Published in final edited form as:
Psychol Test Assess Model. 2016 June 27; 58(2): 403–421.

Measuring social function in diverse cancer populations: Evaluation of measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Ability to Participate in Social Roles and Activities short form

Elizabeth A. Hahn[1], **Michael A. Kallen**[2], **Roxanne E. Jensen**[3], **Arnold L. Potosky**[3], **Carol M. Moinpour**[4], **Mildred Ramirez**[5], **David Cella**[6], and **Jeanne A. Teresi**[7]

[2]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL

[3]Department of Oncology, Georgetown University, and Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center, Washington, D.C

[4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA

[5]Research Division, Hebrew Home at Riverdale, NY, and Division of Geriatrics and Palliative Medicine, Weill Cornell Medical College, New York, NY

[6]Department of Medical Social Sciences and Center for Patient-Centered Outcomes, Northwestern University Feinberg School of Medicine, Chicago, IL

[7]Columbia University Stroud Center and New York State Psychiatric Institute, New York, NY, and Research Division, Hebrew Home at Riverdale, NY

Abstract

Conceptual and psychometric measurement equivalence of self-report questionnaires are basic requirements for valid cross-cultural and demographic subgroup comparisons. The purpose of this study was to evaluate the psychometric measurement equivalence of a 10-item PROMIS® Social Function short form in a diverse population-based sample of cancer patients obtained through the Measuring Your Health (MY-Health) study (*n* = 5,301). Participants were cancer survivors within six to 13 months of a diagnosis of one of seven cancer types, and spoke English, Spanish, or Mandarin Chinese. They completed a survey on sociodemographic and clinical characteristics, and health status. Psychometric measurement equivalence was evaluated with an item response theory approach to differential item functioning (DIF) detection and impact. Although an expert panel proposed that many of the 10 items might exhibit measurement bias, or DIF, based on gender, age, race/ethnicity, and/or education, no DIF was detected using the study's standard DIF criterion, and only one item in one sample comparison was flagged for DIF using a sensitivity DIF criterion. This item's flagged DIF had only a trivial impact on estimation of scores. Social function measures are especially important in cancer because the disease and its treatment can affect the

[1]Correspondence concerning this article should be addressed to: Elizabeth A. Hahn, PhD, Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 633 N. St. Clair St., Suite 1900, Chicago, IL 60611, USA; e-hahn@northwestern.edu.

quality of marital relationships, parental responsibilities, work abilities, and social activities. Having culturally relevant, linguistically equivalent and psychometrically sound patient-reported measures in multiple languages helps to overcome some common barriers to including underrepresented groups in research and to conducting cross-cultural research.

Keywords

patient-reported outcomes; social function; psychometrics; differential item functioning; cancer

Introduction

An Institute of Medicine report recommends development of standardized indicators focused on priority health outcomes (Institute of Medicine. Committee on Public Health Strategies to Improve Health, 2011). Several groups are working to identify and test concepts of health and function that are meaningful across countries and cultures (Taskforce on Health Status, 2005). This includes the Patient Reported Outcomes Measurement Information System® (PROMIS®; www.nihpromis.org) initiative, which is developing items and measures that can be used for making comparisons across ethnically diverse groups differing in sociodemographic and clinical characteristics, as well as diverse medical conditions. PROMIS methodology (DeWalt, Rothrock, Yount, Stone, & PROMIS Cooperative Group, 2007; Reeve et al., 2007) is consistent with the universalist model of cross-cultural equivalence (Herdman, Fox-Rushby, & Badia, 1998; Regnault & Herdman, 2015).

PROMIS adopted the World Health Organization framework to define three components of health: physical, mental, and social (World Health Organization, 1946; see http://www.nihpromis.org/measures/domainframework). Measures of social health will play a key role in applications that use ecologic (or determinants of health) models emphasizing how patients' environments influence their health (Institute of Medicine, 2003; Institute of Medicine. Committee on Public Health Strategies to Improve Health, 2011; Whitehead, 1995). Importantly, social determinants of health is now a topic for Healthy People 2020 (Healthy People 2020, 2015). This renewed emphasis on social health is of particular significance given that, historically, social health has been a relatively neglected domain. This is due to a lack of measures for clinical populations, as well as a fundamental disagreement about how best to define and measure social health (Hahn, Cella, Bode, & Hanrahan, 2010).

Social function measures are especially important in cancer since the disease and its treatment can affect the quality of marital relationships, parental responsibilities, work abilities, and social activities (Bouknight, Bradley, & Luo, 2006; Fantoni et al., 2010; McDowell, 2006; Munir, Yarker, & McDermott, 2009; Taskila, De Boer, Van Dijk, & Verbeek, 2011). Disparities in cancer burden continue to be documented among racial and ethnic minorities, and some cultural groups (American Cancer Society, 2015).

The PROMIS domain framework for Social Health (v2.0) includes two primary subcomponents: Social Function and Social Relationships (Hahn et al., 2014). As described in detail elsewhere, mixed methods were implemented to develop several sets of social function and social relationships items (Castel et al., 2008; DeWalt et al., 2007; Hahn, Cella, et al., 2010; Hahn, Devellis, et al., 2010; Hahn et al., 2014). English and Spanish versions of the social function items were tested in several large, diverse convenience samples manifesting varied clinical problems associated with functional limitation, and several online survey panels of general population respondents. Results revealed highly acceptable psychometric properties providing evidence of reliability and validity, and no evidence of measurement bias by gender, age, education, or language (Hahn et al., 2014).

