Psychological Test and Assessment Modeling, Volume 58, 2016 (2), 371-402

Differential item functioning in Patient Reported Outcomes Measurement Information System® (PROMIS®) Physical Functioning short forms: Analyses across ethnically diverse groups

Richard N. Jones[1], *Doug Tommet*[2], *Mildred Ramirez*[3,4],
Roxanne Jensen[5,6] *& Jeanne A. Teresi*[3,4,7]

Abstract

We analyzed physical functioning short form items derived from the PROMIS® item bank (PF16) using data from more than 5,000 recently diagnosed, ethnically diverse cancer patients. Our goal was to determine if the short form items demonstrated evidence of differential item functioning (DIF) according to sociodemographic characteristics in this clinical sample. We evaluated responses for evidence of unidimensionality, local independence (given a single common factor), differential item functioning, and DIF impact. DIF was evaluated attributable to sex, age (middle aged vs. younger and older), race/ethnicity (White vs. Black or African-American, Asian/Pacific Islander, Hispanic) and level of education. We used a multiple group confirmatory factor analysis with covariates approach, a multiple indicators multiple causes (MIMIC) model. We confirmed essential unidimensionality but some evidence for multidimensionality is present, particularly for basic activities of daily living items, and many instances of local dependence. The presence of local dependence calls for further review of the meaning and measurement of the physical functioning domain among cancer patients.

[1] *Correspondence concerning this article should be addressed to:* Richard N. Jones, Sc.D., Department of Psychiatry and Human Behavior, Department of Neurology, Warren Alpert Medical School, Brown University, Butler Hospital, 345 Blackstone Boulevard, Box G-BH, Providence, Rhode Island 02906, USA; email: richard_jones@brown.edu

[2] Department of Psychiatry and Human Behavior, Brown University

[3] Research Division, Hebrew Home at Riverdale, RiverSpring Health

[4] Weill Cornell Medical Center, Department of Geriatrics and Palliative Medicine

[5] Department of Oncology, Georgetown University

[6] Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center

[7] Columbia University Stroud Center and New York State Psychiatric Institute

Nearly every item demonstrated statistically significant DIF. In all group comparisons the impact of DIF was negligible. However, the Hispanic subgroup comparison revealed an impact estimate just below an arbitrary threshold for small impact. Within the limitations of local dependency violations, we conclude that items from a static short form derived from the PROMIS physical functioning item bank displayed trivial and ignorable DIF attributable to sex, race, ethnicity, age, and education among cancer patients.

Key words: PROMIS, physical function, differential item functioning, cancer, measurement equivalence, ethnicity

Introduction

Physical functioning is a dimension of health and quality of life that is of singular importance in policy research, outcomes evaluation, and clinical practice. Physical functioning is useful as a predictor of health care services utilization, eligibility determination for services and support, and as an outcome variable in population estimates of disease burden, active life, disability, evaluating clinical trials, and monitoring the progress of chronic disease (Feinstein, Josephy, & Wells, 1986; Fried, Ettinger, Lind, Newman, & Gardin, 1994).

The Patient Reported Outcomes Measurement Information System® (PROMIS®; Cella et al., 2007) is a set of measures and an infrastructure for collecting data from patients on their physical, mental, and social health. Within the physical health domain, physical functioning is one of the domains assessed by PROMIS. PROMIS is the result of a watershed project funded as one of the NIH Roadmap initiatives (http://nihroadmap.nih.gov) to re-engineer the research enterprise. The goal of this project was to provide a method to assess patient reported outcomes (PROs) suitable for use in randomized controlled trials enrolling patients with chronic disease. The intent was to develop measures that would capture how diseases and treatments affect the subjective experience of symptoms and their severity and frequency in the areas of physical, emotional, and social well-being. A major objective was to address the lack of standardization produced by the multitude of measures used by various researchers and facilitate the establishment of a common metric for PROs (NIH, 2003). PROMIS has been enormously successful in achieving its aims. The use of PROMIS measures as outcomes in observational studies and clinical trials will accelerate as computerized adaptive testing (CAT) modules are available in electronic data capture systems (Rothrock, 2014) and tools such as REDCap (Harris et al., 2009; Rothrock, 2014). Additionally, efforts are underway to bring PROMIS measures into clinical practice (Jensen et al., 2015; Wagner et al., 2014) and quality monitoring (Glasgow & Riley, 2013). Finally, establishing the validity of a patient reported outcome measure is a critical step in the development of tools to be used in regulated drug and device development (FDA, 2009).

Validation studies of the PROMIS Physical Function (PF) item bank have been described, and while the results are promising the clinical samples were small, well-educated, and predominantly White, or based on internet panels (Bruce et al., 2009;

Rose, Bjorner, Becker, Fries, & Ware, 2008). Therefore, an unmet need is evaluation of PROMIS physical function items in a diverse clinical population in order to detect possible measurement bias (the items do not measure the same thing in the same way across groups). Ruling out possible measurement bias is an important step in construct validation (Reise & Waller, 2009). The primary goal of these analyses was the use of a latent variable measurement model to estimate parameters that may express differential item functioning (DIF) or item bias attributable to patient sex, age, race, and education in 16 PROMIS physical functioning short form items administered to a cohort of ethnically diverse cancer patients. The secondary goal was to illustrate methodological challenges encountered in the process of evaluating measurement equivalence.

Methods

Participants and source of data

Participants in this study were recruited through the Measuring Your Health (MY-Health) Study. Eligible cancer patients were females with breast, uterine, or cervical cancer, males with prostate cancer, or males or females with colorectal cancer, lung cancer, or non-Hodgkin lymphoma. Patients were recruited from three (California, Louisiana, New Jersey) states covered by four Surveillance, Epidemiology, and End Results Program (SEER) cancer registries. Additional eligibility criteria included self-nomination (or registry-supplied) racial designation as Hispanic, White (non-Hispanic), Black, or Asian, aged between 21 and 84 years at the time of diagnosis, and residence within a conscripted study area within the target state. Finally, participants had to be within 6 - 13 months of their first cancer diagnosis and able to read or speak English, Spanish, or Mandarin. For this analysis, we used a listwise complete sample on the basis of socio-demographics, and from among the remainder sample, participants had to have data on at least two of the 16 physical functioning items to be included. The available MY-Health data set contains data for 5,507 persons. We excluded 36 people (0.7 %) with missing data on sex, and from the remainder we excluded 136 (2.5 %) records belonging to persons who were missing data on race/ethnicity (n = 4) or did not self-identify as White, Black or African-American, Asian, or Hispanic (n = 132). From this remainder we excluded 17 (0.3 %) who did not provide at least two non-missing physical functioning item responses. The final analytic sample size was 5,318 (97 % of 5,507).

Measures

In this analysis we consider responses to a 16 item short form of the PROMIS Physical Functioning domain item bank (Fries et al., 2014). The PROMIS physical functioning domain is broadly defined and includes assessments of function in a range of impairment and ability that covers disability through physical fitness, as well as upper and lower extremity functioning, and fine motor coordination. The 16-item short form used in this study included items from PROMIS Physical Function short forms 4a, 6a, 6b, 10a, and seven of eight items from short form 8a (heavy work around the house was excluded). Additionally, 11 items from version 20a were included in the data set. Items were selected according to their frequency of administration in the online PROMIS computerized adaptive testing assessment among persons scoring at two levels indicating impairment, -0.5 and -1.0 standard deviations from the mean PF score.

Hypotheses generation

DIF hypotheses were generated by asking a set of eight content experts to indicate whether or not they expected DIF to be present, and the direction of the 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 hypotheses 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 difficulty (e.g., in ability to do chores) should depend only on the level of the trait (state), e.g., physical function 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., mild) level of physical function should have the same likelihood of reporting difficulty related to chores. If it is theorized that this might not be the case, it would be hypothesized that the item has gender DIF.

The experts were asked to rate individually each of the 16 items with respect to gender, age, race/ethnicity, language, education, and diagnosis. They provided hypotheses in terms of presence and direction of DIF. The goal was to identify items that might have a different meaning or not be understood well and/or equivalently by individuals of any of the groups referenced. The number of directional hypotheses across raters for each demographic subgroup was tallied, entered in the summary table, and reviewed by an expert in qualitative methods, conversant in generation of DIF hypotheses. A second reviewer participated in an adjudication process. The hypothesis was declared if two or more raters posited a directional hypothesis.

Analytic approach

Our approach to analysis involved checking model assumptions, followed by examination of DIF.

Method to test assumptions

Many different statistical models fall under the rubric of item response theory (IRT). Our implementation of the MIMIC model is consistent with a unidimensional ordinal response variable model (Samejima, 1969) and uses a multivariate probit regression framework. Assumptions of this model include unidimensionality (that a single latent variable is sufficient to account for the correlation among the items in the test) and local independence (that pairs of items, after conditioning on a single latent variable, are correlated only trivially). We checked the assumption of unidimensionality using permuted parallel analysis (Buja & Eyuboglu, 1992). This method compares the observed eigenvalues (which we obtained from a matrix of polychoric correlation coefficients) to a quasi-permutation distribution of eigenvalues. The quasi-permutation distribution of eigenvalues is obtained by repeatedly randomly assigning item responses to participants and estimating eigenvalues from the random data. With a permutation distribution of eigenvalues, we can obtain *p*-values for each eigenvalue from the observed data, and use those to guide decisions about the significance of possible latent factors.

