American Journal of Public Health Research, 2015, Vol. 3, No. 4, 136-147
Available online at http://pubs.sciepub.com/ajphr/3/4/3

DOI:10.12691/ajphr-3-4-3

Evaluation of the Short-Form Health Survey (SF-36) Using the Rasch Model

Peter D. Hart[1,2,*], Minsoo Kang[3], Norman L. Weatherby[3], Yun Soo Lee[3], Tom M. Brinthaupt[4]

[1]Health Promotion, Montana State University - Northern, Havre, MT
[2]Research and Statistical Consultant, Health Demographics, Havre, MT
[3]Department of Health and Human Performance, Middle Tennessee State University, Murfreesboro, TN
[4]Department of Psychology, Middle Tennessee State University, Murfreesboro, TN
*Corresponding author: peter.hart@msun.edu

Received May 26, 2015; Revised June 15, 2015; Accepted July 02, 2015

Abstract *Introduction*. Health-related quality of life (HRQOL) is an outcome variable of growing importance in chronic disease research. Many intervention-type studies seek to show improvements in HRQOL based on treatment effects. As interest grows in using HRQOL as an outcome measure, the need to investigate the measurement properties of HRQOL assessments increases in importance. *Objective*. The purpose of this study was to evaluate the SF-36 for proper measurement functioning using the Rasch model. *Methods*. A total of 634 participants completed the SF-36 HRQOL assessment. The Rasch partial credit model was used to analyze the two dominant HRQOL domains (physical and mental) of the assessment *Results*. Majority of the total criteria used for optimal category functioning were met for the physical health domain and all of the total criteria were met for the mental health domain. Both convergent and construct validity evidence provided substantial confirmation for the use of the Rasch physical and mental health person scores as measures of HRQOL. *Conclusion*. Results of this study showed that the SF-36 met stringent modern measurement criteria using the Rasch model.

Keywords: *chronic disease, health-related quality of life, psychometrics, Rasch measurement*

Cite This Article: Peter D. Hart, Minsoo Kang, Norman L. Weatherby, Yun Soo Lee, and Tom M. Brinthaupt, "Evaluation of the Short-Form Health Survey (SF-36) Using the Rasch Model." *American Journal of Public Health Research*, vol. 3, no. 4 (2015): 136-147. doi: 10.12691/ajphr-3-4-3.

1. Introduction

Health-related quality of life (HRQOL) is an outcome measure that has seen considerable attention in public health research [1]. HRQOL is a broad concept that generally includes dimensions of physical, mental, and social well being. Wilson and Cleary expand on the complexity of HRQOL by stating that HRQOL is a function of biological and physiological variables, symptom status, functional status, and general health perceptions [2]. Because HRQOL is such an all-encompassing health measure, it has become a standard outcome variable in public health research [3]. HRQOL has also shown to be a valuable predictor of health status, predicting the number of physician visits, hospitalization events, and mortality among adults [4].

Item response theory (IRT) is a modern approach to measurement theory. IRT works differently from classical test theory (CTT) in that it focuses on each item by examining the response of an individual at a specific ability level and the characteristics of that item [5]. An IRT model that is only concerned with an item's difficulty level (*b*-parameter) and the individuals' ability (θ), is considered a 1-parameter model, and commonly called a Rasch measurement model [6]. Rasch analysis can be applied to health and behavioral assessments containing dichotomous response (yes/no) items, polytomous response (Likert-type) items, or a mix of both [7].

Given the overwhelming interest in HRQOL as a measure in physical activity research, there is a strong need for a better understanding of the measurement properties of HRQOL assessments commonly used in physical activity research. The Short Form-36 Health Survey (SF-36) is the leading HRQOL assessment used in physical activity research. The majority of physical activity researchers use either one or both of the SF-36 domain component scores (physical and mental). There are currently no studies that assess the measurement properties of these two commonly used domains using the Rasch measurement model. Therefore, the purpose of this study was to evaluate the measurement properties of each SF-36 domain using the Rasch model. The results of this study will serve as a critical evaluation of the SF-36 and possibly find needed modifications due to poor measurement properties or validate its continued use.

2. Methods

Participants. Data for this study came from a survey administered to adults in and around a large southeastern U.S. university community. A covenience sample is

appropriate for Rasch analysis because of the sample invariant item and trait characteristics of IRT [8]. Therefore, participants were recruited via public advertisement and announcements to local social group networks. Participants were allowed to complete the survey if they were 18 years of age or older. Human subject clearance was obtained before conducting research from the campus Institutional Review Board. Each HRQOL assessment was converted to electronic form for web-based administration and the ordering of HRQOL assessments was counterbalanced. The online survey took approximately 15 minutes to complete. Participants completed the survey during the months of January-February, 2012.

A total of 634 participants completed the SF-36 HRQOL assessment of which 72.2% were female (see Table 2). For age, 54.3% were between 18 and 24 years, 32.0% between 25 to 49 years, and 13.4% were between 50 and 78 years. For race, majority (83.4%) of the participants were White followed by Black (9.1%). Of the participants, 3.2% reported having only a high school education or less, 50.0% reported having some college education and 46.5% reported having completed a college degree. Finally, 60.3% of participants reported being single, 18.3% reported being married, 13.0% reported being either separated or divorced, 6.7% reported living with a partner, and 1.4% reported being widowed.

HRQOL assessments. The *Short-Form Health Survey (SF-36)* is one of the most widely used HRQOL instruments in physical activity research. The SF-36 was developed from the Medical Outcomes Study (MOS) conducted by RAND [9]. The SF-36 is a multi-dimensional scale consisting of 36 items, 8 health-related dimensions, and two domains (see Table 1). The dimensions include: 1) vitality, 2) physical functioning, 3) bodily pain, 4) general health, 5) physical role functioning, 6) emotional role functioning, 7) social role functioning, and 8) mental health. The physical domain consists of the physical functioning, bodily pain, general health, and physical role functioning dimensions and the mental domain consists of the vitality, emotional role functioning, social role functioning, and mental health dimensions [10].

The SF-36 is intended to measure HRQOL in adults and can be self-administered, administered via computer, with aid of an interviewer, or by telephone. The instrument can be modified to include either a (standard) 4-week recall or a 1-week recall and has been incorporated into both observational as well as intervention-type studies. The *SF-12* is a shorter version of the original that maintains the measurement of all 8 dimensions as well as the two domain-specific summary scores [11].

The *CDC Healthy Days* module (HRQOL-9) was administered for the validity portion of this study. The HRQOL-9 is a widely used module in national surveillance systems such as the National Health and Nutrition Examination Survey (NHANES) and the Behavioral Risk Factor Surveillance System (BRFSS). The HRQOL-4 is a simple 4-item tool for assessing HRQOL in large scale studies and is considered the core (see Table 1). The first item assesses perceived general health and asks respondents to rate their health in general on a 5-point categorical scale ranging from *excellent* to *poor*. The last three items ask for the number of days out of the previous 30 in which (1) your physical health was poor, (2) your mental health was poor, and (3) you were unable to engage in usual activities due to poor health. A fifth summary measure of healthy days (or unhealthy days) can be computed by summing the physical and mental items and creating a ceiling at 30 days [3]. For the current study, the Healthy Days Index was used for a validation of the SF-36 ability scores. The Healthy Days Index contains two items which combined represent both domains of HRQOL [12].