The purpose of this study was to evaluate the psychometric measurement equivalence of a subset of PROMIS social function items (Ability to Participate in Social Roles and Activities) in a diverse population-based sample of cancer patients. Conceptual and psychometric measurement equivalence of self-report questionnaires are basic requirements for valid cross-cultural and demographic subgroup comparisons (Meredith, 1993; Meredith & Teresi, 2006; Stewart & Napoles-Springer, 2000; Teresi, Stewart, Morales, & Stahl, 2006; van de Vijver & Kwok, 1997). To our knowledge, this is the first study to measure social function in a large, ethnically diverse sample of people with cancer using three language versions of the PROMIS items (English, Spanish, and Mandarin Chinese).

Methods

Participant recruitment and assessment procedures

The Measuring Your Health (MY-Health) study recruited a population-based sample of cancer patients from four Surveillance, Epidemiology, and End Results (SEER) Program cancer registries in three states (California, Louisiana, New Jersey). A brief summary of the MY-Health study is provided here; complete details are provided elsewhere (Jensen et al., 2016). Sampling was stratified by four race-ethnicity groups (Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian/Pacific Islander, Hispanic) and three age groups (21–49, 50–64, 65–84). The study was approved by institutional review boards at each participating institution. Eligibility criteria were based on SEER cancer registry records and included: diagnosed with one of seven cancers (prostate, colorectal, non-small cell lung, non-Hodgkin’s lymphoma, female breast, uterine, or cervical); no prior cancer diagnosis (except non-melanoma skin cancer); and currently within six to 13 months of diagnosis. The SEER registry sites mailed English, Spanish, or Mandarin Chinese language surveys to eligible participants. Non-responders were contacted and given the option to complete their survey over the telephone. Survey content included self-reported sociodemographic and clinical information, and health status items. As an overall assessment of understandability and acceptability, participants were asked to indicate whether they needed help answering the written survey questions (*I answered all of the questions with no help; I answered all of the questions with some help from my parent, guardian, spouse, child or significant other; My parent, guardian, spouse, child or significant other answered all of the questions*). Each participant received a $30 incentive.

PROMIS measures

MY-Health focused on eight domains that are important to cancer outcomes and relevant to other chronic diseases: anxiety, cognitive function, depression, fatigue, pain interference, physical function, sleep disturbance, and social function. These domains were selected based on their prevalence, importance, and known variations across age, gender, race/ethnicity, or socioeconomic groups for several of the major cancers included in this study (McHorney & Cook, 2005; Moinpour & Provenzale, 2005; Patrick et al., 2004; Sprangers, Taal, Aaronson, & te Velde, 1995). Customized short form versions of each domain were developed. Item response theory (IRT) methods were used to create PRO-MIS item banks that allow for computer adaptive tests (CAT) and the creation of multiple short forms of varying length that serve to provide accurate measurement while minimizing response burden (Cella, Gershon, Lai, & Choi, 2007; Cella et al., 2007; Hambleton, Swaminathan, & Rogers, 1991; Reeve et al., 2007; Samejima, 1969; Thissen, 1991; van der Linden & Hambleton, 1997).

The 10 items of the custom short form for Social Function: Ability to Participate in Social Roles and Activities (Social Function: Ability-SF10) were chosen by members of the PROMIS Social Health Workgroup and the PROMIS Psychometrics Team (see Table 1). The criteria for item inclusion were content representativeness, maximized range of difficulty (inclusion of items across the IRT calibration range), and acceptable discrimination levels (inclusion of items that distinguish between people across the latent trait). These items were already available in English and Spanish (Hahn et al., 2014) and were translated into Mandarin Chinese for the MY-Health study. PROMIS translation methodology was used, which included a multi-step forward-backward process and cognitive debriefing interviews with five Chinese-speaking individuals (Eremenco, Cella, & Arnold, 2005; Wild et al., 2005).

Differential Item Functioning (DIF) hypotheses

DIF hypotheses were generated by asking a panel of content experts to indicate whether they expected DIF to be present, and the direction of that DIF, with respect to several comparison groups: gender, age, race/ethnicity, language, education, and diagnosis. A definition of DIF was provided, and the following instructions related to hypothesis generation were given:

> Differential item functioning means that individuals in groups with the same underlying trait (state) level will have different probabilities of endorsing an item. Put another way, reporting limitations in social function, e.g., limited social activities outside the home, should depend only on the level of the trait (state), e.g., level of social functioning, and not on membership in a group, e.g., male or female. Very specifically, randomly selected persons from each of the two groups (e.g., males and females) who are at the same (e.g., low) level of social functioning should have the same likelihood of reporting “limited social activities outside the home.” If it is theorized that reporting limitations in social function could depend to some extent on gender group membership, it would be hypothesized that the item has gender DIF.

The social function items were reviewed qualitatively by nine content experts regarding potential sources of DIF. Four members of this panel were clinical or counseling psychologists, three were public health professionals, one was a gerontologist, and one was a health behavior methodologist. The experts were asked to rate individually each of the 10 items with respect to gender, age, race/ethnicity, language, education, and diagnosis. Their summarized ratings provided this study's DIF hypotheses, in terms of both presence and direction of DIF. The goal was to identify items that might have a different meaning or not be understood well or equivalently by individual members of any of the groups referenced. A grid containing a row for each of the 10 items and separate columns for each of the referenced groups was distributed to the experts for completion in order to facilitate their ratings.

Psychometric and statistical analyses

Social Function: Ability-SF10 uses a five-point Likert-type "*never* to *always*" response option set. Item responses are scored as follows: *always* (1), *often* (2), *sometimes* (3), *rarely* (4), and *never* (5). Although all items are framed using language such as "I have trouble" or "I have to limit," higher scores indicate a greater ability to participate in social roles and activities. IRT-based Bayesian *expected a posteriori* (EAP) estimation response pattern scoring was conducted for the scale, employing previously established item parameters derived from the original Social Function: Ability item bank (Hahn et al., 2014). The two-parameter graded response model (GRM) was used for item calibration (Samejima, 1969). Social Function: Ability uses a T-score metric (mean = 50; standard deviation = 10; Hahn et al., 2014). The IRT software package IRTPRO was used for IRT-based scoring (Cai, Thissen, & du Toit, 2011).