We evaluated the assumption of local independence by examining residual item polychoric correlations given a single factor model. We accomplished this by estimating a single factor model iteratively freeing those residual covariances, one at a time, that were identified as suggesting model misfit on the basis of modification indices. Modification indices are estimated from derivatives of the model fitting function, scaled approximately χ^2, and are estimated for model parameters that are not freely estimated (including parameters constrained to be 0 or constrained to be equal to some other parameter estimate; Raykov & Marcoulides, 2000).

Description of the method used for testing for the presence of DIF

Our approach to DIF detection is accomplished in a general latent variable modeling framework that is consistent with IRT, and presented in the overview to this series (Teresi & Jones, 2016). Model estimation and parameter estimates were obtained using Mplus software (version 7.11; Muthén & Muthén, Los Angeles CA). Briefly, we estimated multiple group confirmatory factor analysis models, where the grouping variable is the factor for which we would like to determine the presence of DIF, and the latent variable indicators are items from the assessment scale (Physical Function-16; PF-16). DIF may be present in the form of different measurement slopes or factor loadings (analogous to discrimination parameters in the IRT framework) and/or in the item locations or thresholds (analogous to item difficulty parameters in the IRT framework). Estimation can be achieved in Mplus using full-information maximum likelihood and a logistic regression model, or limited information weighted least squares and a probit regression model. We used a multivariate probit regression framework and the weighted least squares mean and variance adjusted estimator (WLSMV). One of the advantages of conducting DIF detection in the context of a general latent variable modeling framework is the ability to control for the effect of background variables or other covariates that are not of particular interest in terms of DIF detection but may be important for describing differences in the latent trait or otherwise act as potential confounders. For example, when examining DIF due to sex, it would be helpful to know that differences in the distribution of age across sex groups are not confounding the analysis. When a latent variable model includes multiple covariates and multiple indicators of a latent trait, we can describe the model as a MIMIC (multiple indicators, multiple causes) model (Jöreskog & Goldberger, 1975). We describe our approach as a multiple group (MG) MIMIC model. Examples and detailed descriptions of the approach can be found elsewhere (Jones, 2003, 2006a, 2006b).

The multiple group aspect of the model allowed for DIF to be tested for both the factor loading and item thresholds. As mentioned, the MIMIC model approach allows for additional covariates to be controlled beyond the covariate that was used to stratify the model into groups. For example, when testing for DIF due to gender, the model includes control for the effects of age and race in physical functioning. We limit our DIF detection to two groups at a time, a reference group and a focal group. If the factor under study for potential DIF has more than two levels (e.g., age groups, race/ethnicity, education) we compare different levels of that categorical variable to a common reference group. It should be noted that Mplus and other structural equation modeling (SEM) software packages are capable of estimating models in an indefinite number of groups simultaneously, and in sensitivity analyses we estimated models using a multiple groups approach with more than two groups (age, education, race models). Similar results were obtained.

Our DIF detection algorithm starts with a model assuming measurement invariance (i.e., no DIF). Ideally, anchor items are chosen a priori (Jones, 2006b), although we did not do so in this analysis. Following the framework articulated by Kopf, Zeileis, and Strobl (2015), our approach closely resembles the *all-other purified selection* stepwise approach to anchor item selection. DIF is identified in a forward stepwise fashion using scaled derivatives from the model fitting function known as modification indices (MIs). We used MIs in a very restricted way. Only MIs related to cross group equality constraints on latent response variable means or factor loadings were considered. (Latent response variables, y^*, are auxiliary or latent continuous response variables that when present at a level greater than some threshold t_c lead to observing category c or higher on the observed categorical variable (y): e.g., for a dichotomous item, $y = 0$ if $y^* \leq t_c$; 1 if $t_c < y^*$; see Agresti, 2002). Separately, the item parameter corresponding to the largest modification index for the factor loading and the item threshold were allowed to be freely estimated and then tested against the baseline model. The model with the largest test statistic for either the factor loading or response mean became the new baseline model. The process repeats by examining the new modification indices and continues until the freely estimated item parameter does not return a significant test statistic. A more detailed description of the anchor item selection procedure is presented in the methods overview paper in this series (Teresi & Jones, 2016).

Our IRT, confirmatory factor analytic, and eigenvalue models were estimated with Mplus software (version 7.11, Muthén & Muthén, Los Angeles CA), and governed with custom Stata (version 13.1, College Station, TX) modules to automate data preparation and resampling, evaluation of estimated results, and implementation of the forward stepwise model selection procedure (available at http://www.lvmworkshop.org/home/runmplus-stuff). Figures were generated with custom software using Stata and R (version 3.1.2, R Core Team, Vienna, Austria). All scripts and programs are available upon request.

Sample size, effect size, and impact

Sample size requirements for IRT work are not well established, although sample size recommendations and rules of thumb range from 100 to 500 persons per group (reviewed by Orlando-Edelen & Reeve, 2007); however, most two parameter IRT models or multiple group SEM models require 200 to 300 respondents per group. The implications for DIF detection in a general latent variable modeling framework using Mplus software are no different. As with any complex analysis, arguably the best practice to power estimation in the study design stage is to use methods of simulation using data relevant to the study at hand to evaluate the statistical power to detect DIF of a magnitude that either meets some standard effect size threshold or produces a minimally clinically meaningful impact. Simulation is important, because the required sample size to estimate factor loadings (or discrimination parameters), item means/thresholds or difficulty parameters, and group differences in these parameters will be influenced by the instrument (length, number of item response categories, fidelity to the unidimensionality assumption) and the sample (size, coverage of the latent trait, magnitude of group differences in the mean, and variance of the latent trait). However, simulation studies are technically demanding and conceptually challenging as the field is without consensus on definitions of clinically meaningful DIF impact or DIF magnitude (DeMars, 2011; Hidalgo & López-Pina, 2004; Orlando-Edelen, Stucky, & Chandra, 2015; Steinberg & Thissen, 2006; Teresi, Ramirez, Jones, Choi, & Crane, 2012). Power calculations with simulations performed on behalf of the original proposed analyses specified 300 per group as the number of subjects required.

We used tests of statistical significance without correction for multiple comparisons to determine the presence of DIF. We have found, in unpublished simulation studies, that the DIF procedures we describe here maintain a nominal type-I error level of 5 % without multiple comparisons correction. However, we also did not base an inference of DIF on statistical significance alone. Instead, we viewed significance as a first level screen for items with possible DIF, evaluated further with effect size statistics and measurement of DIF impact. We declared DIF to be of negligible impact if the mean difference in the latent trait across groups before and after adjusting for DIF was less than 0.2. Specifically, we compared the estimate of the difference in the mean latent trait in reference and focal groups without versus with DIF adjustment. If this difference in differences was greater than |0.2|, we inferred the overall DIF impact was more than trivial. We used the 0.2 threshold because, as the latent trait is assumed to follow a unit normal distribution in the reference population, a difference of 0.2 amounts to what is the threshold between trivial and small effects in Cohen's effect size taxonomy (Cohen, 1988). A large sample may permit the detection of DIF that while statistically significant due to extremely high power is of trivial impact. Thus, it is important to pair the statistical significance step with the impact assessment when drawing inferences regarding the relevance of detected DIF.

Handling of missing data

Mplus software handles missing data differently depending on the variable's role in the analysis and the method of parameter estimation. For independent variables (covariates), Mplus uses listwise complete samples. Therefore, when presented with data sets with missing data among the covariates, methods of multiple imputation were used to stochastically generate multiple complete copies of the data set. Mplus can use multiple data sets to estimate a model and produce a single and appropriately averaged result set. For missing data among the dependent variables, Mplus uses pairwise complete data to estimate mean and covariance matrices for parameter estimation when least squares parameter estimation is requested. When full information maximum likelihood model estimation is requested, Mplus will use all observations and produce maximum likelihood parameter estimates. The pairwise complete analysis is appropriate when the missing data mechanism can be considered missing completely at random, whereas the maximum likelihood approach is more appropriate when the missing data mechanism is more suitably assumed to be missing at random. Because we used the mean and variance adjusted weighted least squares modeling approach, our analyses invoked the assumption that missing data among the outcomes (PF items) are missing completely at random. Also, we used a listwise complete sample with regard to the covariates, which also invokes the assumption of missing completely at random.

Results

Qualitative: The physical function items were reviewed qualitatively by eight content experts regarding potential sources of differential item functioning. Two of the members of the panel were clinical or counseling psychologists, three were public health professionals. The remaining three were: a gerontologist, an epidemiologist, and a psychiatrist. The hypotheses are discussed below in relation to findings of DIF.

Sample Characteristics: Participant (n = 5,318) characteristics are reported in Table 1. The sample was predominantly female (60 %), and older (41 % were aged 65 to 84), but about a fifth of the sample was under 50 and included persons aged 21-84 at diagnosis. Most of the sample self-identified as minority group members; 43 % were White and not Hispanic, and about a fifth each of the remainder identified themselves as Black or African-American, Asian/Pacific Islander, or Hispanic.

Characteristics of the PROMIS short form are presented in Tables 2a and 2b. The PROMIS Physical Functioning item bank has two types of items. One type of item, which we call health limitation items (Table 2a), asks about whether physical health problems are causing difficulty with physical functioning tasks. For example: “Does your health now limit you in climbing one flight of stairs?” Another type of item, that we call ability items, does not include the health problem qualification (Table 2b). For example: “Are you able to go up and down stairs at a normal pace?” For both sets of items responses are measured on a 5 point scale (see Table 2a-b). The average item response across all 16 items was about 4, indicating that the cancer patients tended to have a little difficulty or were somewhat limited by their physical health in performing various physical activity tasks assessed. It is interesting to note that the MY-Health participants tended to rate the health limitation items (Table 2a) with greater impairment than the difficulty (ability) questions (Table 2b). For example, 63 % responded *not at all* or *very little* (the top two boxes) to the question “Does your health now limit you in climbing one flight of stairs?” (Table 2a), while 74 % responded *without any difficulty* or *with a little difficulty* (the two top categories) to “Are you able to go up and down stairs at a normal pace?” (Table 2b).