Table 1. Characteristics of participants completing the SF-36 HRQOL assessment (N = 634)

Characteristic	*N*	%
Gender		
Male	175	27.6
Female	458	72.2
Age Group		
18-24	344	54.3
25-49	203	32.0
50-78	85	13.4
Race		
White	529	83.4
Black	58	9.1
Hispanic	8	1.3
Asian	12	1.9
Other	24	3.8
Education		
High School or less	20	3.2
Some College	317	50.0
College Degree	295	46.5
Marital Status		
Single	382	60.3
Married	116	18.3
Separated/Divorced	82	13.0
Widowed	9	1.4
Living w/Partner	42	6.7

Table 2. Number of Items by Domain of the SF-36 and CDC HRQOL-4 assessment tools

SF-36		CDC HRQOL-4	
Domain	Items	Domain	Items
Physical Health	21	Physical Health	3
Mental Health	14	Mental Health	1
		Healthy Days Index	1
Total Items	35		5

Note. Healthy Days Index is a composite variable from items 2 and 3 of the HRQOL-4 core.

Rasch model. The Rasch model is a probability model which includes a persons' ability and an item's difficulty as parameters. The Rasch model converts responses from a rating scale to a new scale with interval level measurement properties [7]. The new scale values are called logits (log odds) and are so for a persons' ability (θ) as well as an item's difficulty (*b*). Logits take the same presence as Z-scores, with a mean of zero. A person with

a positive logit generally has a greater "ability" concerning the trait being measured (i.e., has a higher overall HRQOL) and a person with a negative logit generally has a lower ability concerning the trait. An item with a positive logit generally indicates higher item "difficulty" and an item with a negative logit generally indicates lower item difficulty [7]. A larger item difficulty indicates that individuals are less likely to endorse that item.

The primary assumption of the Rasch model is that the measurement scale should be unidimensional. For this study, this means that each scale should measure its respective HRQOL domain and nothing more. This assumption can be examined by examining item fit statistics. Once data are fit to the Rasch model and the assumption of unidimensionality is met, a researcher can proceed in inspecting several of the Rasch model statistics. Person reliability estimates and item reliability estimates are reported from a Rasch analysis and provide analogous information as that of Cronbach's alpha, with a range of 0 to 1.00. Person separation and item separation indices are standard error units representing the spread or separation of persons (or items) on the ability scale, where a larger value indicates the scale's ability to better separate persons (or items). The Infit and Outfit statistics from a Rasch analysis are mean square statistics with expected values of one and an acceptable range of 0.50 to 1.50 [13,14]. Item Infit and Outfit statistics provide evidence of construct validity. Person Infit and outfit statistics represent whether individuals respond in an expected way given their response pattern and item difficulty [7].

Proper category functioning can also be examined by the Rasch model. Item-person map (Wright map) distributions can be examined from a Rasch analysis. The item-person map is a single dimensional graph linking item difficulty and person ability estimates on the same common scale (logits). The item-person map shows both distributions as well as the relative position of an individual's trait (i.e., HRQOL) for the items.

Data analysis. The plan was to run two separate analyses on the two HRQOL domains (physical & mental) of the SF-36 assessment. A 7-step procedure was followed to evaluate each SF-36 domain by Rasch analysis. The first step included evaluating each item for proper category functioning. The evaluation criteria included (1) regular frequency distributions, (2) average logit score measures increasing as categories increase, (3) Infit and Outfit mean square residuals are appropriate for each category, and (4) category thresholds arranged in order [15,16]. The second step included an evaluation of model-data fit. The model-data fit criteria included inspection of the Infit and Outfit statistics for each item. If these fit statistics were greater than 1.5 or less than 0.5, the item was considered misfit [17] and were subsequently discarded. The third step included an inspection of the item-person map. This step evaluates how evenly spread the items are relative to the participants in terms of the HRQOL trait. The fourth step included the evaluation of each item in terms of item difficulty parameters, item separation, and item separation reliability. The fifth step involved the evaluation of individuals fitting the Rasch model in terms of person ability (θ) fit, person separation index, and person separation reliability. The sixth step included convergent validity evidence for the SF-36 domains by computation of bivariate correlations between each of the SF-36 ability (θ) scores and the CDC Healthy Days Index from the HRQOL-4 core. The seventh and final step included construct validity evidence for the SF-36 ability (θ) scores by showing differences in the scores between groups of participants with known theoretical differences in HRQOL. The grouping variables were all dichotomized (yes/no) and included obesity, smoking status, chronic illness, vigorous activity participation, moderate activity participation, strength training participation, hypertension, high cholesterol, and diabetes. All analyses were carried out using SAS version 9.3 and Winsteps v3.65 [18].

3. Results

Table 3 displays distribution information for the SF-36 physical domain items. Twenty one items had responses across all categories. Relative frequencies per category ranged from .005 to .959 across all items in the physical domain. Ten items have categorical rating scales consisting of 3 points, another ten items have a 5-point scale, and one item is a 6-point scale. The overall average response across all 21 items was 3.55, ranging from 2.45 to 4.74. All items were coded to reflect greater HRQOL with higher scores. Table 4 displays item distribution information for the SF-36 mental domain. Relative frequencies per category ranged from .008 to .549 across all items in the mental domain. All 14 items were on a 5-point scale and each had responses across all categories. The overall average response across all 14 items was 3.76, ranging from 2.91 to 4.28. Each item in the mental domain was also coded to reflect greater HRQOL with higher scores.

Table 3. Item category distributions (%) and item means of the physical health domain of the SF-36 HRQOL assessment (N = 634)

		Item Categories					
Item	*M*	1	2	3	4	5	6
SF1	3.56	0.5	9.1	36.4	41.3	12.6	*
SF3a	2.45	10.7	33.4	55.8	*	*	*
SF3b	2.87	2.2	8.2	89.6	*	*	*
SF3c	2.91	1.4	6.2	92.4	*	*	*
SF3d	2.71	4.6	19.9	75.6	*	*	*
SF3e	2.90	1.6	7.3	91.2	*	*	*
SF3f	2.78	3.5	14.7	81.9	*	*	*
SF3g	2.83	3.6	10.1	86.3	*	*	*
SF3h	2.89	2.5	6.2	91.3	*	*	*
SF3i	2.91	2.7	4.1	93.2	*	*	*
SF3j	2.93	2.7	1.4	95.9	*	*	*
SF4a	4.60	0.8	1.7	6.8	18.5	72.2	*
SF4b	4.28	1.6	6.9	10.7	23.7	57.1	*
SF4c	4.60	0.9	2.5	6.2	16.1	74.3	*
SF4d	4.55	0.9	2.4	8.0	17.8	70.8	*
SF7	4.74	0.5	3.3	11.2	18.9	39.1	27.0
SF8	4.52	0.5	3.2	6.6	23.8	65.9	*
SF11a	4.12	3.5	9.6	7.7	30.3	48.9	*
SF11b	3.75	6.0	10.1	15.5	40.1	28.4	*
SF11c	4.06	1.6	7.3	19.4	26.8	45.0	*
SF11d	3.67	5.4	12.0	12.6	50.2	19.9	*

Note. Categories reflect reverse coding with higher categories representing higher HRQOL. *Represents a category which is not present for the item.

