Qual Life Res (2013) 22:2359–2369
DOI 10.1007/s11136-013-0373-1

Measurement bias of the SF-36 Health Survey in older adults with chronic conditions

Hongdao Meng · Bellinda L. King-Kallimanis · Amber Gum · Brenda Wamsley

Accepted: 13 February 2013 / Published online: 6 March 2013

Abstract
Purpose The objectives of this study were to investigate the psychometric properties of the SF-36 in a sample of older adults with chronic conditions and to test whether measurement bias exists based on the levels of comorbidity.
Methods Participants included were 979 cognitively intact older adults with comorbidities who were interviewed at their homes. We examined the psychometric properties of the SF-36 and conducted confirmatory factor analysis (CFA) to investigate the assumption of measurement invariance by the levels of comorbidity.
Results Overall data quality was high and scaling assumptions were generally met with few exceptions. Floor and ceiling effects were present for the role-physical and role-emotional subscales. Using CFA, we found that a three-factor measurement model fits the data well. We identified two violations of measurement invariance. Results showed that participants with high comorbidity level place more emphasis on social functioning (SF) and bodily pain (BP) in relation to physical health-related quality of life (HRQoL) than those with low comorbidity level.
Conclusions Measurement bias was present for the SF and BP components of the SF-36 physical HRQoL measure. Researchers should be cautious when considering the use of SF-36 in clinical studies among older adults with comorbidities.

Keywords Psychometric evaluation · Quality of life · Confirmatory factor analysis · Comorbidity

H. Meng (✉)
University of South Florida, Tampa, FL, USA
e-mail: meng@usf.edu

B. L. King-Kallimanis
Department of Medical Gerontology, Trinity College Dublin, Dublin, Ireland

A. Gum
Department of Mental Health Law and Policy, University of South Florida, Tampa, FL, USA

B. Wamsley
Department of Social Work, West Virginia State University, Dunbar, WV, USA

Introduction

While recent development in the measurement of health outcomes has generated significant interest in adapting new and flexible approaches to measure patient-reported health-related quality of life (HRQoL) [1], the 36-item Medical Outcomes Study Short Form Health Survey (SF-36) has continued to be widely used [2]. SF-36 was originally derived from a 149-item questionnaire measuring the multi-dimensional concept of health status related to physical, psychological, and social well-being among the general adult population [3]. While the SF-36 has been shown to have satisfactory psychometric properties in the general older population [4, 5], its validity and reliability have not been established in older adults with chronic conditions. Recent research has demonstrated that the negative impact of comorbid chronic conditions on HRQoL varies across disease and comorbidity status [6]. In addition, there is also evidence suggesting that the SF-36 may exhibit differential item functioning (DIF), where respondents with comorbidities (hypertension, diabetes, and respiratory diseases) subgroups selectively endorse certain items differently conditional on having the same

underlying latent trait. As a result, DIF may alter the effect of exogenous variables such as race/ethnicity and education [7].

The population of older adults with comorbidities has become an increasingly important focus for health outcomes researchers and policymakers because of the increases in the prevalence of chronic conditions with population aging, the complexity of medical and/or supportive care needs, as well as substantial public/private resources needed in providing such care. For example, research has shown that 83 % of Medicare beneficiaries have at least one chronic condition, and that 23 % of Medicare beneficiaries with five or more chronic conditions account for 68 % of the program's spending [8]. Therefore, understanding the measurement properties of the SF-36 in the population of older adults with comorbidities is especially important for intervention studies designed to improve HRQoL in this population. In addition, valid and reliable measures of patient-reported outcomes in the Medicare population are essential to determining the effectiveness of medical and other types of interventions in comparative effectiveness and patient-centered outcomes research. Specifically, it is important to ensure that outcome measures based on instruments such as the SF-36 remain reliable and valid for use among older adults with comorbid chronic medical conditions.