Frequency distributions were evaluated for each item for range and completeness of category responses and for potential ceiling and floor effects. Internal consistency reliability was estimated using Cronbach's coefficient alpha (Nunnally & Bernstein, 1994). Reliability evidence was sought to support the use of Social Function: Ability-SF10 in this cancer patient sample for making appropriate group and individual case comparisons based on scale performance differences. Previous dimensionality assessments of Social Function: Ability included conducting an exploratory factor analysis (EFA) on one half of a randomly split sample and a confirmatory factor analysis (CFA) on the other half of the sample (Hahn et al., 2014). Findings from those complementary analyses supported the essential unidimensionality of Social Function: Ability required for DIF analyses, with the EFA displaying a dominant first factor, the single-factor CFA showing good model fit, and no residual correlations from the CFA analysis meeting or exceeding the 0.20 criterion for local dependence (Hahn et al., 2014). New dimensionality assessments were conducted with this study's cancer patient sample to confirm previous dimensionality findings and to provide additional support for Social Function: Ability-SF10's unidimensionality. Single-factor CFAs were conducted in LISREL using polychoric correlations and diagonally weighted least squares estimation (Joreskog & Sorbom, 2006). The following criteria were identified as representing "good" model fit: comparative fit index (CFI) > 0.95; non-normed fit index (NNFI) > 0.95; root mean square error of approximation (RMSEA) < 0.08; standardized root mean square residual (SRMR) < 1.0. Residual correlations meeting or exceeding a 0.20 criterion indicated inter-item local dependence. Analyses to evaluate criterion-related validity (Scientific Advisory Committee of the Medical Outcomes Trust et al., 2002) of Social Function: Ability-SF10 were conducted using Pearson correlations with PROMIS measures of physical function, sleep disturbance, anxiety, depression, fatigue, and pain interference.

DIF analysis was implemented to assess psychometric measurement equivalence (Camilli & Shepard, 1994; Holland & Wainer, 1993; Teresi, 2006; van de Vijver & Kwok, 1997). DIF was evaluated in a two-step process: Step One – detection, and Step Two – impact. Step One of the DIF analysis (detection) was to identify whether any Social Function: Ability-SF10 items displayed DIF by 18 sample characteristic groupings, e.g., gender, age. To conduct a DIF analysis, the minimum sample size for each DIF subgroup was set at $n = 200$. A novel hybrid "logistic ordinal regression (LOR)-plus-IRT" approach to DIF detection was implemented, using both a standard criterion and a more conservative sensitivity criterion. The DIF method uses an IRT-derived ability score for the LOR modeling, rather than the traditionally modeled summed-score ability term (Choi, Gibbons, & Crane, 2011). For standard DIF detection, a liberal McFadden pseudo-R^2 (McFadden, 1974) change criterion of 0.010 was used (see, for example, the McFadden pseudo-R^2 change criterion of 0.020 used by Paz and colleagues, 2013). For sensitivity DIF detection, to increase the ability to detect potential item bias, this criterion was then lowered by half to 0.005. LOR-based DIF detection employed model comparisons to identify DIF. Three relevant models were involved: Model 1, which used only ability to predict item performance; Model 2, which used ability plus group status (e.g., cancer stage) to predict item performance; and Model 3, which used ability, group status, and the ability-by-group status interaction to predict item performance. When comparing Models 1 vs. 2, if the McFadden pseudo-R^2 change criterion was met, uniform DIF was considered to be present (i.e., the biasing effect was constant across varying trait levels). If the McFadden pseudo-R^2 change criterion was met when comparing Models 2 vs. 3, non-uniform DIF was considered present (i.e., the biasing effect varied conditional on trait level). Thus, the use of logistic ordinal regression, a widely recommended DIF methodology, provided a flexible and comprehensive approach to DIF detection (Camilli & Shepard, 1994; Zumbo, 1999). It allowed for (a) the incorporation of IRT-derived ability estimates that were "purified" or adjusted in real time for any DIF items identified during the analytic process, (b) the addition of regression model terms (independent variables) that could identify both uniform (significant group status term) and non-uniform (significant ability-by-group status term) DIF, and (c) access to a wide-ranging set of accompanying model statistics and measures to address questions of statistical significance (e.g., chi-squared-based p values) and effect size (e.g., model change in regression beta coefficient and pseudo-R^2).

Step Two of the DIF analysis (impact) involved conducting score difference analyses to evaluate the impact of identified DIF on Social Function: Ability-SF10 total scores. A series of analyses were conducted, comparing unadjusted or "initial" Social Function: Ability-SF10 scores to DIF-adjusted or "purified" Social Function: Ability-SF10 scores. Unadjusted initial scores were based on the use of a common-across-groups set of item parameters for all items, while DIF-adjusted purified scores were based on the use of (a) common-across-groups item parameters for all non-DIF items and (b) group-specific item parameters for DIF-identified items. DIF impact evidence included: 1) Pearson correlation (initial vs. purified theta scores); 2) a median theta standard error (SE) assessment (the number and percentage of individual difference scores, i.e., initial theta minus purified theta that exceeded initial theta's median SE; 3) an individual theta score SE assessment (the number and percentage of individual difference scores that exceeded initial individual theta score

SEs); and 4) a comparison of Cohen's *d* group factor effect sizes across competing analyses of variance (ANOVA; i.e., initial theta scores by group factor vs. purified theta scores by group factor; Cook et al., 2011). The R package lordif (Choi, Gibbons, & Crane, 2012) and the statistical program SPSS (IBM Corporation, 2013) were used for conducting the DIF detection and impact analyses.