Table 1:
PROMIS MY-Health subsample (*n* = 5,318) participant characteristics

Characteristic	**n or Mean (*SD*)**	**(%)**
Total [*n* (%)]	5,318	(100)
Sex [*n* (%)]		
Men	2,152	(41)
Women	3,166	(60)
Age at cancer diagnosis (years) [*M* (*SD*)]	60 (13)	
Age group at cancer diagnosis [*n* (%)]		
21-49 years	1,160	(22)
50-64	1,967	(37)
65-84	2,191	(41)
Race/ethnicity [*n* (%)]		
White	2,267	(43)
Black or African-American	1,094	(21)
Hispanic	1,051	(20)
Asian/Pacific Islander	906	(17)
Educational attainment [*n* (%)]		
Less than high school	949	(18)
High school graduation	1,031	(19)
More than high school, without 4-year degree	1,705	(32)
College graduate (with 4 year degree)	955	(18)
More than college graduate	621	(12)
Not reported	57	(1)
Average of 16 PF items (scored 1 - 5, 1 = *cannot do* 5 = *not at all limited*) [*M* (*SD*)]	3.9 (1)	

Table 2a:
PROMIS MY-Health Physical Functioning 16 item (PF16) short form, item description and distribution: *Health limitation items* (items 1-8; PROMIS short form item numbers are in parentheses under the item label)

Item description		**Category distribution (%)**					
Label	**Question (order)**	**1**	**2**	**3**	**4**	**5**	**Mean**
Usual physical activities (6a,8a)	How much do physical health problems now limit your usual physical activities (such as walking or climbing stairs)? (6)	5	20	21	20	34	3.6
Moderate work (6a, 6b, 8a, 8b)	Does your health now limit you in doing moderate work around the house like vacuuming, sweeping floors or carrying in groceries? (7)	6	13	19	18	44	3.8
Climbing stairs (10a, 20a)	Does your health now limit you in climbing one flight of stairs? (3)	7	14	16	16	47	3.8
Carrying groceries (8a, 8b, 10a, 20a)	Does your health now limit you in lifting or carrying groceries? (4)	7	14	19	17	42	3.7
Bending, kneeling, stooping (10a, 20a)	Does your health now limit you in bending, kneeling, or stooping? (5)	7	17	21	18	37	3.6
Physical labor (6b, 8b, 20a)	Does your health now limit you in doing two hours of physical labor? (8)	17	16	19	17	30	3.3
Walking more than one mile 10a, 20a	Does your health now limit you in walking more than a mile? (2)	19	15	17	13	36	3.3
Vigorous activities (10a, 20a)	Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports? (1)	26	24	22	13	16	2.7

Note: Response category labels are *not at all* (5), *very little* (4) *somewhat* (3), *quite a lot* (2), and *cannot do* (1)

Table 2b:
PROMIS MY-Health Physical Functioning 16 item (PF16) short form, item description and distribution: *Ability items* (items 9-16)

Item description		**Category distribution (%)**					
Label	**Question (order)**	**1**	**2**	**3**	**4**	**5**	**Mean**
Wash, dry body (10a, 20a)	Are you able to wash and dry your body? (12)	1	2	4	9	83	4.7
Get on, off toilet (10a. 20a)	Are you able to get on and off the toilet? (13)	1	2	4	10	83	4.8
Dress yourself (10a, 20a)	Are you able to dress yourself, including tying shoelaces and doing buttons? (10)	1	3	7	13	76	4.6
Shampoo hair (10a, 20a)	Are you able to shampoo your hair? (11)	2	2	4	7	84	4.7
Run errands, shop (4a, 6a, 8a, 6b, 8b)	Are you able to run errands and shop? (16)	6	6	12	16	59	4.2
Walk 15 minutes (4a, 6a, 6b, 8a, 8b)	Are you able to go for a walk of at least 15 minutes? (15)	7	6	11	15	60	4.2
Go up, down stairs (4a, 6a, 6b, 8a, 8b)	Are you able to go up and down stairs at a normal pace? (14)	7	7	12	20	54	4.1
Chores (4a, 6a, 6b, 8a, 8b, 10a, 20a)	Are you able to do chores such as vacuuming or yard work? (9)	11	9	17	23	39	3.7

Note: Response category labels are *without any difficulty* (5), *with a little difficulty* (4) *with some difficulty* (3), *with much difficulty* (2), and *unable to do* (1)

Model assumptions

We evaluated the assumption of unidimensionality using permuted parallel analysis. Results of this analysis suggest that the set of items do conform to a unidimensional scale. Relative to the quasi-permutation distribution of eigenvalues, the first eigenvalue was significant (*p* <0.01) and the second and subsequent eigenvalues were not (*p* >0.90).

Despite this evidence of unidimensionality, upon estimating a single factor confirmatory factor analysis (CFA) model, we obtained sub-optimal model fit statistics (Table 3). The confirmatory fit index (CFI) for this model was 0.983 (values over 0.95 indicate good fit; Hu & Bentler, 1999), but the root mean squared error of approximation (RMSEA) was 0.113, where values less than 0.06 indicate good fit (Hu & Bentler). However, it is not uncommon to observe values of around 0.10 to 0.12 for PROMIS measures; and there are some recommendations to move away from reliance on fit statistics alone to determine model adequacy (see Cook, Kallen, & Amtmann, 2009).

The sub-optimal fit observed for a single factor model when the eigenvalues were consistent with a single factor model suggests the presence of multiple specific factors that leave large residual item correlations; in other words, violations of the assumption of local independence. We report selected residual correlations in Table 3. The correlations selected for this table were identified using an iterative backwards stepwise procedure

Table 3:
PROMIS MY-Health Physical Functioning 16 item (PF16) short form residual correlations given a single factor model

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Vigorous activities	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2. Walking more than one mile	.44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3. Climbing stairs	-	.46	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4. Carrying groceries	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5. Bending, kneeling, stooping	-	.14	.15	-	-	-	-	-	-	-	-	-	-	-	-	-
6. Usual physical activities	.13	.47	.45	-.19	.27	-	-	-	-	-	-	-	-	-	-	-
7. Moderate work	-	-	-.26	.12	-.31	-.21	-	-	-	-	-	-	-	-	-	-
8. Physical labor	.42	.40	.12	-	-	.21	.37	-	-	-	-	-	-	-	-	-
9. Chores	.20	.26	-	-	-	-	.43	.43	-	-	-	-	-	-	-	-
10. Dress yourself	-.15	-	-	-	-	-	-	-	.19	-	-	-	-	-	-	-
11. Shampoo hair	-.15	-	-	-	-	-	-	-	.25	.67	-	-	-	-	-	-
12. Wash, dry body	-.17	-	-	-	-	-	-	-	.25	.75	.84	-	-	-	-	-
13. Get on, off toilet	-.18	-	-	-.14	.12	-	-	-	.10	.61	.61	.71	-	-	-	-
14. Go up, down stairs	-.20	.14	.20	-.85	-.13	-	-.89	-	-	-	-	.13	.30	-	-	-
15. Walk 15 minutes	-.14	.48	-	-.79	-.30	-	-.75	-	-	-	-	-	-	-	-	-
16. Run errands, shop	-	.24	-	-.13	-.15	-	-	.21	-	.23	.33	.36	.22	-	.20	-

that involved examining the model modification indices for the residual correlations, identifying the largest on the basis of expected improvement in model χ^2, and then freeing that residual correlation and re-examining the model modification indices from the resulting model. This procedure was repeated until no model modification indices remained that exceeded 10. A number of substantial residual correlations were so identified (*rs* > 0.50), but the final model that included these residual correlations did not address the misfit implied by the RMSEA, which remained at 0.11. Item pairs with high residual correlations included carrying groceries and walking 15 minutes (*r* = -0.79) shampoo hair and dress self (*r* = 0.67) dress self and wash and dry body *(r* = 0.75), and shampooing hair and wash and dry body (*r* = 0.84). A large negative residual correlation implies that the single factor model over-estimates the correlation between those two

Table 4:
CFA parameter estimates and indications of DIF from a single factor, multiple group MIMIC model. PROMIS MY-Health Physical Functioning 16 item (PF16) short form

Item	**Factor**		**DIF**			
	Loading	**Thresholds**	**Sex**	**Race**	**Age**	**Education**
Vigorous activities	0.85	-0.68 -0.02 0.58 1.03	a, b	a, b	a	a, b
Walking more than one mile	0.90	-0.94 -0.45 0.02 0.38	a, b	a, b	b	a
Climbing stairs	0.91	-1.63 -0.90 -0.37 0.08	a	b	b	b
Carrying groceries	0.90	-1.60 -0.87 -0.27 0.20	a, b	b	a, b	b
Bending, kneeling, stooping	0.86	-1.59 -0.76 -0.14 0.34	a	a, b	b	
Usual physical activities	0.93	-1.79 -0.71 -0.10 0.43	a, b	a	b	a
Moderate work	0.93	-1.69 -0.93 -0.32 0.16	b	a, b	a, b	b
Physical labor	0.90	-0.99 -0.44 0.08 0.54	a, b	a, b	a, b	b
Chores	0.92	-1.28 -0.86 -0.31 0.30	b	a	a, b	
Dress yourself	0.87	-2.40 -1.81 -1.29 -0.75	b	b	a, b	a
Shampoo hair	0.91	-2.26 -1.97 -1.54 -1.13		a, b	a, b	b
Wash, dry body	0.95	-2.53 -2.01 -1.55 -1.05	b	a, b	a, b	b
Get on, off toilet	0.86	-2.70 -2.11 -1.63 -1.03	b	a, b	b	a, b
Go up, down stairs	0.91	-1.63 -1.17 -0.69 -0.10		a, b	b	
Walk 15 minutes	0.91	-1.56 -1.20 -0.76 -0.29	a, b	a, b	b	
Run errands, shop	0.91	-1.61 -1.23 -0.73 -0.24	a, b	a, b	b	b