Previous research has suggested that there are a number of potential issues related to the applicability of SF-36 to community-living older adults in terms of the mode of administration and its sensitivity to change over time [4, 9, 10]. However, evidence on the potential measurement bias of the SF-36 among older adults with chronic conditions is very limited and inconclusive. For example, when used in a self-reported format in general practice, three early studies found that the measure resulted in substantial missing data and therefore recommended that interviewer-administered SF-36 would be better suited for older adults [4, 11, 12]. In addition, when used in disease-specific populations, studies suggested that SF-36 is not suitable for older patients with stroke or Parkinson's disease [13, 14]. Similarly, Stadnyk et al. [15] have found that the SF-36 lacked sufficient internal consistency and test–retest reliability for clinical use among a sample of frail older adults. On the other hand, a large survey study conducted in a community-dwelling older adult population in the United Kingdom have found that the SF-36 yielded good response rate (82 %) and completion rate, as well as good internal consistency for most subscales (except for social functioning) [16]. More recently, Yu et al. [7] found evidence that DIF could lead to different results for individuals with chronic conditions using data obtained from a large integrated health system. Therefore, as the aging of the US population continues, it is important to ensure that the widely used SF-36 is suitable to measure and differentiate HRQoL among older adults with comorbidities.

There are a number of reasons why the use of SF-36 in older adults with comorbidities may be more susceptible to measurement bias. First, SF-36 is a self-reported measure of HRQoL; therefore, items may be interpreted using different frames of reference than would be used by a younger population [4]. In addition, older adults with comorbidities may have undergone a response shift where they may have re-conceptualized the importance of certain domains of the SF-36 [17]. If this is the case, the relationships between the observed items and the SF-36 will not remain constant across subgroups of individuals with different levels of comorbidities. As a result, the assumptions of measurement invariance may be violated, resulting in measurement bias and preventing the researchers from drawing valid conclusions from the data. In the context of comparative effectiveness research, testing for the assumption of measurement invariance and measurement bias is especially important for a high risk population (such as the older adults with comorbidities) to ensure that valid information is available to policymakers when determining the comparative effectiveness of various interventions on HRQoL when targeting older adults with comorbid chronic conditions.

The present study assessed the following psychometric properties of the SF-36 in a sample of older Medicare beneficiaries with chronic conditions: data quality, scaling assumptions, and reliability and validity according to the methods described by the developers of the scale [18]. We then examined the factorial validity based on the hypothesized factor structure, as well as testing for measurement invariance by levels of comorbidity.

Methods

Data

Data came from the baseline interview of the Medicare Primary and Consumer-Directed Care (PCDC) demonstration, which has been described in detail elsewhere [19]. In brief, the PCDC demonstration was a two-year community-based randomized controlled trial conducted in a convenience sample in New York, West Virginia, and Ohio. Medicare beneficiaries who were recruited through their primary care physicians and met the following criteria: (1) enrollment in Medicare Parts A and B; (2) needing or receiving help for at least two activities of daily living (ADLs) or at least three instrumental activities of daily living (IADLs); and (3) recent significant health care utilization (had been a patient in a hospital, nursing home, or Medicare home health care agency within the past 12 months, or had two or more emergency room visits in

the past 6 months). The exclusion criteria included: living in a nursing home, receipt of Medicare hospice or end-stage renal disease benefits, or enrollment in an HMO or a state Medicaid home and community-based waiver program. A total of 1,605 Medicare beneficiaries from 19 counties in upstate New York and the West Virginia–Ohio border area entered the study between June 1998 and June 2000. The study protocol was approved by University of Rochester Research Subjects Review Board. For the purpose of the present study, we excluded participants who were younger than 65 ($n = 164$), participants who provided responses via proxies ($n = 256$), and participants who were cognitively impaired ($n = 197$) based on the cognitive performance scale [20], and finally, participants without comorbidities ($n = 9$). As a result, baseline assessments of 979 individuals were included in the study.