Results

Study participants

Over 5,000 people with diverse cancer diagnoses participated in the study and provided responses to Social Function: Ability-SF10 (see Table 2). There were fewer men than women, 40 % were age 65 or older, 42 % were non-Hispanic White, and 36 % had an educational attainment of High School or lower. The majority completed the questionnaire on paper in English without assistance. A small proportion of participants (< 2 %) completed the questionnaire by telephone interview.

Distributional, reliability, dimensionality, and validity analyses

All five response choices were observed (*always* to *never*) across all items of Social Function: Ability-SF10. About one-third of the responses for each item were *never*, indicating a report of no limitations in specific aspects of the ability to participate in social roles and activities (see Table 1). A total of 1,041 respondents (20 %) reported no limitations for all 10 Social Function: Ability-SF10 items, and 2 % (n = 123) reported that they *always* have limitations in all 10 items. The never-limited respondents plus the always-limited respondents (total n = 1,164) are referred to below as "extreme-score" respondents.

Social Function: Ability-SF10 exhibited excellent internal consistency reliability (Cronbach's coefficient alpha = 0.98); no item deletion improved alpha. Excluding extreme-score respondents for the internal consistency reliability analysis, Social Function: Ability-SF10 continued to exhibit excellent internal consistency reliability (alpha = .96); again, no item deletion improved alpha.

Results from the single-factor CFAs confirmed the previous finding of essential unidimensionality in the Social Function: Ability item bank (Hahn et al., 2014). In this study's Social Function: Ability-SF10 CFA analysis, factor loadings ranged from 0.91 to 0.95; overall model fit statistics suggested acceptable-to-good fit (CFI = 0.99, NNFI = 0.99, RMSEA = 0.097, SRMR = 0.025); and no residual correlations met or exceeded the 0.20 criterion for local dependence. Excluding extreme-score respondents from the Social Function: Ability-SF10 CFA analysis, factor loadings ranged from 0.83 to 0.92; overall model fit statistics continued to suggest acceptable-to-good fit (CFI = 0.98, NNFI = 0.98, RMSEA = 0.119, SRMR = 0.045); and, again, no residual correlations met or exceeded the 0.20 criterion for local dependence.

In the validity analyses, Pearson correlations between Social Function: Ability-SF10 and a set of six related PROMIS measures ranged from -0.784 to 0.765, with the following individual correlations providing specific evidence of Social Function: Ability-SF10's criterion-related validity: physical function (r = 0.765), sleep disturbance (r = −0.495), emotional distress-anxiety ($r = -0.614$), emotional distress-depression ($r = -0.635$), fatigue ($r = -0.784$), and pain interference ($r = -0.679$). Excluding extreme-score respondents from the Social Function: Ability-SF10 validity analysis, Pearson correlations between Social Function: Ability-SF10 and the set of six PROMIS measures ranged from −0.705 to 0.689, with the following individual correlations continuing to provide specific evidence of Social Function: Ability-SF10's criterion-related validity: physical function ($r = 0.689$), sleep disturbance ($r = -0.400$), emotional distress-anxiety ($r = -0.504$), emotional distress-depression ($r = -0.534$), fatigue ($r = -0.705$), and pain interference ($r = -0.595$).

DIF hypotheses

Hypotheses proposed by the expert panel are briefly summarized in Table 3. Gender-DIF hypotheses were that women (for reasons unrelated to social function) will tend to report more trouble doing all the family activities, doing all the activities with friends, and keeping up with family responsibilities; and will tend to report greater limitations doing fun things, doing social activities outside the home, and doing work, including work at home. Directional age-DIF was hypothesized for all items except for one (limit social activities at home), suggesting that older individuals will be more likely to report more trouble or limitation than younger individuals. Race/ethnicity-DIF was posited for four items suggesting that at the same level of social function, Asians and Hispanics would be more likely than other groups to report trouble keeping up with family responsibilities, Asians would be more likely to report greater limitation with doing fun things with others, and Hispanics would be more likely to report greater limitation with social activities as well as more trouble doing all of the family activities. Language-DIF hypotheses were not posited for any of the items. Education-DIF was posited for one item suggesting that individuals with higher levels of education will be likely to report more trouble doing activities with friends than those with lower levels of education.

Psychometric analyses: DIF detection and impact

DIF was evaluated for all five factors reviewed by the expert panel and for 13 additional factors. Using the study's standard DIF criterion (a McFadden pseudo-R^2 change of 0.01 or greater), none of the Social Function: Ability-SF10 items were flagged for DIF in any of the 18 sample characteristic comparisons (see Table 3). Using the study's sensitivity DIF criterion (a McFadden pseudo-R^2 change of 0.005 or greater), only one item ("I have trouble keeping up with my family responsibilities") was flagged for DIF in only one of the 18 sample characteristic comparisons (Stage 1 by cancer type: breast [$n = 712$] vs. prostate [$n = 273$] vs. uterus [$n = 290$]).

DIF impact analyses involving this one flagged item indicated a trivial impact. The Pear-son correlation of initial vs. purified theta scores was $r = 0.99$; 0.08 % ($n = 1$) of individual difference scores (initial theta minus purified theta; mean = −0.03, SD = .06) exceeded initial theta's median SE of 0.173; 0 % ($n = 0$) of individual difference scores exceeded initial individual theta score SEs; and Cohen's d effect sizes (initial theta scores by Stage 1 cancer type vs. purified theta scores by Stage 1 cancer type) differed minimally: Stage 1 breast vs. prostate (0.39 vs. 0.39); Stage 1 breast vs. uterus (0.10 vs. 0.12); Stage 1 prostate vs. uterus (0.28 vs. 0.26). No other DIF impact analyses were conducted because no other items were flagged for DIF, either in standard or sensitivity DIF detection analyses.