Table 4. Item category distributions (%) and item means of the mental health domain of the SF-36 HRQOL assessment (N = 634)

		Item Categories				
Item	*M*	1	2	3	4	5
SF5a	4.22	1.4	3.3	18.3	25.7	51.3
SF5b	3.96	2.1	10.3	17.7	30.0	40.1
SF5c	4.26	0.8	5.8	14.4	24.1	54.9
SF6	4.15	1.7	5.8	13.9	32.8	45.7
SF9a	3.49	2.1	13.1	28.7	46.5	9.6
SF9b	3.69	2.2	11.4	24.4	39.4	22.6
SF9c	4.28	1.1	4.6	14.4	25.6	54.4
SF9d	3.30	3.6	17.2	28.9	45.7	4.6
SF9e	3.22	4.4	17.2	36.3	36.4	5.7
SF9f	3.95	2.2	6.5	18.6	39.4	33.3
SF9g	3.22	7.3	17.0	31.4	34.9	9.5
SF9h	3.68	0.8	8.4	24.4	55.2	11.2
SF9i	2.91	10.6	22.1	36.4	27.3	3.6
SF10	4.27	0.8	4.1	15.9	25.9	53.3

Note. Categories reflect reverse coding with higher categories representing higher HRQOL.

Table 5 displays the criteria for the SF-36 physical health domain. Overall, the item categories for the physical health domain functioned well, meeting 87.0% of the total criteria used to evaluate proper functioning. A total of ten items were flagged for negative validity. Two items (4a, 8) had Outfit mean square values greater than 2.0. Seven items (3b, 3c, 3g, 3h, 3i, 11a, 11d) lacked ordered thresholds. And one item (3j) had both an Outfit mean square greater than 2.0 and disordered thresholds. Figure 1 displays the category probability curves for the SF-36 physical health domain items. The graphs depict unordered thresholds for the eight items mentioned. For the mental health domain, the item categories functioned very well, meeting 100% of the total criteria used to evaluate proper functioning. Table 6 displays the criteria for the SF-36 mental health domain and Figure 2 displays proper ordering of the category thresholds.

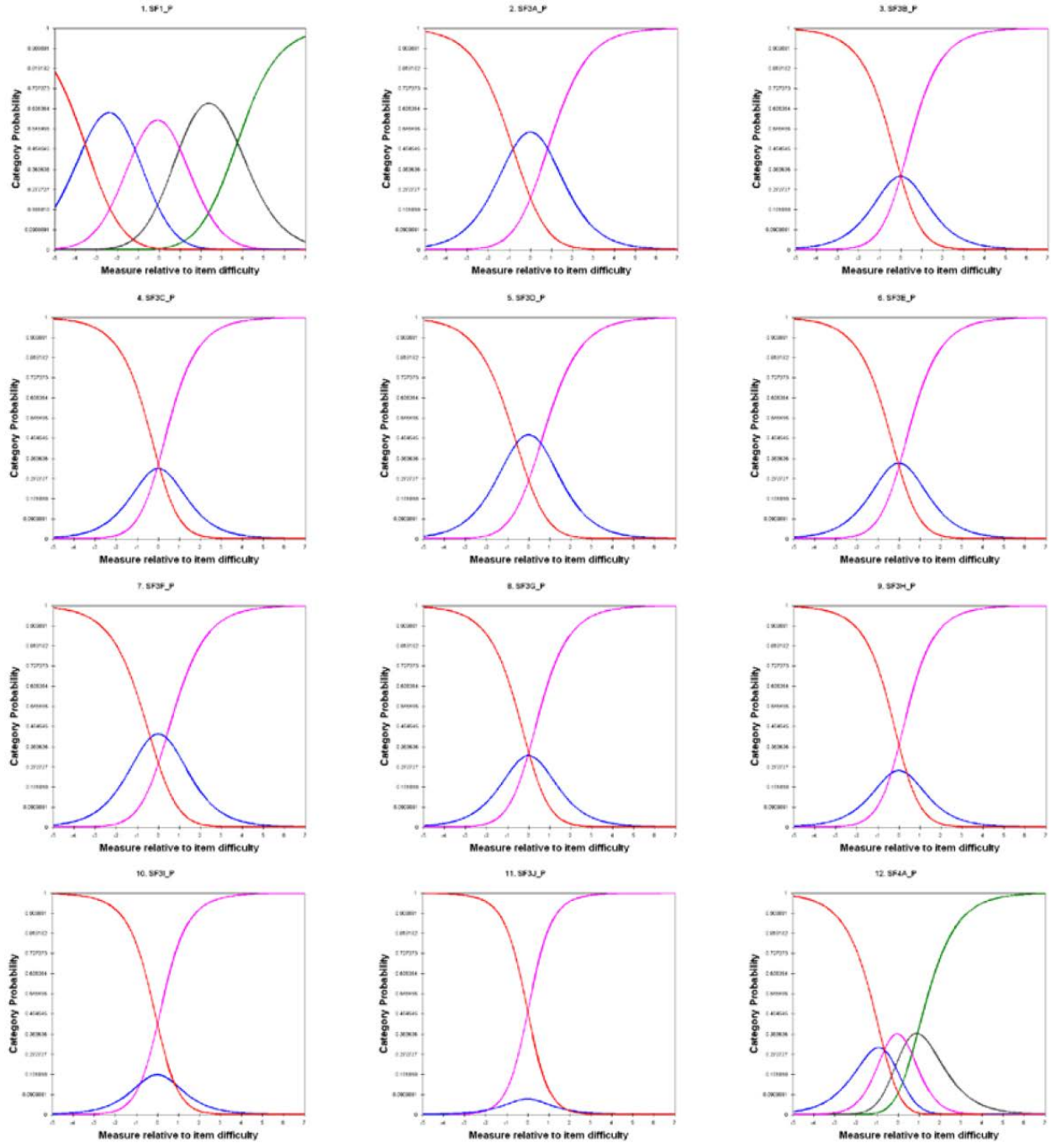

Figure 1.