Notes: In the DIF columns, “a” implies “a-DIF”, or a statistically significant difference across at least one of the groups in the item discrimination parameter (factor loading). “b” implies b-DIF, or a statistically significant difference across groups in the item thresholds. Factor loading and thresholds are presented for the entire sample. All models include adjustment for the effects of age, sex, race/ethnicity, and level of educational attainment.

activities, and a large positive residual correlation implies that the single factor model under-estimates the correlation between these two items. The large positive residual correlations among the basic activities of daily living items (dressing, washing hair, body, toileting) suggest a secondary factor. This was investigated *post hoc* by fitting a bifactor model with a specific factor loading in items 10-13. This factor significantly improved model fit (Mplus difference test, $\Delta\chi^2 = 1{,}141$ (*df* 4), $p < .001$) but did not resolve issues of poor fit (RMSEA = 0.094, CFI = 0.989). Because essential unidimensionality held, and identified secondary factors did not appreciably affect model fit, we retained neither the secondary factors nor the residual correlations in our differential item functioning detection models.

DIF hypotheses and results

Summary of results of the hypothesis generation

Language DIF hypotheses were posited for five items with no direction indicated: limitation with vigorous activities, limitation in walking more than a mile, limitation in doing two hours of physical labor, ability to do vacuuming or yard work, and ability to run errands and shop. Education DIF hypotheses were posited for four items, only two with directionality: conditional on physical function, individuals with lower education would be more likely to report less limitation with physical activities and with their ability to vacuum or do yard work due to physical health problems. Race/ethnicity DIF was posited for five items, suggesting for example, that conditional on physical function, Black or African-American respondents would be more likely to report less limitation in doing vigorous activities and in doing two hours of physical labor, and greater limitation in walking more than a mile than the contrast group. Some raters posited that at the same level of physical function, women will report greater limitation in walking more than a mile, in doing moderate work around the house, in doing two hours of physical labor, and less limitation in doing vigorous activities as well as in their ability to run errands and shop in contrast with men. Age DIF was posited regarding most of the items; the majority of the hypotheses were in the same direction suggesting that conditional on the physical function trait, older individuals will report greater limitation in: performing vigorous activities; in walking more than a mile; climbing one flight of stairs; in bending, kneeling, or stooping; doing usual physical activities; doing moderate work around the house; doing two hours of physical labor; vacuuming or doing yard work; going up and down stairs at a normal pace; and in ability to run errands and shop.

Summary of results of the DIF tests

DIF findings are summarized with expected item score characteristic curves (Figures 1-9). These curves display the expected item score (y-axis) as a function of the latent trait (x-axis). (See also the methods overview article in this series; Teresi & Jones, 2016.) Expected item scores ranged from 0 to 4, defined from the response categories of the PF16. We plotted expected item score characteristic curves over a range of the latent trait from [-3,+3]. Detailed tables of model parameter estimates are available upon request. As shown, group differences in the curves were small.

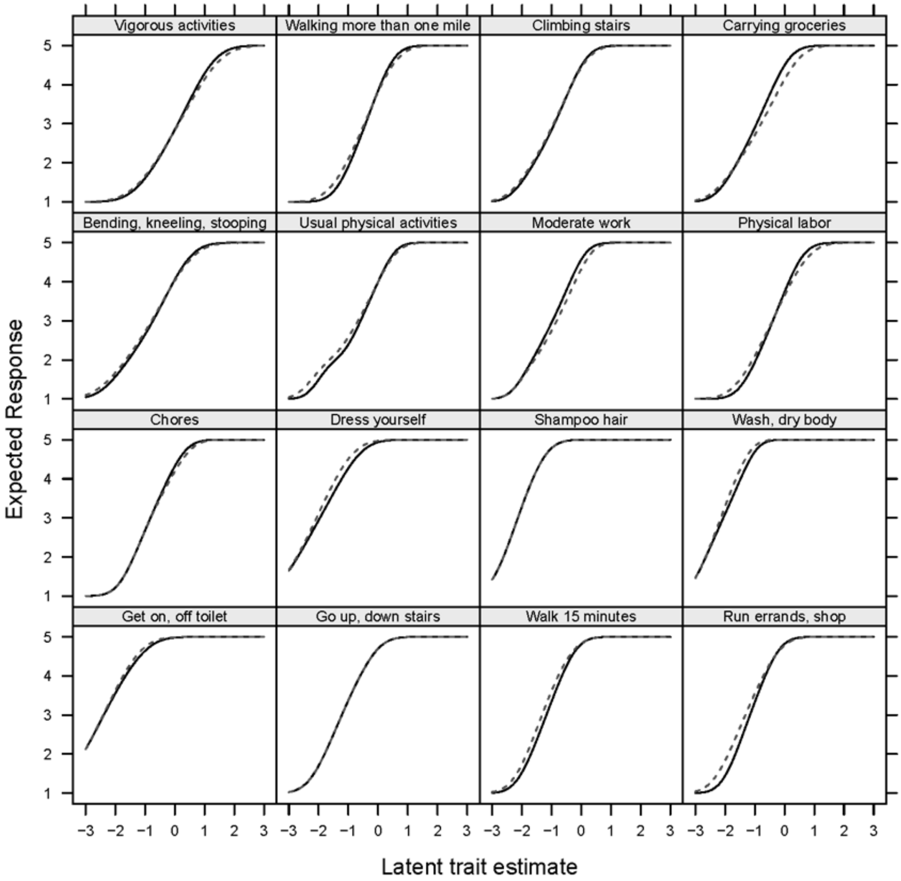

Figure 1:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Women vs. men*

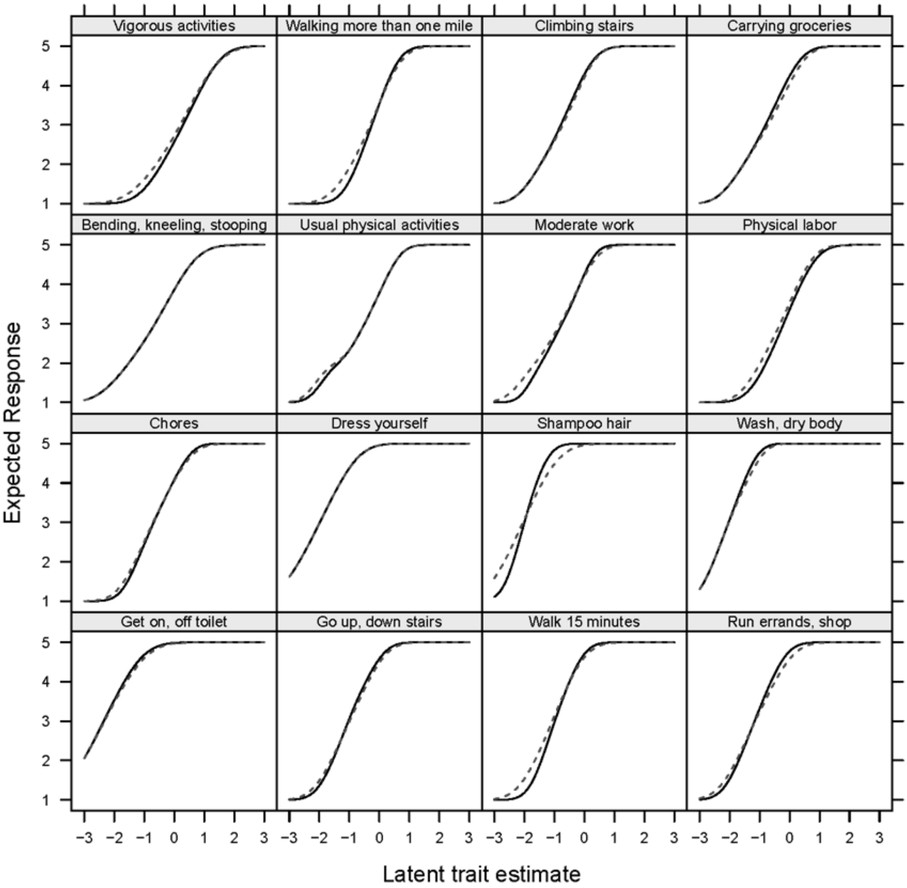

Figure 2:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Black or African-American vs. White*