Measurements

SF-36 Health Survey

The SF-36 Health Survey is designed to measure HRQoL [3]. It consists of eight multi-item (range from 2 to 10 items) domains: physical functioning (PF), role limitations due to physical health (role-physical, RP), bodily pain (BP), general health perceptions (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (role-emotional, RE), and mental health (MH) [3]. Responses are on 2-, 3-, 5-, or 6-point scales, and nine items require reverse coding so that all items are scored in the same direction. After simple summation of all individual items within a scale, transforming, and averaging, each scale ranges from 0 (worst health) to 100 (best health). Finally, scores from the eight domains can be converted into two summary measures, the physical component summary (PCS) and the mental component summary (MCS). The Medicare demonstration used the standard SF-36 with a four-week recall period.

Covariates

Participants were interviewed in their own homes by trained interviewers at baseline. All interviewers were trained and completed two reliability assessments of videotaped standardized interview sessions. Information on individual socio-demographics, health and disability status, self-reported health and physician diagnoses were collected. Socio-demographic variables include age (in years), gender, ethnicity (white versus others), living arrangement (whether the participant lived alone), informal caregiver status (whether a caregiver had been identified), health insurance status (Medicaid), and rural status. Baseline health and disability status variables include number of activities of daily living (ADL), instrumental activities of daily living (IADL) and self-rated health. For the purposes of the confirmatory factor analysis (CFA), levels of chronic medical conditions were classified as either low (1–3 conditions) or high (4 or more conditions) based on the sample median of 4. Physician-diagnosed chronic medical conditions included the following: arthritis, hypertension, angina, heart disease related to valves/rhythm, sciatica, chronic obstructive pulmonary disease (COPD), congestive heart failure,diabetes, myocardial infarction, stroke, and cancer.

Analysis

We evaluated the following five criteria as recommended by Ware and Gandek [18]: data quality, scaling properties and score distributions, item internal consistency, item discriminant validity, internal consistency reliability, and construct validity. Data quality included the extent of missing and out-of-range data. Poor data quality is commonly associated with item wording, format, and respondent understanding. Missing data were examined by summarizing the percentage of patients missing for each item.

We tested the following assumptions: equal item variance, equality of item-scale correlations, item internal consistency, and item discriminant validity. Equal variance refers to the assumption that items within the same scale should have similar standard deviations. Equal item-scale correlation means that the correlation between each item and its scale is hypothesized to be similar across all items because each of the items should be measuring a similar proportion of information about the concept being measured. Correction for overlap was performed to ensure that the item-scale correlation is not inflated [21]. Internal consistency reliability indicates whether items within a scale are measuring the same concept. For each of the eight scales, Cronbach's alpha was calculated with a threshold of 0.7 or above as the criterion of acceptable reliability [22]. Item discriminant validity suggests that the correlation of each item with its hypothesized scale should be higher than its correlation with non-hypothesized items. An item-scale correlation of two standard errors above the item-other scale correlation was considered as evidence of discriminate validity [23].

To test the factor structure of the SF-36, we fit two frequently used structures using confirmatory factor analysis (CFA) with maximum likelihood [24]. In the first model (Model 1), HRQoL was represented by two latent variables, physical HRQoL (PHYS HRQoL: PF, RP, BP, and GH) and mental HRQoL (MENT HRQoL: VT, SF, RE, and MH). The second model (Model 2) included three latent variables: PHYS HRQoL (PF, RP, and BP); MENT HRQoL (SF, RE, and MH); and General Well-Being (GEN WB), with GH associated with both PHYS HRQoL and GEN WB, VT associated with both GEN WB and MENT HRQoL. We used the chi-square test of exact fit, the root mean square error of

approximation (RMSEA), and the expected cross-validation index (ECVI) to assess goodness of fit. A RMSEA of less than or equal to .08 is considered satisfactory fit and a RMSEA of less than or equal to .05 this suggests good fit [25].