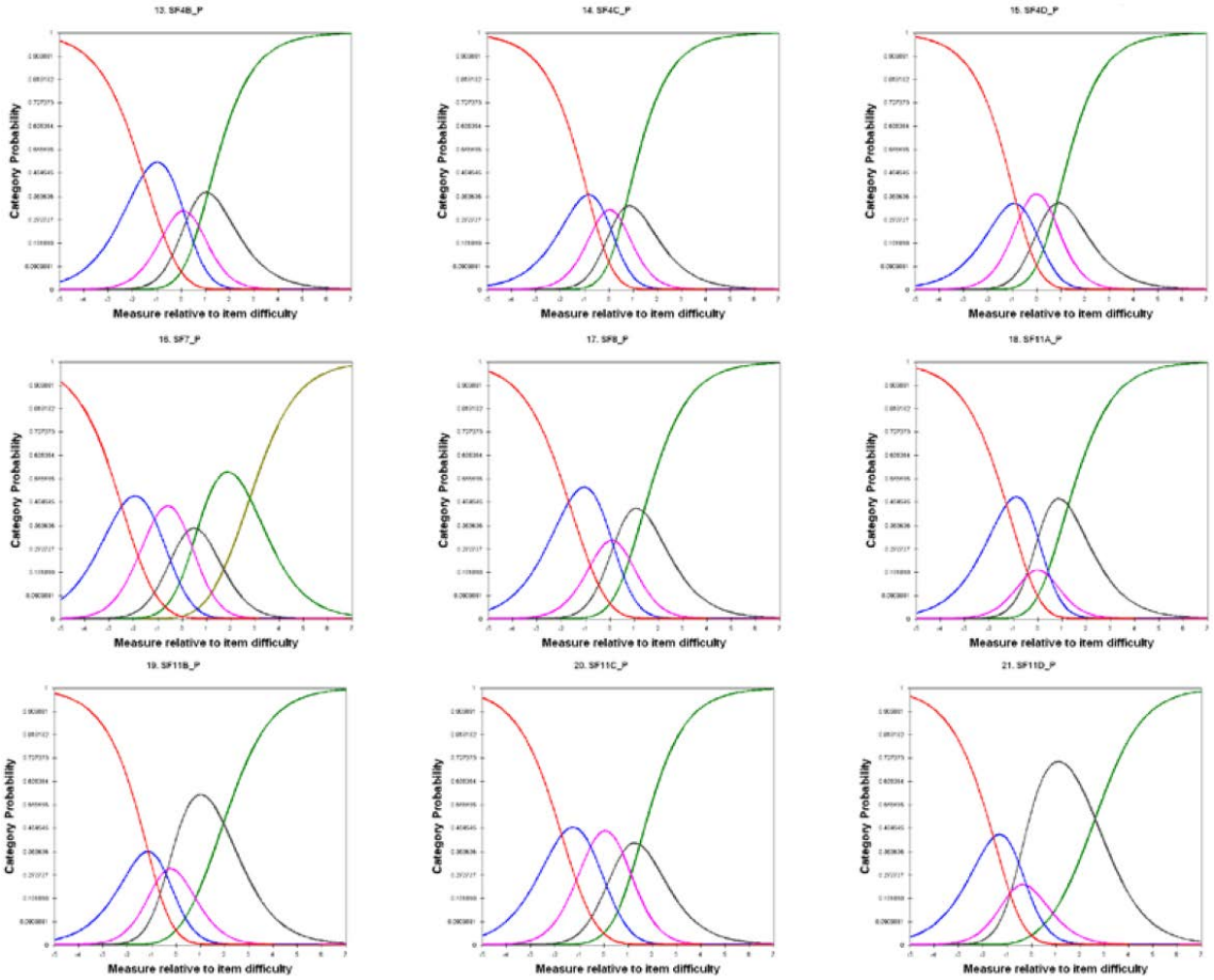

Figure 1. (continued) Category probability curves of the SF-36 HRQOL physical domain items

Note. Horizontal axes represent Rasch ability score. Vertical axes represent category probability

Table 5. Rating scale properties and decisions resulting from Rasch analysis of the SF-36 physical health domain (Items = 21)

Item	Response Scale	Regular Frequency Distribution	Average Advances w/Categories	Outfit MS < 2.0	Thresholds Ordered	Valid Criteria
SF1	5	Yes	Yes	Yes	Yes	4
SF3a	3	Yes	Yes	Yes	Yes	4
SF3b	3	Yes	Yes	Yes	**No**	3
SF3c	3	Yes	Yes	Yes	**No**	3
SF3d	3	Yes	Yes	Yes	Yes	4
SF3e	3	Yes	Yes	Yes	Yes	4
SF3f	3	Yes	Yes	Yes	Yes	4
SF3g	3	Yes	Yes	Yes	**No**	3
SF3h	3	Yes	Yes	Yes	**No**	3
SF3i	3	Yes	Yes	Yes	**No**	3
SF3j	3	Yes	Yes	**No**	**No**	2
SF4a	5	Yes	Yes	**No**	Yes	3
SF4b	5	Yes	Yes	Yes	Yes	4
SF4c	5	Yes	Yes	Yes	Yes	4
SF4d	5	Yes	Yes	Yes	Yes	4
SF7	6	Yes	Yes	Yes	Yes	4
SF8	5	Yes	Yes	**No**	Yes	3
SF11a	5	Yes	Yes	Yes	**No**	3
SF11b	5	Yes	Yes	Yes	Yes	4
SF11c	5	Yes	Yes	Yes	Yes	4
SF11d	5	Yes	Yes	Yes	**No**	3

Note. Seventy-three out of 84 (87.0%) total criteria met. Fifty out of 56 (89.3%) total criteria met after misfit items discarded.

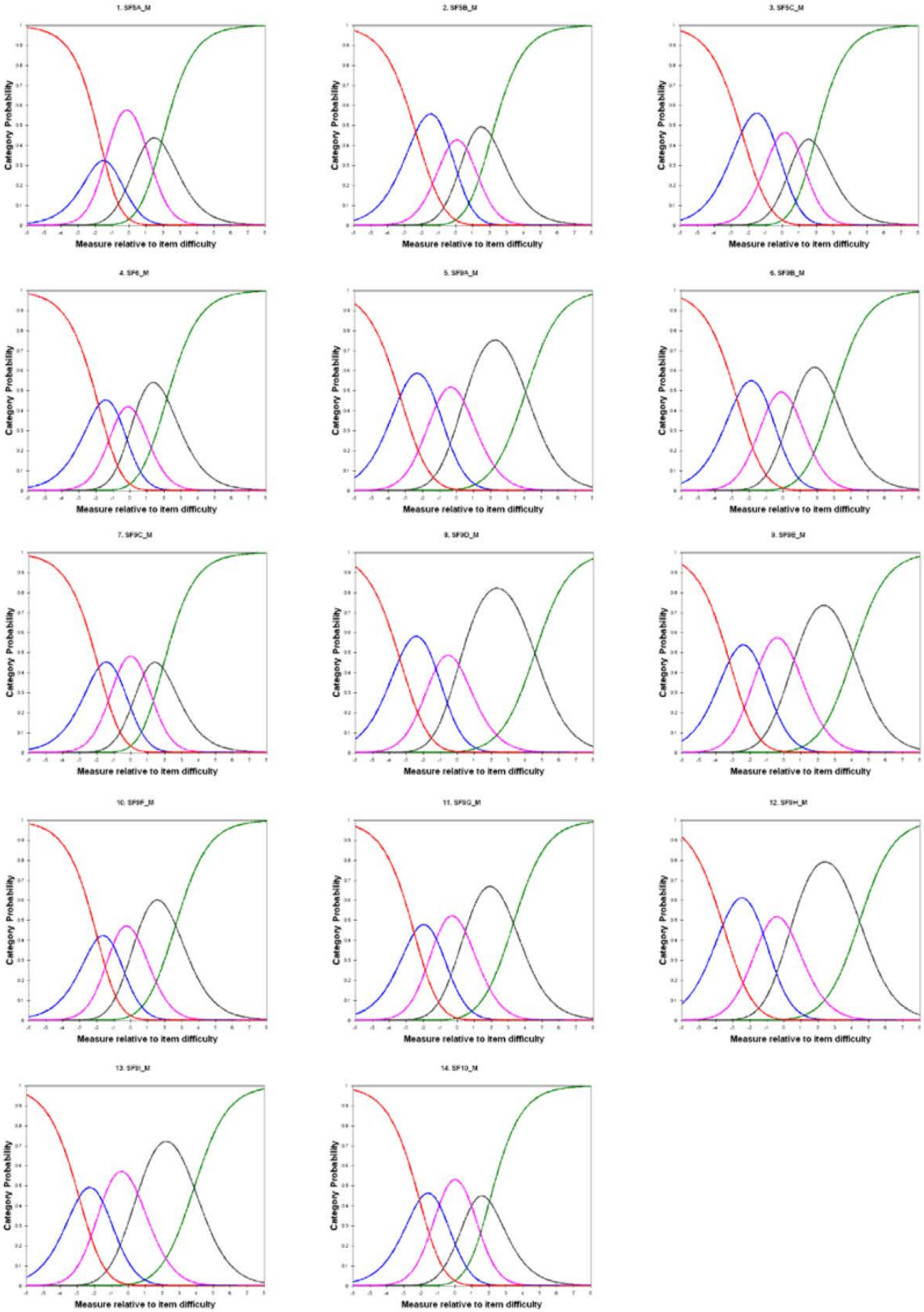

Figure 2. Category probability curves of the of the SF-36 HRQOL mental domain items