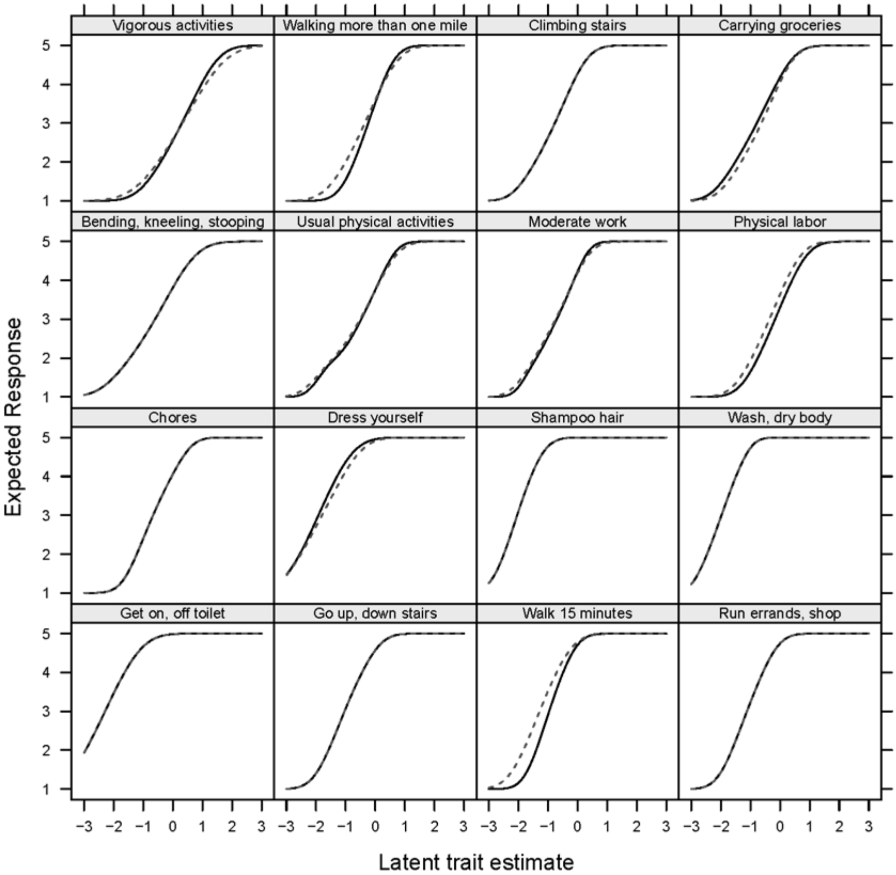

Figure 3:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Hispanic vs. White*

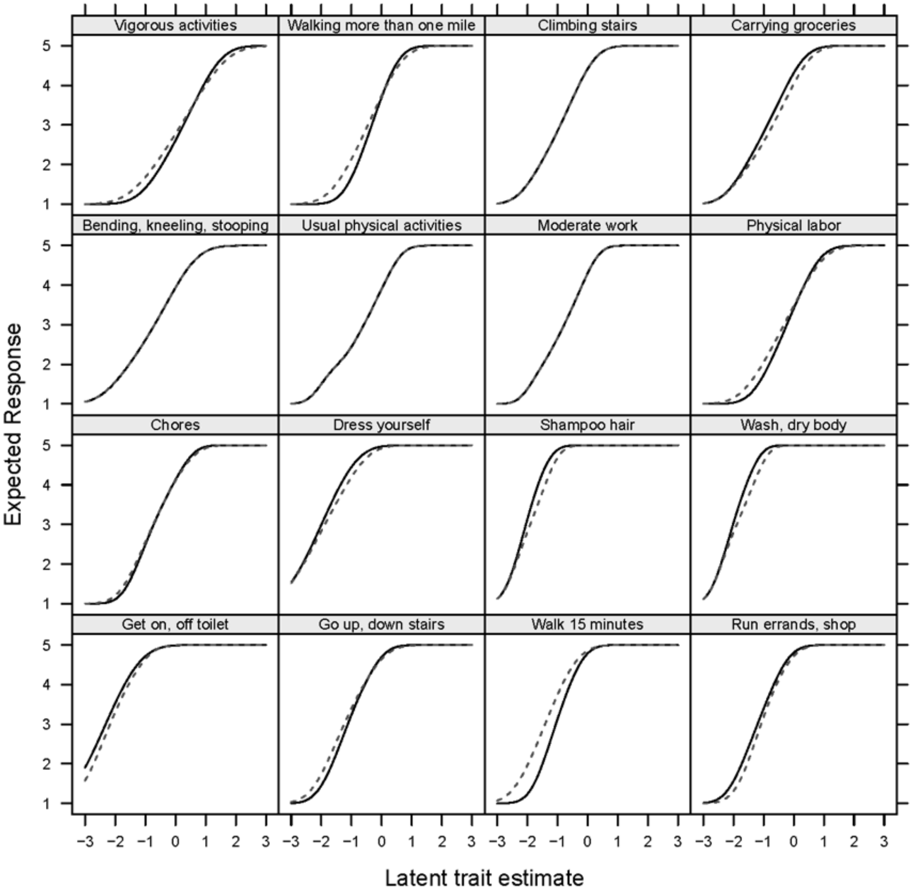

Figure 4:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Asian/Pacific Islander vs. White*

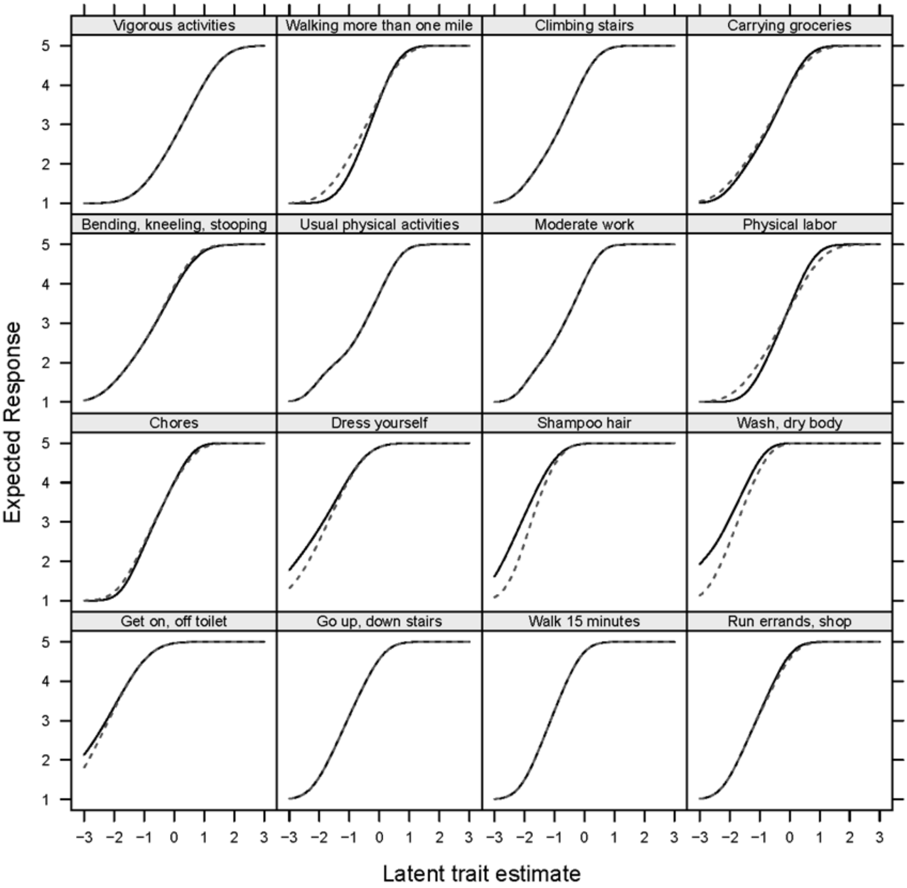

Figure 5:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Young vs. middle age*

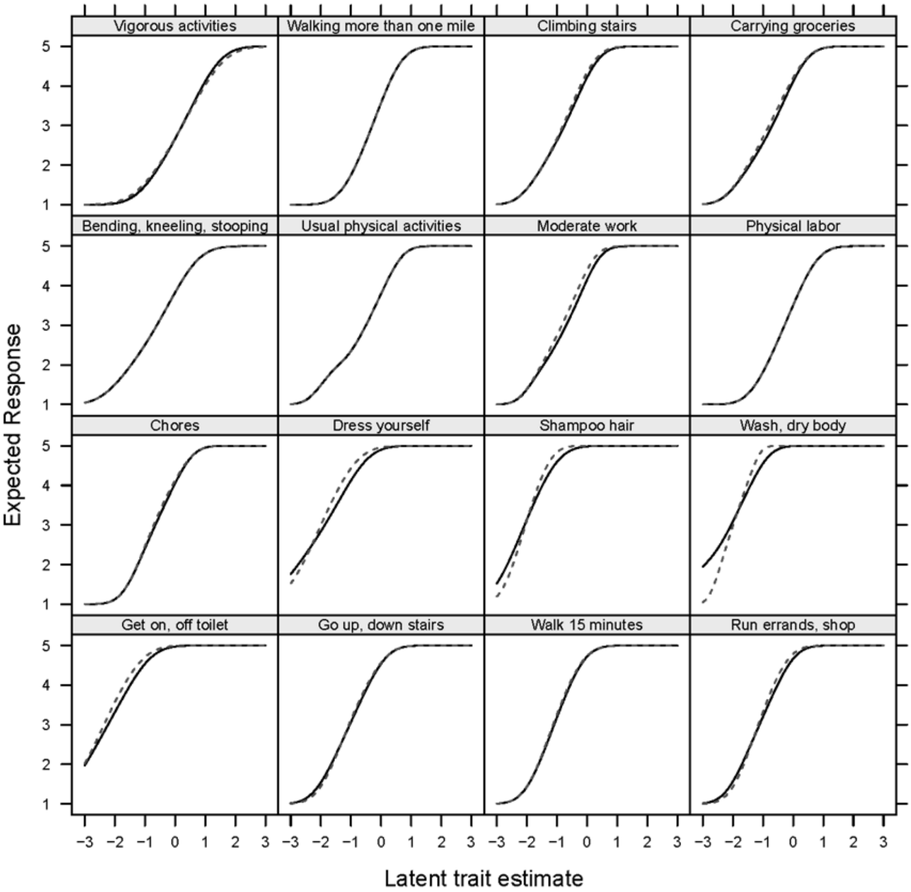

Figure 6:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Older vs. middle age*