Discussion

To our knowledge, this was the first study to measure social function in people with cancer across three languages: English, Spanish, and Chinese. Comparisons of level of social function and of item response characteristics were conducted on the PROMIS Social Function 10-item Ability short form. Over 5,000 people with diverse cancer diagnoses participated in the study; the majority (91 %) completed the questionnaires in English. Many respondents (20 %) reported no limitations in social function and a few (2 %) reported extreme limitations. The social function short form exhibited excellent internal consistency reliability and essential unidimensionality, with and without the extreme-score respondents, providing evidence that the scale's use in this cancer patient sample was of sufficient reliability to allow appropriate group and individual comparisons based on scale performance differences. As with any assessment of self-reported health, measurement of individual-level change should be performed with careful attention paid to the accumulation of error over time (Donaldson, 2008; McHorney & Tarlov, 1995; Ware, Brook, Davies, & Lohr, 1981). Criterion-related validity was also supported.

Although an expert panel proposed that many of the 10 items might exhibit measurement bias, or DIF, based on gender, age, race/ethnicity, and/or education, no DIF was detected using state-of-the-science methods. Across 18 different sample characteristic groupings, no items were flagged for DIF using the study's prespecified DIF criterion, and only one item in one sample characteristic comparison was flagged for DIF using a sensitivity DIF criterion. This item's flagged DIF had only a trivial impact on estimation of scores.

Having culturally relevant, linguistically equivalent and psychometrically sound patient-reported measures in multiple languages helps to overcome some common barriers to including underrepresented groups in research and to conducting cross-cultural research (Stewart & Napoles-Springer, 2000). This will permit better examination of cultural differences in patient-reported outcomes and health disparities among vulnerable populations. In particular, there is a need for standardized measures of social health and participation that are applicable to a broad range of conditions and clinical settings (Whiteneck, 2010). The use of common indicators of social health will facilitate measurement consistency and comparison across studies and populations, and should enhance understanding of how these variables relate to other aspects of health.

Social function measures are especially important in cancer since the disease and its treatment can affect the quality of marital relationships, parental responsibilities, work abilities, and social activities (Bouknight et al., 2006; Fantoni et al., 2010; McDowell, 2006; Munir et al., 2009; Taskila et al., 2011). Optimal care for people with cancer thus includes obtaining a complete picture of their physical and psychosocial health status (Alfano & Rowland, 2006; Aziz, 2007a, 2007b; Bloom, Petersen, & Kang, 2007; Gotay & Muraoka, 1998; Hewitt, Greenfield, Stovall, Institute of Medicine [U.S.], & American Society of Clinical Oncology, 2006; Moinpour, Donaldson, & Redman, 2007). Although symptom status is very important, it is also important to capture the "reach" of symptoms and toxicity effects on day-to-day functioning (Jensen, Moinpour, & Fairclough, 2012). The diversity of social function issues during and after cancer treatment made this study's participants an excellent sample for evaluating the PROMIS Social Function: Ability measure. The results from this study are consistent with prior work in non-cancer populations (Hahn et al., 2014), and provide strong evidence that little to no DIF might be present among populations with other chronic conditions.

There are some limitations to this study. The sample size for participants who completed the survey in Chinese was too small to permit language DIF analysis. Although the terms measurement equivalence, differential item functioning (DIF), and bias are used interchangeably in this article, it should be noted that they have slightly different meanings. Typically, the term bias is reserved for findings of differential item functioning that have been both hypothesized to show DIF and for which there is other evidence in the literature lending confirmation to the findings. Given that a large proportion of participants (20 %) reported no limitations in social function, it would be useful to conduct additional studies with people with more limitations.

The Medical Outcomes Trust outlined eight recommended attributes for multi-item measures of latent traits: 1) a conceptual and measurement model, 2) reliability, 3) validity, 4) responsiveness, 5) interpretability, 6) low respondent and administrative burden, 7) alternative forms, and 8) cultural and language adaptations (Scientific Advisory Committee of the Medical Outcomes Trust et al., 2002). The results from this study, combined with the results from previous studies (Hahn, Cella, et al., 2010; Hahn et al., 2014), add to the accumulating evidence regarding the measurement properties of the PROMIS Social Function: Ability item bank and short forms.

Acknowledgments

This study was supported by grant #U01-AR057971-S1 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases, and grant #P30-CA051008-S1 from the National Cancer Institute. The authors thank all of the patients who participated in this study.

References

Alfano CM, Rowland JH. Recovery issues in cancer survivorship: A new challenge for supportive care. Cancer Journal. 2006; 12(5):432–443.

American Cancer Society. Cancer facts & figures 2015. Atlanta: American Cancer Society; 2015.

Aziz NM. Cancer survivorship research: State of knowledge, challenges and opportunities. Acta Oncologica. 2007a; 46(4):417–432. DOI: 10.1080/02841860701367878 [PubMed: 17497308]

Aziz NM. Late effects of cancer treatment. In: Chang AE, Ganz PA, Hayes DF, Kinsella T, Pass HI, Schiller JH, Stone RM, Strecher V, editorsOncology: An evidence-based approach. New York, NY: Springer Science & Business Media; 2007b. 1768–1790.

Bloom JR, Petersen DM, Kang SH. Multi-dimensional quality of life among long-term (5+ years) adult cancer survivors. Psycho-Oncology. 2007; 16(8):691–706. DOI: 10.1002/pon.1208 [PubMed: 17628036]

Bouknight RR, Bradley CJ, Luo Z. Correlates of return to work for breast cancer survivors. Journal of Clinical Oncology. 2006; 24(3):345–353. DOI: 10.1200/JCO.2004.00.4929 [PubMed: 16421415]

Cai L, Thissen D, du Toit S. IRTPRO 2.01. Lincolnwood, IL: Scientific Software International; 2011.

Camilli G, Shepard LA. Methods for identifying biased test items. London, England: Sage Publications; 1994.

