRO measures were collected, which may affect results. We also do not know the extent to which these results generalize to PB and PRO measures not used in our study or to other therapeutic areas. We used an intent-to-treat approach and have no information about compliance with WWE, which could bias results if 1 group was less likely to comply. There were some missing data in the HAQ, PROMIS, and PB measures, with the self-directed group having a greater proportion of missing data than the instructor-led group. We could not find a relationship between missing data and demographics or baseline PF (i.e., people with worse PF having more missing data). We were also unable to adjust for clinical characteristics that could affect PF because these measures are not available in our dataset.

The parent study evaluation of the effectiveness of 2 WWE modalities would have yielded similar findings had it used 8 PB measures or 1 PRO measure. If time, costs, and participant and administrator burdens are irrelevant, investigators may wish to include both PB and PRO measures to provide comprehensive evaluations of the effect of the intervention on PF. However, time, costs, and burden are often challenges for studies. PB measures require (1) a trained assessor, thus necessitating measurements take place in a clinic or assessment site, (2) training of assessors in multisite studies and followup to maintain data collection consistency, (3) time burden to complete tasks, and (4) funds to pay assessors and participant incentives. Relatively, PRO (1) do not require an observer; however, an observer may have to be available for technical problems accessing surveys, (2) are shorter to complete (e.g., PROMIS CAT administered about 5 questions), and (3) questionnaires can be completed by participants more often in the convenience of the clinic or at home. Biased responses to PRO measures based on group characteristics such as age, sex, and race/ethnicity can be reduced using strong PRO measure design principles and psychometric evaluations.

Noting previously discussed limitations including that the study sample self-reported their arthritis conditions, our study provides support for the use of PRO measures of PF as indicators of treatment effectiveness in research studies. Costs and time are saved, relative to PB measures, to collect PRO data from patients at their convenience, especially when collecting other PRO endpoints such as fatigue and pain. Multisite trials will also benefit from consistent measures used across sites with electronic PRO data automatically stored in coordinating centers.

The use of PRO PF measures may allow us to glean insight into aspects of PF that are not identified by PB measures alone. Use of PRO measures allows patients to communicate their own perceptions of PF, which may lead to more accurate representations. Although our conclusions do not suggest abandoning the use of PB measures to identify PF, they suggest that PRO measures serve as complementary or surrogate endpoints.

REFERENCES

1. McDowell I. The theoretical and technical foundation of health measurement. In: Measuring health: a guide to rating scales and questionnaires, 3rd ed. New York: Oxford University Press; 2006.
2. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61:17-33.
3. Hootman JM, Helmick CG. Projections of US prevalence of arthritis and associated activity limitations. Arthritis Rheum 2006;54:226-9.
4. Guillemin F, Briançon S, Pourel J. Validity and discriminant ability of the HAQ Functional Index in early rheumatoid arthritis. Disabil Rehabil 1992;14:71-7.
5. Arthritis Foundation. Walk with ease: your guide to walking for better health, improved fitness and less pain. Atlanta: Arthritis Foundation; 1999.

6. Callahan LF, Shreffler JH, Altpeter M, Schoster B, Hootman J, Houenou LO, et al. Evaluation of group and self-directed formats of the Arthritis Foundation's Walk With Ease Program. Arthritis Care Res 2011;63:1098-107.
7. Bruno M, Cummins S, Gaudiano L, Stoos J, Blanpied P. Effectiveness of two Arthritis Foundation programs: Walk With Ease, and YOU Can Break the Pain Cycle. Clin Interv Aging 2006;1:295-306.
8. Gershon R, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment Center. Qual Life Res 2010;19:677-85.
9. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061-6.
10. van der Linden WJ, Ren H. Optimal Bayesian adaptive design for test-item calibration. Psychometrika 2015;80:263-88.
11. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph 1969;1:i169.
12. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
13. Lorig KS, Stewart A, Ritter P, Gonzalez VM, Laurent D, Lynch J. Outcome measures for health education and other health care interventions. Thousand Oaks: SAGE Publications Inc.; 1996.
14. Stewart AL, Ware JE Jr. Health perceptions, energy/fatigue, and health distress measures. In: Stewart AL, Ware JE Jr, ed. Measuring functioning and well-being: the medical outcomes study approach. Durham: Duke University Press Books; 1992:143-72.
15. Gill TM, Richardson ED, Tinetti ME. Evaluating the risk of dependence in activities of daily living among community-living older adults with mild to moderate cognitive impairment. J Gerontol A Biol Sci Med Sci 1995;50:M235-41.
16. Gill TM, Williams CS, Mendes de Leon CF, Tinetti ME. The role of change in physical performance in determining risk for dependence in activities of daily living among nondisabled community-living elderly persons. J Clin Epidemiol 1997;50:765-72.
17. Rikli RE, Jones CJ. Development and validation of criterion-referenced clinically relevant fitness standards for maintaining physical independence in later years. Gerontologist 2013;53:255-67.
18. Vellas BJ, Wayne SJ, Romero L, Baumgartner RN, Rubenstein LZ, Garry PJ. One-leg balance is an important predictor of injurious falls in older persons. J Am Geriatr Soc 1997;45:735-8.
19. Dancey CP, Reidy J. Statistics without maths for psychology: using SPSS for Windows. London: Prentice Hall; 2004.
20. Lipsey MW. Design sensitivity: statistical power for experimental research. Newbury Park: SAGE Publications; 1990.
21. Schwartz CE, Ayandeh A, Motl RW. Investigating the minimal important difference in ambulation in multiple sclerosis: A disconnect between performance-based and patient-reported outcomes? J Neurol Sci 2014;347:268-74.
22. Konan S, Hossain F, Patel S, Haddad FS. Measuring function after hip and knee surgery: the evidence to support performance-based functional outcome tasks. Bone Joint J 2014;96-B:1431-5.
23. Nixon A, Kerr C, Doll H, Naegeli AN, Shingler SL, Breheny K, et al. Osteoporosis Assessment Questionnaire-Physical Function (OPAQ-PF): a psychometrically validated osteoporosis-targeted patient reported outcome measure of daily activities of physical function. Osteoporos Int 2014;25:1775-84.