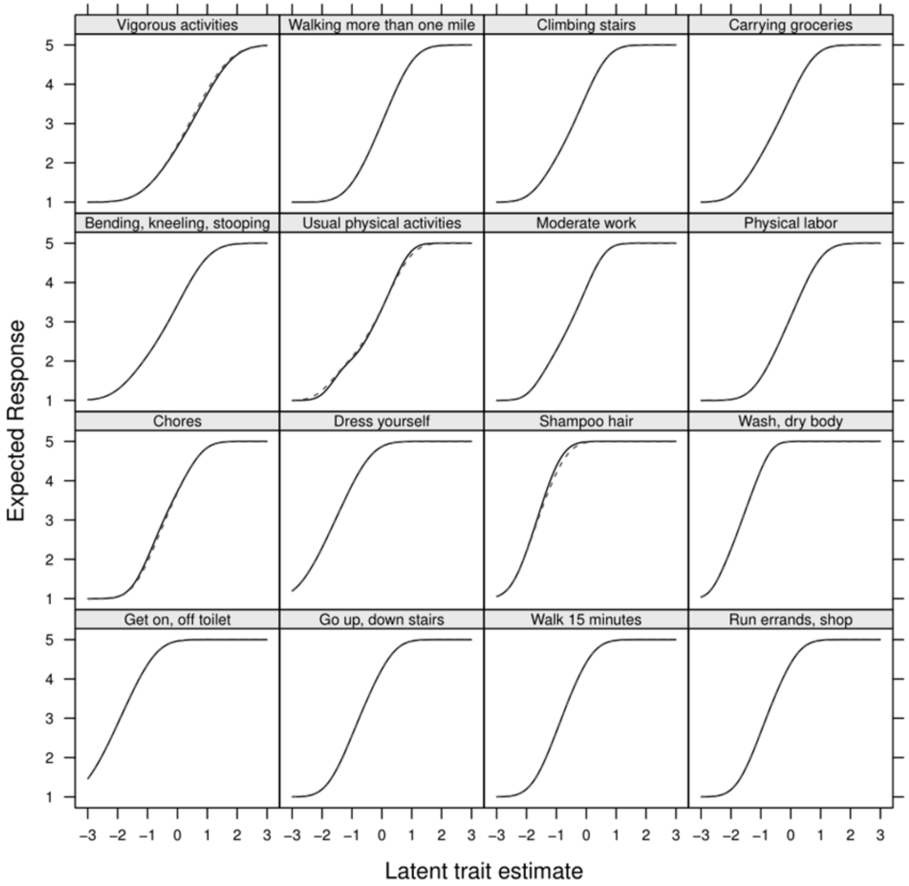

Figure 7:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Less than high school vs. high school graduate*

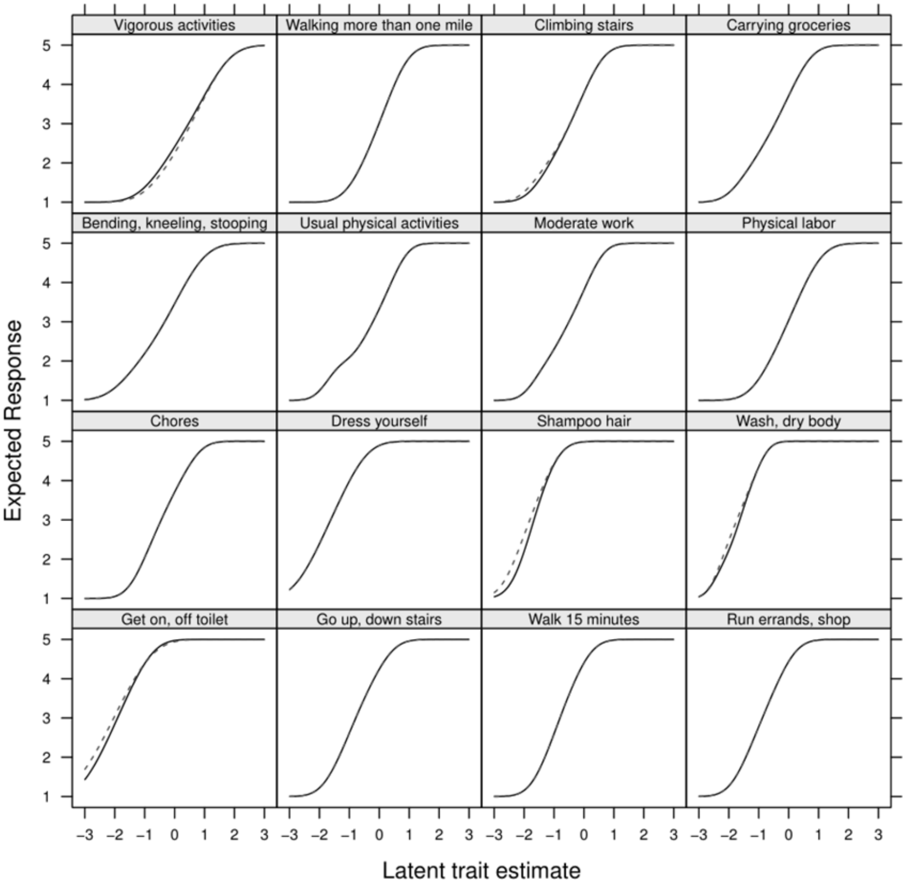

Figure 8:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *Some college vs. high school graduate*

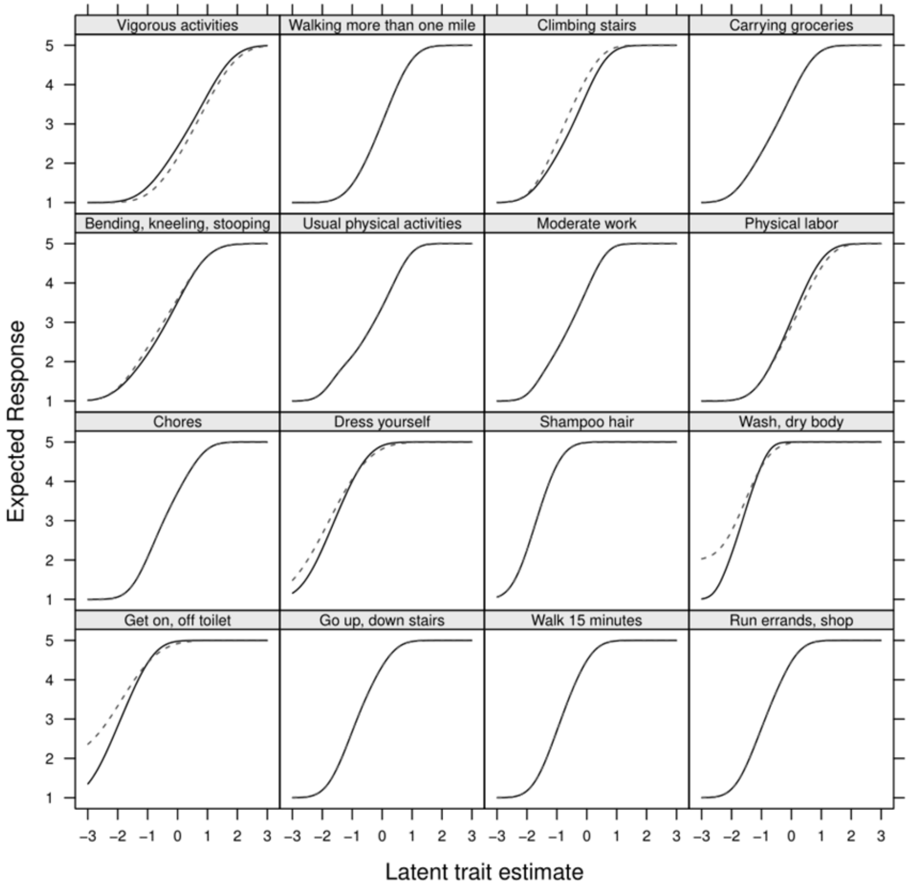

Figure 9:
Expected item score characteristic curves for a 16 item PROMIS physical functioning short form among the MY-Health subsample: *College graduate vs. high school graduate*

Sex: Most items showed DIF with respect to sex; six of these items were hypothesized to show gender DIF (see Table 4). Comparing women to men (Figure 1), we found 14 out of 16 items to evidence DIF. Of the 14 items, two evidenced DIF on the factor loading only (climb flight of stairs, kneeling/bending/stooping), four showed DIF on the thresholds only (moderate work, dress yourself, wash/dry body, get on/off toilet), and eight evidenced DIF on both the factor loading and thresholds (vigorous activities, walk a mile, carry groceries, usual physical activities, physical labor, chores, walk 15 minutes, do errands). The mean difference in physical functioning prior to DIF adjustment was -0.33 (lower scores for women). This difference was higher after adjusting for DIF (-0.34). The difference-in-differences with and without DIF adjustment is trivial, so we conclude the detected DIF was of negligible impact.