Note. Horizontal axes represent Rasch ability score. Vertical axes represent category probability

Table 6. Rating scale properties and decisions resulting from Rasch analysis of the SF-36 mental health domain (Items = 14)

Item	Response Scale	Regular Frequency Distribution	Average Advances w/Categories	Outfit MS < 2.0	Thresholds Ordered	Valid Criteria
SF5a	5	Yes	Yes	Yes	Yes	4
SF5b	5	Yes	Yes	Yes	Yes	4
SF5c	5	Yes	Yes	Yes	Yes	4
SF6	5	Yes	Yes	Yes	Yes	4
SF9a	5	Yes	Yes	Yes	Yes	4
SF9b	5	Yes	Yes	Yes	Yes	4
SF9c	5	Yes	Yes	Yes	Yes	4
SF9d	5	Yes	Yes	Yes	Yes	4
SF9e	5	Yes	Yes	Yes	Yes	4
SF9f	5	Yes	Yes	Yes	Yes	4
SF9g	5	Yes	Yes	Yes	Yes	4
SF9h	5	Yes	Yes	Yes	Yes	4
SF9i	5	Yes	Yes	Yes	Yes	4
SF10	5	Yes	Yes	Yes	Yes	4

Note. Fifty-six out of 56 (100%) total criteria met.

Table 7. Summary of Rasch calibration of the SF-36 physical health domain

Item	Response Scale	Calibration logits	SE logits	Infit MnSq	Outfit MnSq
SF1	5	0.82	0.07	0.98	0.97
SF3d	3	-0.21	0.09	0.95	0.85
SF3f	3	-0.5	0.1	1.05	1.1
SF3g	3	-0.56	0.1	0.91	0.75
SF3h	3	-0.89	0.12	1.04	1.24
SF4a	5	-0.52	0.07	0.93	0.9
SF4b	5	0.1	0.06	0.88	0.93
SF4c	5	-0.45	0.07	0.72	0.55
SF4d	5	-0.39	0.07	0.74	0.73
SF7	6	0.31	0.05	1.28	1.41
SF8	5	-0.54	0.07	0.88	0.82
SF11a	5	0.5	0.05	1.49	1.43
SF11b	5	1.1	0.05	1.18	1.15
SF11d	5	1.24	0.05	0.9	0.8

Note. Items 3a, 3b, 3c, 3e, 3j, 3i, & 11c were discarded due to misfit criteria.

Table 8. Summary of Rasch Calibration of the SF-36 mental health domain

Item	Response Scale	Calibration logits	SE logits	Infit MnSq	Outfit MnSq
SF5A	5	-0.9	0.06	1	0.91
SF5B	5	-0.39	0.06	0.91	0.82
SF5C	5	-1.1	0.06	1.07	0.88
SF6	5	-0.72	0.06	0.89	0.82
SF9A	5	0.47	0.06	0.86	0.85
SF9B	5	0.04	0.06	1.37	1.37
SF9C	5	-1.04	0.06	0.85	0.79
SF9D	5	1.06	0.06	0.92	0.95
SF9E	5	1.15	0.06	1.07	1.08
SF9F	5	-0.36	0.06	0.82	0.81
SF9G	5	1.13	0.06	1.25	1.24
SF9H	5	-0.09	0.07	0.85	0.88
SF9I	5	1.88	0.06	1.1	1.08
SF10	5	-1.13	0.06	1.03	0.90

Note. No items were discarded for mis-fitting.

In terms of model-data fit, the physical health domain data did not initially fit the Rasch model well. Table 7 displays the individual fit statistics for each item (mis-fit items not shown). Seven (3a, 3b, 3c, 3e, 3j, 3i, & 11c) out of the 21 items had fit statistics out of the acceptable range (i.e., 0.5 to 1.50). After the misfit items were discarded, a 14-item physical health domain fit the Rasch model well. The mean (SD) of the Infit and Outfit statistics were 1.00 (0.20) and 0.98 (0.30), respectively. The mental health domain did initially fit the Rasch model well. The mean

(SD) of the Infit and Outfit statistics were 1.00 (0.16) and 0.96 (0.17), respectively. Table 8 displays the individual fit statistics for each item in the mental health domain.

The item-person map for the adjusted physical health domain is shown in Figure 3. The leftmost vertical axis represents the logit scale where larger values signify better HRQOL. The pound signs (#) represent the distribution of person-level HRQOL relative to the logit scale. The rightmost side of the graph represents each item relative to its difficulty. The map shows that the distribution of items, with mean (SD) of 0.0 (0.67) was not well matched to the persons' HRQOL, with mean of 2.21 (1.61). The item locations indicate that the items are not targeting people of high HRQOL (> 1.5 logits) or low HRQOL (< -1.0 logits).

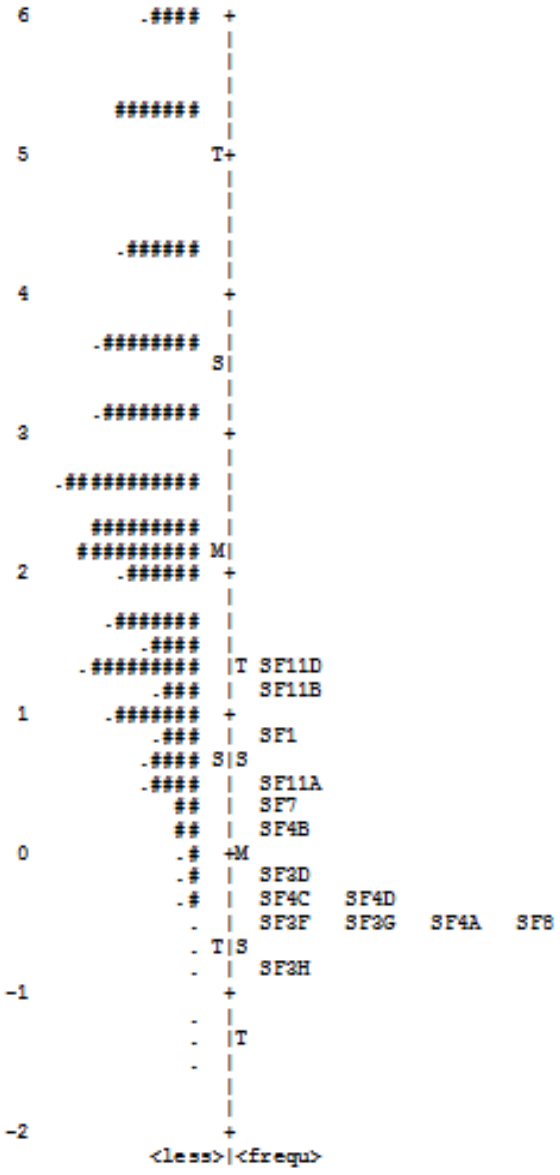

Figure 3. Person-item map of the SF-36 physical HRQOL domain

Note Each # represents 5 participants. M=mean. S=1 SD. T = 2 SD.