Castel LD, Williams KA, Bosworth HB, Eisen SV, Hahn EA, Irwin DE, … DeVellis RF. Content validity in the PROMIS social-health domain: A qualitative analysis of focus-group data. Quality of Life Research. 2008; 17(5):737–749. DOI: 10.1007/s11136-008-9352-3 [PubMed: 18478368]

Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. Quality of Life Research. 2007; 16(Suppl 1): 133–141. DOI: 10.1007/s11136-007-9204-6 [PubMed: 17401637]

Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, … Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of a NIH roadmap cooperative group during its first two years. Medical Care. 2007; 45(5 Suppl 1):S3–S11. DOI: 10.1097/01.mlr.0000258615.42478.55

Choi SW, Gibbons LE, Crane PK. 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. 2011; 39(8):1–30. DOI: 10.18637/jss.v039.i08

Choi SW, Gibbons LE, Crane PK. lordif: Logistic regression differential item functioning using IRT. 0.2–2. 2012. Retrieved from http://cran.r-project.org/web/packages/lordif/lordif.pdf

Cook KF, Bombardier CH, Bamer AM, Choi SW, Kroenke K, Fann JR. Do somatic and cognitive symptoms of traumatic brain injury confound depression screening? Archives of Physical Medicine and Rehabilitation. 2011; 92(5):818–823. DOI: 10.1016/j.apmr.2010.12.008 [PubMed: 21530731]

DeWalt DA, Rothrock N, Yount S, Stone AA. PROMIS Cooperative Group. Evaluation of item candidates: The PROMIS qualitative item review. Medical Care. 2007; 45(5 Suppl 1):S12–S21. DOI: 10.1097/01.mlr.0000254567.79743.e2 [PubMed: 17443114]

Donaldson G. Patient-reported outcomes and the mandate of measurement. Quality of Life Research. 2008; 17(10):1303–1313. DOI: 10.1007/s11136-008-9408-4 [PubMed: 18953670]

Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Evaluation and the Health Professions. 2005; 28(2):212–232. DOI: 10.1177/0163278705275342 [PubMed: 15851774]

Fantoni SQ, Peugniez C, Duhamel A, Skrzypczak J, Frimat P, Leroyer A. Factors related to return to work by women with breast cancer in northern France. Journal of Occupational Rehabilitation. 2010; 20(1):49–58. DOI: 10.1007/s10926-009-9215-y [PubMed: 19902340]

Gotay CC, Muraoka MY. Quality of life in long-term survivors of adult-onset cancers. Journal of the National Cancer Institute. 1998; 90(9):656–667. DOI: 10.1093/jnci/90.9.656 [PubMed: 9586662]

Hahn EA, Cella D, Bode RK, Hanrahan RT. Measuring social well-being in people with chronic illness. Social Indicators Research. 2010; 96(3):381–401. DOI: 10.1007/s11205-009-9484-z

Hahn EA, Devellis RF, Bode RK, Garcia SF, Castel LD, Eisen SV, … Cella D. Measuring social health in the patient-reported outcomes measurement information system (PROMIS): Item bank development and testing. Quality of Life Research. 2010; 19(7):1035–1044. DOI: 10.1007/s11136-010-9654-0 [PubMed: 20419503]

Hahn EA, DeWalt DA, Bode RK, Garcia SF, DeVellis RF, Correia H, Cella D. New English and Spanish social health measures will facilitate evaluating health determinants. Health Psychology. 2014; 33(5):490–499. DOI: 10.1037/hea0000055 [PubMed: 24447188]

Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: SAGE Publications, Inc; 1991.

Healthy People 2020. Social Determinants of Health. 2015. Retrieved from http://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-health

Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: The universalist approach. Quality of Life Research. 1998; 7(4):323–335. DOI: 10.1023/A:1024985930536 [PubMed: 9610216]

Hewitt M, Greenfield S, Stovall E. Institute of Medicine (U.S.), & American Society of Clinical Oncology. From cancer patient to cancer survivor: Lost in transition. Washington, DC: National Academies Press; 2006.

Holland PW, Wainer H. Differential item functioning. Hillsdale, NJ: Lawrence Earlbaum Associates; 1993.

IBM Corporation. IBM SPSS statistics for Windows, version 22.0. Armonk, NY: IBM Corporation; 2013.

Institute of Medicine. The future of the public's health in the 21st century. Washington, DC: National Academies Press; 2003.

Institute of Medicine. Committee on Public Health Strategies to Improve Health. For the public's health: The role of measurement in action and accountability. Washington, DC: National Academies Press; 2011.

Jensen RE, Moinpour CM, Fairclough DL. Assessing health-related quality of life in cancer trials. Clinical Investigation. 2012; 2(6):563–577. DOI: 10.4155/cli.12.48

Jensen RE, Moinpour CM, Keegan THM, Cress RD, Wu XC, Paddock LA, … Potosky AL. The measuring your health study of cancer survivors. Psychological Test and Assessment Modeling. 2016; 58:99–117.

Joreskog K, Sorbom D. LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc; 2006.

McDowell I. Measuring health: A guide to rating scales and questionnaires. 3. Oxford, New York: Oxford University Press; 2006.

McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editorFrontiers in Econometrics. New York: Academic Press; 1974. 105–142.

McHorney CA, Cook KF. Invited Paper B. The ten Ds of health outcomes measurement for the twenty-first century. In: Lipscomb J, Gotay CC, Snyder C, editorsOutcomes assessment in cancer: Measures, methods, and applications. Cambridge, UK: Cambridge University Press; 2005. 590–609.

McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? Quality of Life Research. 1995; 4(4):293–307. DOI: 10.1007/BF01593882 [PubMed: 7550178]

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58(4):525–543. DOI: 10.1007/BF02294825

Meredith W, Teresi JA. An essay on measurement and factorial invariance. Medical Care. 2006; 44(11 Suppl 3):S69–S77. DOI: 10.1097/01.mlr.0000245438.73837.89 [PubMed: 17060838]

Moinpour C, Provenzale D. Outcomes assessment in cancer: Measures, methods, and applications. In: Lipscomb J, Gotay CC, Snyder C, editorsTreatment for colorectal cancer: Impact on health-related quality of life. Cambridge, UK: Cambridge University Press; 2005. 178–200.

Moinpour CM, Donaldson GW, Redman MW. Do general dimensions of quality of life add clinical value to symptom data? JNCI Monographs. 2007; 2007(37):31–38. DOI: 10.1093/jncimonographs/lgm007

Munir F, Yarker J, McDermott H. Employment and the common cancers: Correlates of work ability during or following cancer treatment. Occupational Medicine. 2009; 59(6):381–389. DOI: 10.1093/occmed/kqp088 [PubMed: 19692524]

Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill, Inc; 1994.

Patrick DL, Ferketich SL, Frame PS, Harris JJ, Hendricks CB, Levin B, … Vernon SW. National Institutes of Health State-of-the-Science Conference Statement: Symptom management in cancer: Pain, depression, and fatigue, July 15–17, 2002. Journal of the National Cancer Institute Monographs. 2004; (32):9–16. DOI: 10.1093/jnci/djg014 [PubMed: 15263035]

Paz SH, Spritzer KL, Morales LS, Hays RD. Evaluation of the patient-reported outcomes information system PROMIS(R) Spanish-language physical functioning items. Quality of Life Research. 2013; 22(7):1819–1830. DOI: 10.1007/s11136-012-0292-6 [PubMed: 23124505]

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, … Cella D. 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 Suppl 1):S22–S31. DOI: 10.1097/01.mlr.0000250483.85507.04 [PubMed: 17443115]

Regnault A, Herdman M. Using quantitative methods within the Universalist model framework to explore the cross-cultural equivalence of patient-reported outcome instruments. Quality of Life Research. 2015; 24(1):115–124. DOI: 10.1007/s11136-014-0722-8 [PubMed: 24894383]

Samejima F. Estimation of latent ability using a response pattern of graded scores; Psychometrika Monograph Supplement. 1969. 17Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Lohr KN, Aaronson N, Alonso J, Burnam A, Patrick DL, … Stein RE. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research. 2002; 11(3):193–205. DOI: 10.1023/A:1015291021312 [PubMed: 12074258]

Sprangers MA, Taal BG, Aaronson NK, te Velde A. Quality of life in colorectal cancer. Stoma vs. nonstoma patients. Diseases of the Colon and Rectum. 1995; 38(4):361–369. DOI: 10.1007/BF02054222 [PubMed: 7720441]

Stewart AL, Napoles-Springer A. Health-related quality-of-life assessments in diverse population groups in the United States. Medical Care. 2000; 38(9 Supplement II):102–124. DOI: 10.1097/00005650-200009002-00017

Taskforce on Health Status. Criteria for and selection of domains for the measurement of health status. Paper presented at the Conference of European Statisticians; Budapest, Hungary. 2005.

Taskila T, De Boer A, Van Dijk FJH, Verbeek J. Fatigue and its correlates in cancer patients who had returned to work – a cohort study. Psycho-Oncology. 2011; 20(11):1236–1241. DOI: 10.1002/pon.1843 [PubMed: 20821375]

Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Medical Care. 2006; 44(11 Suppl 3):S152–S170. DOI: 10.1097/01.mlr.0000245142.74628.ab [PubMed: 17060822]

Teresi JA, Stewart AL, Morales LS, Stahl SM. Measurement in a multi-ethnic society: Overview to the special issue. Medical Care. 2006; 44(11 Suppl 3):S3–S4. DOI: 10.1097/01.mlr.0000245437.46695.4a [PubMed: 17060831]

Thissen D. Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory. Lincolnwood, IL: Scientific Software International, Inc; 1991.

van de Vijver F, Kwok L. Methods and data analysis for cross-cultural research. Thousand Oaks: Sage Publications; 1997.

van der Linden WJ, Hambleton RK. Handbook of modern item response theory. New York: Springer-Verlag; 1997.

Ware JE, Brook RH, Davies AR, Lohr KN. Choosing measures of health status for individuals in general populations. American Journal of Public Health. 1981; 71(6):620–625. DOI: 10.2105/AJPH.71.6.620 [PubMed: 7235100]

Whitehead M. Tackling inequalities: A review of policy initiatives. In: Benzeval M, Judge K, Whitehead M, editorsTackling inequalities in health: An agenda for action. London: King's Fund; 1995. 22–52.

Whiteneck GG. Issues affecting the selection of participation measurement in outcomes research and clinical trials. Archives of Physical Medicine and Rehabilitation. 2010; 91(9):S54–S59. DOI: 10.1016/j.apmr.2009.08.154 [PubMed: 20801281]

Wild D, Grove A, Martin M, Eremenco S, Ford S, Verjee-Lorenz A, Erickson P. Principles of good practice for the translation and cultural adaptation process for patient reported outcomes (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in Health. 2005; 8(2):94–104. [PubMed: 15804318]

World Health Organization. Constitution of the World Health Organization. Geneva: World Health Organization; 1946.

Zumbo BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for Binary and Likert-Type (Ordinal) Item Scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.