Race and Ethnicity: Most of the five items which evidenced DIF for Black or African-American respondents as contrasted with the White non-Hispanic reference group were also hypothesized to show DIF (see Table 4). For example, doing two hours of physical labor, vigorous activities and walking more than a mile were posited to show DIF, and the latter two items were also found in other studies to evidence DIF for Black in contrast to White groups. Comparing Blacks to Whites (Figure 2), we found 14 items to have DIF. Of the 14 items, one evidenced DIF on the factor loading only (chores), six on the threshold only (climb flight of stairs, carry groceries, usual physical activities, physical labor, wash/dry body, get on/off toilet), and seven on both the factor loading and threshold (vigorous activities, walk a mile, moderate work, shampoo hair, go up/down stairs, walk 15 minutes, do errands). The estimated Black/White difference in physical function was unchanged (-0.14) by adjusting for DIF. We therefore conclude the DIF was of negligible impact.

Comparing Hispanics to Whites, we found 11 items to have DIF (see Table 4). Of the 11 items, four had DIF on the factor loading only (vigorous activities, physical activities, moderate work, chores), five on the threshold only (e.g., carry groceries, kneeling/bending/stooping, physical labor, wash/dry body), and two on both the factor loading and threshold (walk a mile, walk 15 minutes). The difference in physical function between Whites and Hispanics was -0.07 before adjusting for DIF, but was -0.11 after adjusting for DIF. Proportionally the effect is substantial (a 38 % under-estimation of Hispanic group deficits in physical functioning), and in absolute terms the mean difference of 0.18 when controlling for DIF. This magnitude approached but did not meet our arbitrary threshold to flag meaningful item bias.

Comparing Asians/Pacific Islanders to Whites, we found 15 items to evidence DIF (see Table 4). Of the 15 items, two evidenced DIF on the factor loading only (chores, get on/off toilet), seven on the threshold only (climb a flight of stairs, carry groceries, physical activities, dress yourself, shampoo hair, wash/dry body, do errands), and six on both the factor loading and threshold (vigorous activities, walk a mile, kneeling/bending/stooping, physical labor, go up/down stairs, walk 15 minutes). The estimated Asian/White difference in physical function was unchanged (-0.05) by adjusting for DIF. We therefore conclude the DIF was of negligible impact. Although there were no confirmatory hypotheses regarding Asians/Pacific Islanders, the finding that most items evidenced DIF was similar to the findings reported in this series with respect to other domains, e.g., fatigue, depression, and anxiety (Reeve et al., 2016; Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016a; Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016b).

Age: Comparing the 21-49 year age group to the 50-64 year age group, we found nine items to show DIF (see Table 4). Of these nine items, one evidenced DIF on the factor loading only (carry groceries), four on the threshold only (vigorous activities, walk a mile, wash/dry body, get on/off toilet) and three on the response mean only (physical labor, dress yourself, shampoo hair). The pre-DIF adjustment mean physical functioning level for young age group is 0.15, and the post-DIF adjustment mean is 0.16, revealing DIF impact to be negligible.

Comparing the 65-84 year age group to the 50-64 year age group, we found 13 items to evidence DIF. Of these 13 items, three evidenced DIF on the factor loading only (vigorous activities, kneeling/bending/stooping, physical labor), seven on the threshold only (carry groceries, physical activities, moderate work, get on/off toilet, go up/down stairs, walk 15 minutes, do errands), and three on both the factor loading and threshold (dress yourself, shampoo hair, wash/dry body). As with the young/middle-age comparison, DIF impact was small (-0.16 vs. -0.21 with DIF adjustment).

In summary, all items showed age DIF for one or more comparisons and all items except for two were posited to show age DIF. For example, consistent with the hypotheses positing that most items would show DIF in the direction of older individuals reporting more limitation, conditional on physical function, nine items were found to have age DIF comparing those 21 to 49 vs. those 50 to 64. The items that were both found to have DIF and hypothesized to show DIF were: physical labor, dressing self, shampooing hair, toileting, vigorous activities, walking more than a mile, and carrying groceries.

Education: Comparing the less than high school group to the graduated high school group, we found three items to evidence DIF (see Table 4). Of these three items, two evidenced DIF on the threshold only (shampoo hair, wash/dry body), and one on both the factor loading and threshold (vigorous activities). The pre-DIF adjustment mean physical functioning level for the less than high school group is -0.26, and the post-DIF adjustment mean is -0.26, revealing negligible DIF impact.

Comparing the some college education group to the high school graduate group, we found six items with DIF. Of these six items, one evidenced DIF on the factor loading only (get on/off toilet), and five on the threshold only (vigorous activities, climb a flight of stairs, shampoo hair, wash/dry body, do errands). As with the less than high school/high school graduate comparison, DIF impact was trivial (0.19 vs. 0.19 with DIF adjustment).

Comparing the college graduate group to the high school graduate group, we found six items with DIF. Of these six items, one evidenced DIF on the factor loading only (dress yourself), and five on the threshold only (vigorous activities, climb a flight of stairs, physical activities, moderate work, physical labor). As with the some college/high school graduate comparison, DIF impact was trivial (0.54 vs. 0.55 with DIF adjustment).

Discussion

Key findings

We found that in this large sample of cancer patients, many of the PROMIS physical functioning items included in our 16-item measure showed evidence of DIF that satisfied criteria of statistical significance across race, ethnicity, sex, age, and education groups. However, in no case did DIF result in a substantial impact on the estimation of group mean differences in the underlying latent trait. We take this result as evidence that there is negligible DIF in the PROMIS physical functioning items among cancer patients. The possible exception is for the Hispanic reference group comparison where the impact difference estimate of 0.18 was just below the threshold.

Our logic for inferring non-negligible DIF is based on tests of statistical significance without multiple corrections correction, paired with an impact test based on differences-in-differences of the latent trait mean across reference and focal groups with and without adjustment for detected DIF. With the high statistical power afforded by the large sample size, significance testing without multiple comparisons correction, and relatively low threshold for inferring the presence of non-negligible DIF (0.2 standard normal units), the net we cast for possible DIF effects is wide. Despite this wide net, in our models examining sex, race, ethnicity, age, and education group DIF effects we find no evidence for DIF that is beyond a trivial magnitude as evidenced by the expected item score curves, and small scale-level impact at the group level. It should be noted that our logic for inferring the presence of DIF is based on a particular planned use of the test as a continuous outcome measure or index of physical functioning to be used as a continuous predictor. Other uses (e.g., screening for inclusion in a treatment or rehabilitation program) might have an impact due to DIF that is noticeable and localized around a particular region of the underlying trait relevant for making placement or eligibility decisions. The meaning of such effects can be informed by examining the item characteristic curves.

Relationship of findings to previous research

The finding that statistically significant DIF nonetheless produces negligible impact is not uncommon (Stark, Chernyshenko, & Drasgow, 2004). Teresi and colleagues (2007) evaluated physical functioning items for DIF in oncology patients assessed as part of the Quality of Life Evaluation in Oncology Project. This analysis included items drawn from four quality of life instruments, and found large DIF attributable to race for walking, energy, vigorous activities, and lifting/carrying groceries. Large sex DIF was found only for strenuous activities. However, this group also showed low impact (c.f., Figure 4 in Teresi et al., 2007). The many items with DIF observed for Hispanics are consistent with the results of Paz, Spritzer, Morales, and Hays (2013) who found numerous PROMIS items with DIF for Spanish speakers, resulting in large impact as evidenced by the differences in expected scale score functions.

Several of the items found here to evidence DIF (vigorous activities, carry groceries, walk more than a mile) were similar to those observed in previous studies to evidence DIF for race/ethnicity (Perkins, Stump, Monahan, & McHorney, 2006; Teresi et al., 2007), and were also hypothesized to show DIF. Although several items were posited to show DIF for language for the current analyses, no specific directional hypotheses were given for Spanish as contrasted with English speakers. However, the findings of DIF in physical function items in this study and that of Paz et al. (2013) suggest caution in the use of some of these items with Hispanic groups.

Although the focus of this article was measurement equivalence, concurrent analyses of validity were performed by Jensen and colleagues (2015). Using the same data set, the performance of several physical function short forms was examined: 4a, 6b, 10a and the total 16 item version examined here. These authors provided evidence supporting the convergent, discriminant and known groups validity for all forms. However, ceiling effects across race/ethnic and age groups were observed at higher levels of function, particularly with respect to the 4a short form. Studies of these effects (Fries, Krishman, Rose, Lingala, & Bruce, 2011) and efforts to redress this problem through the use of CAT (Rose et al., 2014) and addition of new items (Bruce, Fries, Lingala, Hussain, & Krishman, 2013; Fries et al., 2014) are ongoing.

Limitations

The MY-Health sample was based on the presence of five cancers, some of which are exclusive to females, others males. Because our analyses did not control for cancer type, it is plausible that the observed sex differences instead reflect differences attributable to specific cancer experiences. The sample we use is a mixed clinical sample and not representative of any single clinical sample. The different cancer patient populations have strong selection biases due to sex, and because the different cancers may have different treatments, course and prognosis, cancer type remains an important potential confounder not addressed in this analysis. Our goal was to address evidence for bias due to socio-demographic characteristics. Additional research is necessary to address potential bias due to cancer diagnosis.

Our analytic approach assumes that data are missing completely at random. This is almost certainly not true. Nevertheless, missing data were very rare (< 3 %) and violations of this assumption are not likely to have biased our results.

Another limitation we identify is that item responses to the PF16, as applied to cancer patients in this sample, are unidimensional but many item pairs do not adhere to the local independence assumption. Moreover, we have some evidence for a non-trivial secondary factor specific to items assessing basic activities of daily living (ability to wash and dry body, get on and off the toilet, dress, and shampoo hair), and our analysis does not address this multidimensionality. Jensen and colleagues (2015) also observed the presence of a secondary factor comprised of the four self-care items. In their analysis, this factor correlated highly with the first factor, and the items also loaded substantially on the first factor. Had we found evidence for practically meaningful DIF, examining the extent that these results are due to relationships between the background variable and specific factor would have been an important sensitivity analysis.