The item-person map for the mental health domain is shown in Figure 4. The map shows that the distribution of items with mean (SD) of 0.0 (0.95) was matched better to the persons' HRQOL with mean of 1.46 (1.77), than the physical health domain. The item locations also indicate that the items have better coverage across persons than the physical health domain, with coverage between -1.25 to 2.00 logits.

Item difficulty values resulting from the Rasch calibration are displayed in Table 7 and Table 8 for the physical and mental health domains, respectively. The larger an item's logit value is the higher the trait (HRQOL) must be for a person to endorse the item. Physical domain item difficulty ranged from -0.89 to 1.24 logits. The most difficult item was item 11d (How true is the following statement: My health is excellent.). The least difficult item was item 3h (Does your health now limit you in: Walking several hundred yards?). Mental domain item difficulty ranged from -1.13 to 1.88 logits. The most difficult item was item 9i (How much of the time during the past 4 weeks: Did you feel tired?). The least difficult item was item 10 (During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities [like visiting friends, relatives, etc.]?).

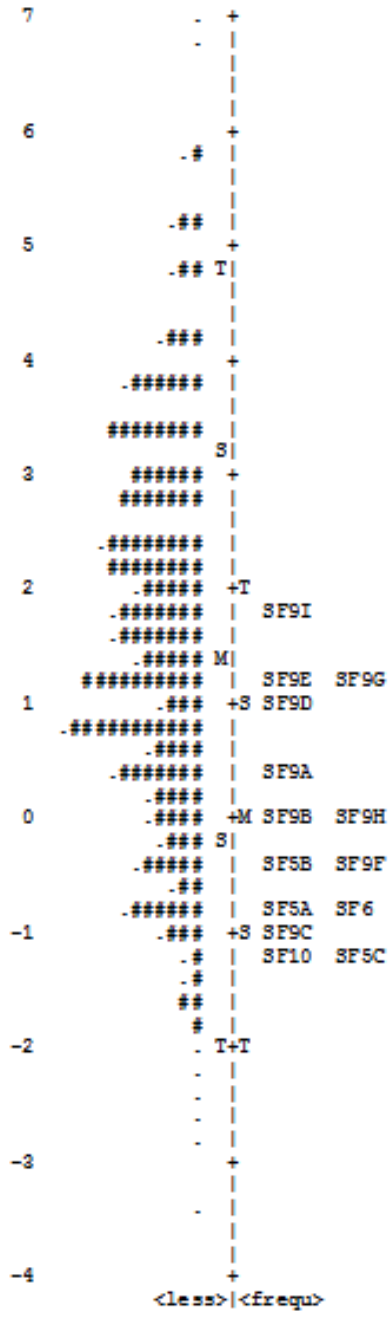

Figure 4. Person-item map of the SF-36 MENTAL HRQOL domain

Note Each # represents 5 participants. M=mean. S=1 SD. T = 2 SD.

Item separation and item separation reliability were examined next (see Table 9). The item separation index indicates how well the scale separates the items along the ability continuum. The item separation index was 8.81 and 15.26 for the final physical and mental health Rasch models, respectively. A high item separation index (2.0 or greater) indicates adequate discrimination by the items. The item separation reliability indicates the capability to replicate item placement within measurement error for another sample. The item separation reliability was. 99 and 1.00 for the final physical and mental health Rasch models, respectively. An item separation reliability close to 1.00 indicates a high degree of confidence for items [7].

Table 9. Model data fit statistics for each stage of the Rash analysis

						Person		Item			
Analysis	Domain	Items	%PVE	%IF	%PF	Separation	Reliability	Separation	Reliability	r_{RM}	Alpha
1	Physical	21	87.0	85.7	85.0	2.42	.85	8.03	.98	.87	.90
	Mental	14	100	100	84.0	3.55	.93	15.26	1.00	.96	.94
2	Physical	14	89.3	100	86.0	2.27	.84	8.81	.99	.88	.88

Note. Analysis #1 is the initial analysis with all items. Analysis #2 is after dropping misfit items. Alpha is Cronbach alpha and r_{RM} is the correlation between raw scores and person abilities (θ). %PVE is percent of validity criteria met for Rasch optimal category analysis. %IF is percent of items fitting the Rasch model. %PF is percent of person abilities fitting the Rasch model.

The persons' HRQOL was estimated for each domain during the Rasch calibration process in logits, where a higher logit value indicated a greater (positive) level of HRQOL. Table 10 displays descriptive statistics for the person-level HRQOL trait (θ). The mean (SD) level of persons' physical HRQOL was 2.21 (1.61). The range of persons' physical HRQOL was from -1.72 to 5.21, indicating a large spread of physical HRQOL. Participant HRQOL was consistent across gender and age. The overall person fit was examined by evaluating the percentage of persons with acceptable fit criteria. Of the total sample, 545 (86%) ability estimates were well fit. Person separation was 2.27, which indicates that people were well spread along the physical HRQOL continuum. The person separation reliability was .84, which indicates an acceptable degree of confidence in replicating person placement within measurement error. The mean (SD) level of persons' mental HRQOL was 1.46 (1.77). The range of persons' mental HRQOL was from -3.50 to 6.78, indicating a large spread of mental HRQOL. Of the total sample, 533 (84%) ability estimates were well fit. Person separation was 2.27, which indicates acceptable spread along the mental HRQOL continuum. The person separation reliability was .84, which was acceptable.

Table 10. Descriptive statistics for person HRQOL trait (θ) from Rasch analyses

	Mental HRQOL		Physical HRQOL	
	M	*SD*	*M*	*SD*
Overall	1.46	1.77	2.21	1.61
Gender				
Males	2.03	1.80	2.71	1.68
Females	1.24	1.69	2.00	1.53
Age (years)				
18-24	1.20	1.59	1.97	1.43
25-49	1.46	1.90	2.43	1.81
50-78	2.47	1.72	2.55	1.63

Note. HRQOL trait is measured in logits.

Table 11 contains results of the convergent validity evidence for the physical and mental HRQOL person scores resulting from the Rasch analyses. Overall, the physical health scores were moderately correlated (r_s = -.53) with the CDC's Healthy Days Index. Analysis by gender and age for the physical health scores showed similar results with correlations ranging from -.44 to -.59. For mental health, person scores were strongly correlated (r_s = -.73) with the CDC's Healthy Days Index. Analysis by gender and age for the mental health scores showed similar results with correlations ranging from -.66 to -.78.

Table 11. Bivariate Spearman correlations between person HRQOL trait (θ) from Rasch analyses and CDC Healthy Days Index

		Mental HRQOL	Physical HRQOL
	N	r_s	r_s
Overall	634	-.729	-.532
Gender			
Males	175	-.681	-.440
Females	458	-.724	-.527
Age (years)			
18-24	344	-.658	-.458
25-49	203	-.777	-.588
50-78	85	-.765	-.449

Note. All correlations were significant (*p*'s < .001). CDC Healthy Days Index represents number of unhealthy days.