Table 1

PROMIS Social Function: Ability to Participate in Social Roles and Activities (10-item Short Form, English Version)[a]

	Ceiling effect (% "Never")	**Floor effect (% "Always")**
I have to limit the things I do for fun with others	32.1	7.3
I have trouble doing all of the activities with friends that I want to do	31.8	7.8
I have to limit social activities outside of my home	34.6	8.3
I am limited in doing my work (include work at home)	34.2	8.6
I have trouble keeping up with my work responsibilities (include work at home)	33.8	8.3
I have trouble doing all of the family activities that I want to do	33.4	7.6
I have trouble doing all of the activities with friends that are really important to me	35.9	7.4
I have to limit social activities at home	39.5	6.5
I have to limit my regular family activities	40.8	5.6
I have trouble keeping up with my family responsibilities	41.6	5.5

[a]This custom 10-item short form was created for this project, prior to the creation of the current PROMIS 4-, 6- and 8-item short forms (Hahn et al., 2014).

Items are listed in order of administration.

Response scale for all items: *never* = 5, *rarely* = 4, *sometimes* = 3, *often* = 2, *always* = 1 (*often* was changed to *usually* in a later version of the items; Hahn et al., 2014)

Table 2

Sociodemographic and Clinical Characteristics of Study Participants (*n* = 5,301)

Gender	
Female	3,134 (59.1 %)
Male	2,133 (40.2 %)
Missing	34 (0.6 %)
Age at Cancer Diagnosis	
21–49	1,177 (22.2 %)
50–64	1,947 (36.7 %)
65–84	2,143 (40.4 %)
Missing	34 (0.6 %)
Ethnicity, Race	
Non-Hispanic White	2,203 (41.6 %)
Non-Hispanic Black	1,081 (20.4 %)
Non-Hispanic Asian/Pacific Islander	879 (16.6 %)
Hispanic, any race	1,006 (19.0 %)
Other	128 (2.4 %)
Missing	4 (0.1 %)
Survey Language	
Chinese	136 (2.6 %)
English	4,843 (91.4 %)
Spanish	322 (6.1 %)
Highest Education	
< High School	923 (17.4 %)
High School Diploma or GED	1,012 (19.1 %)
Some college	1,714 (32.3 %)
College degree	957 (18.1 %)
Advanced degree	627 (11.8 %)
Missing	68 (1.3 %)
Cancer Diagnosis	
Breast	1,586 (29.9 %)
Prostate	1,126 (21.2 %)
Colorectal	896 (16.9 %)
Lung	684 (12.9 %)
Non-Hodgkin’s Lymphoma	445 (8.4 %)
Uterine	382 (7.2 %)
Cervical	148 (2.8 %)
Missing	34 (0.6 %)
Help Answering Survey Questions	
I answered all of the questions with no help	4,421 (83.4 %)

My parent, guardian, spouse, child or significant other helped me with some or all of the questions	744 (14.0 %)
Missing	136 (2.6 %)

Table 3

Planned Comparisons and Hypothesized Differential Item Functioning (DIF)

Factor	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4	Subgroup 5	Subgroup 6	DIF Hypothesis[a]	# DIF Items (standard criterion)[b]	# DIF Items (sensitivity criterion)[c]
Gender	Male (n = 2,133)	Female (n = 3,134)					Women will report more limitations on most items	0 items	0 items
Age at diagnosis	21 to <65 (n = 3,124)	65 or older (n = 2,143)					Older people will report more limitations on most items	0 items	0 items
Ethnicity, Race	Non- Hispanic White (n = 2,203)	Non- Hispanic Black (n = 1,081)	Hispanic (n = 1,006)	Non-Hispanic Chinese (n = 314)	Non-Hispanic Filipino (n = 248)		Chinese and/or Hispanics will report more limitations on 4 items	0 items	0 items
Language	English (n = 4,843)	Spanish (n = 322)					No items will have DIF	0 items	0 items
Born in US	Yes (n = 3,736)	No (n = 1,518)						0 items	0 items
Highest Education	H.S. or less (n = 1,935)	Some college or more (n = 3,298)					Higher educated will report more limitations on 1 item	0 items	0 items
Income 1	<$60,000 ($n$ = 2,650)	$60,000+ ($n$ = 1,744)						0 items	0 items
Income 2	<$20,000 ($n$ = 1,184)	$20,000+ ($n$ = 3,210)						0 items	0 items
Comorbidity	0 or 1 (n = 2,497)	2 or more (n = 2,804)						0 items	0 items
Cancer	Breast (n = 1,586)	Colorectal (n = 896)	Lung (n = 684)	NHL (n = 445)	Prostate (n = 1,126)	Uterus (n = 382)		0 items	0 items
Breast cancer by stage	Stage 1 (n = 712)	Stage 2 (n = 572)						0 items	0 items
Colorectal cancer by stage	Stage 2 (n = 225)	Stage 3 (n = 278)						0 items	0 items
Prostate cancer by stage	Stage 1 (n = 273)	Stage 2 (n = 617)						0 items	0 items
Stage 1 by cancer	Breast (n = 712)	Prostate (n = 273)	Uterus (n = 290)					0 items	1 item[d]
Stage 2 by cancer	Breast (n = 572)	Colorectal (n = 225)	Prostate (n = 617)					0 items	0 items
Depression[e]	Yes (n = 1,020)	No (n = 4,080)						0 items	0 items

Factor	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4	Subgroup 5	Subgroup 6	DIF Hypothesis[a]	# DIF Items (standard criterion)[b]	# DIF Items (sensitivity criterion)[c]
Anxiety[f]	Yes (n = 1,013)	No (n = 4,105)						0 items	0 items
Help answering questions	No (n = 4,421)	Some or all (n = 744)						0 items	0 items

[a] See text for description of how hypotheses were generated and which items were hypothesized to have DIF.

[b] McFadden pseudo-R^2 change of 0.01 or greater

[c] McFadden pseudo-R^2 change of 0.005 or greater

[d] "I have trouble keeping up with my family responsibilities" was flagged for DIF in only one of the 18 sample characteristic comparisons (Stage 1 by cancer type: breast vs. prostate vs. uterus)

[e] "Has a doctor ever told you that you had depression?"

[f] "Has a doctor ever told you that you had anxiety?"