The discrepancy observed between the CFI and RMSEA results can be explained conceptually. The CFI assesses the fit of the model relative to a model where all included variables are mutually uncorrelated, whereas the RMSEA assesses fit of the model relative to a model with optimally chosen parameters (Hooper, Coughlan, & Mullen, 2008). The good fit by CFI for this single group CFA reflects that, as a whole, the items in the PF16 correlate highly with one another (i.e., far from zero) and the single factor model is much preferred over a null model to capture these correlations. The higher than optimal RMSEA, in our case, implies that there may be some unmeasured associations among the items. In our sample, the model-implied variance/covariance matrix derived from the single factor model is not close to an optimal solution.

A (residual) correlation between two test items is conceptually equivalent to the effect of a specific factor that loads equally in the pair of items. Therefore, finding evidence of violations of the local independence assumption may imply that among ethnically diverse cancer patients, pairs of items are highly correlated and some of this correlation has nothing to do with a common factor that drives impairment in other areas of physical functioning. The over-estimation of the correlation between pairs of items implies that some cancer patients may have a specific impairment in some activities and not others. Such speculative hypotheses are not able to be tested in the current sample, but speak to the need for further construct validation of the physical functioning domain in cancer patients. Because local dependencies such as those observed in these analyses can result in biased parameter estimates including inflated slopes as well as inaccurate standard error estimates, there is a potential for false DIF detection (Houts & Edwards, 2013). Taken together, these findings limit our ability to draw conclusive inferences based on the models of DIF presented.

Conclusion

Given the assumptions of unidimensionality and local independence (which these data tend to challenge), we only find evidence for DIF of trivial importance in the PROMIS physical functioning domain assessed with a static 16 item short form, with the possible exception of the comparison involving the Hispanic subgroup. Additional work is needed to understand the meaning and importance of the failure of the selected item set to comply with the local independence assumption.

Acknowledgements

Partial funding for these analyses was provided by the National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971 (PI: Potosky, Moinpour). We are grateful for the helpful comments of a reviewer in pointing out the multidimensionality suggested by the high correlation among the basic activities of daily living items. We thank Stephanie Silver, MPH for editorial assistance with the preparation of this manuscript.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Bruce, B., Fries, J. F., Ambrosini, D., Lingala, B., Gandek, B., Rose, M., & Ware Jr, J. E. (2009). Better assessment of physical function: Item improvement is neglected but essential. *Arthritis Research and Therapy, 11*(6), R191. doi: 10.1186/ar2890.

Bruce, B., Fries, J. F., Lingala, B., Hussain, Y. N., & Krishman, E. (2013). Development and assessment of floor and ceiling items in the PROMIS physical function item bank. *Arthritis Research and Therapy, 15*(5), R144. doi: 10.1186/ar4327.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509. Retrieved from http://ezp-prod1.hul.harvard.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=6377889&site=ehost-live&scope=site doi: 10.1207/s15327906mbr2704_2

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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. doi: 10.1007/s11136-009-9464-4

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education, 24*(3), 189-209. doi:10.1080/08957347.2011.580255

FDA. (2009). *Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Retrieved from Silver Spring, Maryland: http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm193282.pdf

Feinstein, A. R., Josephy, B. R., & Wells, C. K. (1986). Scientific and clinical problems in indexes of functional disability. *Annals of Internal Medicine, 105*(3), 413-420. doi:10.7326/0003-4819-105-3-413

Fried, L. P., Ettinger, W. H., Lind, B., Newman, A. B., & Gardin, J. (1994). Physical disability in older adults: A physiological approach. Cardiovascular Health Study Research Group. *Journal of Clinical Epidemiology, 47*(7), 747-760. doi:10.1016/0895-4356(94)90172-4

Fries, J. F., Krishman, E., Rose, M., Lingala, B., & Bruce, B. (2011). Improved responsiveness and reduced sample size, size requirements of PROMIS physical function scales, with item response theory. *Arthritis Research & Therapy, 13*(5), R147. doi: 10.1186/ar3461

Fries, J. F., Lingala, B., Siemons, L., Glas, C. A., Cella, D., Hussain, Y. N., ...Krishman, E. (2014). Extending the floor and the ceiling for assessment of physical function. *Arthritis & Rheumatology, 66*(5), 1378-1387. doi: 10.1002/art.38342

Fries, J. F., Witter, J., Rose, M., Cella, D., Khanna, D., & Morgan-DeWitt, E. (2014). Item response theory, computerized adaptive testing, and PROMIS: Assessment of physical function. *The Journal of Rheumatology, 41*(1), 153-158. doi:10.3899/jrheum.130813

Glasgow, R. E., & Riley, W. T. (2013). Pragmatic measures: What they are and why we need them. *American Journal of Preventive Medicine, 45*(2), 237-243. doi: 10.1016/j.amepre. 2013.03.010

Harris, P., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. (2009). Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377-381. Retrieved from http://www.sciencedirect.com/science/ article/B6WHD-4TJTX88-1/2/e77f1e54bba8c75de397340dad11aac6. doi: 10.1016/j.jbi. 2008.08.010

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915. doi:10.1177/0013164403261769

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Buisness Reseach Methods, 6*(1), 53-60. doi: 10.5402/2012/301325

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37*(7), 541-562. doi: 10.1177/0146621613491456

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. doi:10.1080/10705519909540118

Jensen, R. E., Potosky, A. L., Reeve, B. B., Hahn, E., Cella, D., Fries, J., …. Moinpour, C. M. (2015). Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Quality of Life Research, 24,* 2333-2344. doi: 10.1007/s11136-015-0992-9

Jensen, R. E., Rothrock, N. E., DeWitt, E. M., Spiegel, B., Tucker, C. A., Crane, H. M., . . . Shulman, L. M. (2015). The role of technical advances in the adoption and integration of patient-reported outcomes in clinical care. *Medical Care, 53*(2), 153-159. doi: 10.1097/MLR.0000000000000289

Jones, R. N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health, 7*(2), 83-102. doi:10.1080/1360786031000045872

Jones, R. N. (2006a). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *Medical Care, 44*(11 Suppl 3), S124-S133. doi: 10.1097/01.mlr.0000245250.50114

Jones, R. N. (2006b). Technical appendix for identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting differential item functioning using MIMIC modeling. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/235705103_Technical_Appendix_for_Identification_of_Measurement_Differences_Between_English_and_Spanish_Language_Versions_of_the_Mini-Mental_State_Examination_Detecting_Differential_Item_Functioning_Using_MIMIC_Modeling._Jones_RN._Med_Care_200644_S124S133/file/9fcfd512c7ba22abb9.pdf

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 10*, 631-639. doi: 10.2307/2285946

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis review, assessment, and new approaches. *Educational and psychological measurement, 75*(1), 22-56. doi: 10.1177/0013164414529792

NIH. (2003). RFA-RM-04-011: Dynamic assessment of patient-reported chronic disease outcomes. Retrieved from http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-011.html

Orlando-Edelen, M., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*, 5-18. doi:10.1007/s11136-007-9198-0

Orlando-Edelen, M., Stucky, B. D., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: Application to a cancer stigma index. *Quality of Life Research, 24*(1), 95-103. doi:10.1007/s11136-013-0540-4

Paz, S. H., Spritzer, K. L., Morales, L. S., & Hays, R. D. (2013). Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items. *Quality of Life Research, 22*(7), 1819-1830. doi: 10.1007/s11136-012-0292-6

Perkins, A. J., Stump, T. E., Monahan, P. O., & McHorney, C. A. (2006). Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Quality of Life Research, 15*(3), 331-348. doi:10.1007/s11136-005-1551-6

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Reeve, B. B., Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., … & Chen, W-. H. (2016). Psychometric evaluation of the PROMIS fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling, 58,* 119-139.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48. doi: 10.1146/annurev.clinpsy.032408.153553

Rose, M., Bjorner, J. B., Becker, J., Fries, J., & Ware, J. (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. doi:http://dx.doi.org/10.1016/j.jclinepi.2006.06.025

Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. (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. doi: http://dx.doi.org/10.1016/j.jclinepi.2013.10.024

Rothrock, N. (2014). *Accessing and using PROMIS instruments*. Presentation. Northwesterm University. Evanston, IL. Retrieved from http://www.geneticalliance.org/sites/default/files/webinararchive/062414Rothrock.pdf

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100. doi: 10.1007/BF02290599

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*(3), 497. doi: 10.1037/0021-9010.89.3.497

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402. doi: 10.1007/s11136-011-9969-5

Teresi, J. A., & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, "Measurement Equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) Short Form Measures". *Psychological Test and Assessment Modeling*, *58,* 37-78.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K., Gibbons, L. E., . . . Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research, 16*, 43-68. doi: 10.1007/s11136-007-9186-4

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System (PROMIS®) depression short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58,* 141-181.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016b). Measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS®) anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling, 58,* 183-219.

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of Aging and Health, 24*(6), 1044-1076. doi:10.1177/0898264312436877

Wagner, L. I., Schink, J., Bass, M., Patel, S., Diaz, M. V., Rothrock, N., . . . Cella, D. (2014). Bringing PROMIS to practice: Brief and precise symptom screening in ambulatory cancer care. *Cancer, 121,* 927-934. doi: 10.1002/cncr.29104