Table 12 contains results of the construct validity evidence for the physical and mental HRQOL person scores. Dichotomous groups were compared (by ANCOVA) with known differences in HRQOL. For physical health, HRQOL person scores were significantly greater for those participants who were not obese, non-smokers, did not have an illness, did engage in vigorous activity, did engage in moderate activity, did engage in strength training exercises, did not have hypertension, did not have high blood cholesterol, and did not have diabetes (all p's < .01). For mental health, HRQOL person scores

were also significantly greater for those participants who were not obese, non-smokers, did not have an illness, did engage in vigorous activity, did engage in moderate activity, did engage in strength training exercises, did not have hypertension, did not have high blood cholesterol, and did not have diabetes (all p's < .01).

Table 12. Mean differences between known groups in person HRQOL trait (θ) from Rasch analyses

	Mental HRQOL			Physical HRQOL		
Health Status	*M*	*SD*	*p*	*M*	*SD*	*p*
Obesity						
Yes	1.16	1.89	.008	1.56	1.51	< .001
No	1.52	1.73		2.33	1.60	
Current smoker						
Yes	0.69	1.54	.002	1.46	1.34	.001
No	1.51	1.76		2.26	1.61	
Has an illness						
Yes	.43	1.48	< .001	0.67	1.07	< .001
No	1.52	1.76		2.32	1.58	
Vigorously active						
Yes	1.88	1.69	< .001	2.69	1.72	< .001
No	0.96	1.71		1.61	1.24	
Moderately active						
Yes	1.68	1.73	< .001	2.38	1.61	< .001
No	.49	1.56		1.40	1.33	
Strength trains						
Yes	1.71	1.73	< .001	2.49	1.68	< .001
No	1.06	1.75		1.76	1.39	
Hypertension						
Yes	1.15	2.05	< .001	1.96	1.48	.009
No	1.50	1.70		2.24	1.62	
High cholesterol						
Yes	1.32	1.92	< .001	1.92	1.50	< .001
No	1.46	1.73		2.23	1.62	
Diabetes						
Yes	0.32	1.45	.001	0.99	1.13	< .001
No	1.49	1.76		2.24	1.61	

Note. p-values are from age-adjusted analysis of covariance (ANCOVA).

4. Discussion

The purpose of this study was to separately evaluate the two HRQOL domains (physical and mental) of the SF-36 assessment using the Rasch model. The initial stages of the analysis evaluated the category functioning of each item. Using four criteria per item, it was found that majority of the total criteria were met for the physical health domain and all of the total criteria were met for the mental health domain. Despite the high percentage of validity evidence in the physical domain, eight items were flagged for disordered thresholds. The issue of ordered thresholds is an important item category characteristic. Order in the thresholds indicates that persons responding to higher levels (or lower levels) of a categorical scale in fact posses higher levels (or lower levels) of the trait being assessed. When thresholds are disordered, it is possible that some categories in the scale are unnecessary and/or redundant [16].

Six of the 8 disordered items came from the physical functioning section of the physical health domain. These items were 3b, 3c, 3g, 3h, 3i, and 3j. All physical functioning items share the same stem (Does your health now limit you in these activities?) and the same categorical scale: 1) Yes, limited a lot, 2) Yes, limited a little, and 3) No, not limited at all. One solution in this case may be to collapse the two "Yes" categories (i.e., 112) for each of these items. This would form a dichotomous item of 1) Yes, limited at least a little and 2) No, not limited at all.

The other two items with disordered thresholds came from the general health section of the physical health domain. These items were 11a and 11d. All general health items share the same stem (How true or false is each of the following statements for you?) and the same categorical scale: 1) Definitely true, 2) Mostly true, 3) Don't know, 4) Mostly false, and 5) Definitely false. One solution in this case may be to remove the "Don't know" category completely from the scale. This option could be explored by combining the "Don't know" category with the "Mostly true" category (i.e., 12234) or combining the "Don't know" category with the "Mostly false" category (i.e., 12334). The exploration of collapsing categories and re-running the Rasch model is a process that should be backed by a confirmatory stage (Linacre, 2002b) and is beyond the scope of this paper. This exploratory and confirmatory procedure is, however, needed and suggested for future research on the SF-36 HRQOL assessment.

Model-data fit was evaluated next and found that the mental health domain items adequately fit the Rasch

model. This provides evidence that the SF-36 assesses a unidimensional mental HRQOL domain. The physical health domain data, however, did not initially fit the Rasch model well. Seven (3a, 3b, 3c, 3e, 3j, 3i, & 11c) out of the 21 items had fit statistics out of the acceptable range. After the misfit items were discarded, a 14-item physical health domain fit the Rasch model well and provided evidence for a unidimensional physical HRQOL domain. Six of the 7 items deleted were from the physical functioning scale (3a, 3b, 3c, 3e, 3j, and 3i). These items assessed participant's limitations in vigorous activity, moderate activity, lifting or carrying groceries, climbing one flight of stairs, walking one hundred yards, and bathing or dressing yourself, respectively.

One of two factors might be the underlying cause of these mis-fitted items. One factor is the 3-point scale previously mentioned regarding the physical functioning section of the SF-36. Four of the 6 mis-fitted physical functioning items were also flagged for having disordered thresholds. This type of category dysfunction is likely to explain the item's mis-fitting the Rasch model [7]. The other factor is the possibility that the mis-fitted items are not of the same unidimensional construct as the other items. However, since these items concerning limitations in movement-related activities are of similar nature to other well-fitted items (i.e., Climbing several flights of stairs, Walking more than a mile, etc.), it is more likely they are mis-fitting due to improper category functioning.

The item-person maps for both the physical and mental health domains showed that the item locations were not well matched to persons of very high (very good) HRQOL or very low (very poor) HRQOL. The items were targeted well to persons of moderately poor to moderately good HRQOL. In other words, the items were too easy for the many of the participants in both domains. The most difficult physical domain item was item 11d (How true is the following statement: My health is excellent.) followed by item 11b (How true is the following statement: I am as healthy as anybody I know.). The least difficult physical domain item was item 3h (Does your health now limit you in: Walking several hundred yards?) followed by item 8 (During the past 4 weeks, how much did pain interfere with your normal work [including both work outside the home and housework]?). Discrimination among persons of better physical HRQOL may be increased by the addition of more difficult items [7].

The most difficult mental domain item was item 9i (How much of the time during the past 4 weeks: Did you feel tired?) followed by item 9e (How much of the time during the past 4 weeks: Did you have a lot of energy?). The least difficult mental domain item was item 10 (During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities [like visiting friends, relatives, etc.]?) followed by item 5c (how much of the time have you had any of the following problems… as a result of any emotional problems: Did work or activities less carefully than usual.). As well, discrimination among persons of better mental HRQOL may be increased by the addition of more difficult items. This should be explored in future research on the SF-36 HRQOL assessment.

A final stage of the present study was to show validity evidence for the persons' HRQOL scores resulting from the Rasch analyses. Results showed that the physical health Rasch scores and the CDC's Healthy Days Index were moderately correlated with each other. This provides convergent validity evidence in that both measures theoretically attempt to assess the same construct [19]. For the mental health domain, the Rasch scores and the CDC's Healthy Days Index were moderately to strongly correlated with each other. Overall, the convergent validity evidence provided substantial confirmation for the use of the Rasch person scores as a measure of HRQOL.

Construct validity evidence was also tested in this study by showing differences in Rasch HRQOL person scores between groups of participants with known differences. Results showed significantly greater physical and mental HRQOL scores for participants who were not obese compared to participants who were obese. This relationship has been confirmed before in large scale studies showing decreased physical HRQOL as well as mental HRQOL among obese adults compared to normal weight adults [20]. The same relationship was also found in a large population-level study using the CDC's Healthy Days HRQOL core [21]. Significantly greater physical and mental HRQOL scores were also seen for participants who did not smoke compared to participants who did smoke. This relationship has also been shown in national data with smokers who made no attempts to quit having significantly lower mental and physical HRQOL [22].

Our finding of greater physical and mental HRQOL among participants with some form of illness has also been confirmed before by others showing that adults with some form of chronic illness were significantly more likely to report lower levels of HRQOL [23]. Significantly greater physical and mental HRQOL scores were also seen for participants who engaged in various physical activities as compared to participants who did not engage in those activities. This relationship has also been shown in quality of life research where adults who engaged in physical activity were less likely to report poor HRQOL than their non-active counterparts [24]. Finally, this study showed that participants who reported having hypertension, high blood cholesterol, or diabetes had significantly lower physical and mental HRQOL compared to those participants who did not report those health problems. These findings have also been confirmed before [25].

This study has many strengths worth mentioning. The large sample size was useful and necessary for proper Rasch parameter estimates and fit statistics. Samples of size 200 and greater are suggested for proper estimation [26]. Another strength of this study is the use of the partial credit Rasch model to allow for the evaluation of proper category functioning per item [7]. This was essential for the SF-36 HRQOL assessment because the instrument contains 35 items (21 for physical health and 14 for mental health) with three different categorical scales (3-point, 5-point, and 6-point). A final strength in this study was the administration of the CDC's Healthy Days HRQOL core to the same sample of participants for its use in validating the Rasch person HRQOL scores.

A limitation of this study was the use of the general population as a sampling frame. The SF-36, like many HRQOL assessments, is often used to differentiate perceived health among people suffering from disease states [27]. It was found in the current study that, for a general sample of adults, the SF-36 items were too easy (ceiling effect). However, it is useful for researchers to

know that when administering the SF-36 to a general sample of adults, the assessment may not be useful in effectively separating those individuals in terms of HRQOL.

5. Conclusion

In conclusion, a Rasch partial credit model was used to analyze the two dominant HRQOL domains (physical and mental) of the SF-36 assessment. The majority of the total criteria used for optimal category functioning were met for the physical health domain and all of the total criteria were met for the mental health domain. Eight items were flagged for disordered thresholds, of which 6 items came from the physical functioning subscale. Seven physical health items had fit statistics out of the acceptable range and were dropped from the final Rasch analysis. It is suggested that exploratory and confirmatory re-categorization of the 8 identified items be investigated. Also, adding more difficult items to the SF-36 should be investigated to help target healthier individuals. Finally, both convergent and construct validity evidence provided substantial confirmation for the use of the Rasch physical and mental health person scores as measures of HRQOL.

References

[1] Heath, G. W., & Brown, D. W. (2009). Recommended levels of physical activity and health-related quality of life among overweight and obese adults in the United States, 2005. *Journal of Physical Activity and Health, 6*(4), 403-411.

[2] Wilson, I. B., & Cleary, P. D. (1995). Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *Journal of the American Medical Association, 273*(1), 59-65.

[3] Centers for Disease Control and Prevention. *Measuring healthy days: Population assessment of health-related quality of life*. Centers for Disease Control and Prevention, Atlanta, Georgia 2000.

[4] Dominick, K. L., Ahern, F. M., Gold, C. H., & Heller, D. A. (2002). Relationship of health-related quality to health care utilization and mortality among older adults. *Aging Clinical and Experimental Research, 14*(6), 499-508.

[5] Embretson, S. E., & Steven, P. Reise. 2000. Item response theory for psychologists.

[6] Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.

[7] Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model : fundamental measurement in* the human sciences (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

[8] Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Medical care, 38(9 Suppl), II28.

[9] Ware, J.E., Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30*(6), 473.

[10] Ware, J.E. (2004). SF-36 Health Survey Update. Retrieved August 15, 2011, from http://www.sf-36.org/announcements/Updated_SF36_bookChapter_Sept04.pdf.

[11] QualityMetric. (2011). SF Health Surveys. Retrieved August 15, 2011, from http://www.qualitymetric.com/WhatWeDo/GenericHealthSurveys/tabid/184/Default.aspx.

[12] Mielenz, T., Jackson, E., Currey, S., DeVellis, R., & Callahan, L. F. (2006). Psychometric properties of the Centers for Disease Control and Prevention Health-Related Quality of Life (CDC HRQOL) items in adults with arthritis. *Health and Quality of Life Outcomes, 24*(4), 66.

[13] Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. Rasch measurement transactions, 8(3), 370.

[14] Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 103–122.

[15] Linacre, J. M. (2002a). What do Infit and Outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

[16] Linacre, J. M. (2002b). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

[17] Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity of examination scores. *Applied Measurement in Education, 3*(4), 331-345.

[18] Linacre, J. (2006). *A user's guide to Winsteps: Rasch model computer programs*. Chicago: Winsteps.

[19] Hennessy, C. H., Moriarty, D. G., Zack, M. M., Scherr, P. A., & Brackbill, R. (1994). Measuring health-related quality of life for public health surveillance. *Public Health Reports, 109*(5), 665-672.

[20] Jia, H., & Lubetkin, E. I. (2005). The impact of obesity on health-related quality-of-life in the general adult US population. *Journal of Public Health (Oxf), 27*(2), 156-164.

[21] Hassan, M. K., Joshi, A. V., Madhavan, S. S., & Amonkar, M. M. (2003). Obesity and health-related quality of life: a cross-sectional analysis of the US population. *International Journal of Obesity and Related Metabolic Disorders, 27*(10), 1227-1232.

[22] McClave, A. K., Dube, S. R., Strine, T. W., & Mokdad, A. H. (2009). Associations between health-related quality of life and smoking status among a large sample of U.S. adults. *Preventive Medicine, 48*(2), 173-179.

[23] Strine, T. W., Chapman, D. P., Balluz, L. S., Moriarty, D. G., & Mokdad, A. H. (2008). The associations between life satisfaction and health-related quality of life, chronic illness, and health behaviors among U.S. community-dwelling adults. *Journal of Community Health, 33*(1), 40-50.

[24] Brown, D. W., Balluz, L. S., Heath, G. W., Moriarty, D. G., Ford, E. S., Giles, W. H., & Mokdad, A. H. (2003). Associations between recommended levels of physical activity and health-related quality of life. Findings from the 2001 Behavioral Risk Factor Surveillance System (BRFSS) survey. *Preventive Medicine, 37*(5), 520-528.

[25] Hayes, D. K., Denny, C. H., Keenan, N. L., Croft, J. B., & Greenlund, K. J. (2008). Health-related quality of life and hypertension status, awareness, treatment, and control: National Health and Nutrition Examination Survey, 2001-2004. *Journal of Hypertension, 26*(4), 641-647.

[26] Kline, T. (2005). *Psychological testing : a practical approach to design and evaluation*. Thousand Oaks, Calif.: Sage Publications.

[27] Ware, J. E., Jr. (2000). SF-36 health survey update. *Spine (Phila Pa 1976), 25*(24), 3130-3139.