Once a satisfactory fitting measurement model has been identified, CFA was used to investigate measurement bias [26]. We fit our model to the two 8 × 8 variance/covariance matrices of the eight HRQoL scales based on the level of comorbidity, that is, we specified one matrix for the low comorbidity group and another for the high comorbidity group. We then constrain the factor loadings and intercepts for the low and high comorbidity groups to be equal and test for measurement bias. If the fit of the constrained model significantly deteriorates compared to that of the unconstrained model, then we conclude that at least one of the equality constraints is not tenable and therefore measurement bias is present. When assessing the absence or presence of measurement bias with respect to the levels of comorbidity, we used a series of models in which the equality constraints associated with each of the observed variables were removed one at a time, with global model fit tests conducted to assess the impact of their removal on model fit based on the following: (a) the chi-square difference test, (b) observed parameter changes (OPC), and (c) ECVI difference test [27]. The chi-square difference test evaluates whether the alternative model fit (with the equality constraints for one observed variable removed) is significantly better than the null model. We used a Bonferroni correction to account for multiple comparisons [28]. The OPCs allowed us to assess whether model parameters being tested changed between the fully constrained model and the standardized parameters of the alternative model. Finally, the ECVI difference test was used to examine the difference in the ECVI values of the null and alternative model (ECVI difference is considered significant if the 90 % confidence interval does not include zero).

Results

Sample characteristics

A total of 979 individuals were included in the study sample. The mean age of the sample was 79.2 years (standard deviation 7.4, range 65–100). The majority of the sample was female (74.2 %). More than two-thirds (69.8 %) of the sample had 12 or fewer years of schooling. About a quarter of the sample had annual household income of $10,000 or less. Participants reported an average of 1.9 ADL dependencies and 2.9 IADL dependencies (both were 0–6 scales, with 0 signifying no dependence). The mean number of chronic conditions reported was 4.1 (out of ten, standard deviation 1.9) and 53.9 % reported fair or poor self-rated health. Table 1 shows the baseline individual characteristics for the study sample.

Table 1 Baseline characteristics of the sample, $n = 979$

Variable	Mean
Socio-demographics	
Age (mean ± SD)	79.2 ± 7.4
Female (%)	74.2
Minority (%)	2.8
Married (%)	38.8
Education (%)	
<High school	38.6
High school	31.4
>High school	30.0
Income (%)	
<$10,000	25.6
$10,000–$19,999	38.6
>$20,000	35.8
Live alone (%)	45.4
Has informal caregiver (%)	66.0
Medicaid (%)	8.6
Rural (%)	29.4
Health and functional status	
ADL (mean ± SD)	1.9 ± 1.5
IADL (mean ± SD)	2.9 ± 1.6
# of chronic conditions (mean ± SD)	4.7 ± 2.1
1	7.4
2	14.2
3	19.3
4	20.3
5	16.1
6	11.9
7+	10.8
Fair/poor self-rated health (%)	53.9

Chronic conditions include the following (in descending order of prevalence): arthritis, hypertension, angina, heart disease related to valves/rhythm, sciatica, chronic obstructive pulmonary disease (COPD), congestive heart failure, diabetes, myocardial infarction, stroke, and cancer

Data quality

The overall proportion of missing data is quite small, with 89.7 % of respondents had complete data on all 36 items, 7.9 % had missing values on one item, and only 2.4 % had missing values on more than one item. Table 2 shows the item-level percent of missing data, as well as descriptive information on the eight subscales. Missing item data ranged from 0 to 2.6 % (median 0.2 %). The General Health items had the most missing data and the question "I expect my health to get worse" had the highest percentage of missing data (GH2: 1.2 %, GH3: 1.4 %, GH4: 2.6 %).

Table 2 Item percent missing, item means, and standard deviations (SD)

Variable	% missing	Mean	SD	Range	% floor	% ceiling
Physical functioning (PF)		23.59	21.25	0–95	10.8	0
Vigorous activities (PF1)	0	1.09	0.38			
Moderate activities (PF2)	0	1.36	0.62			
Lift or carry groceries (PF3)	0.4	1.45	0.66			
Climb several flights of stairs (PF4)	0.5	1.24	0.53			
Climb one flight of stairs (PF5)	0.2	1.63	0.74			
Bend, kneel, or stoop (PF6)	0.1	1.50	0.68			
Walk more than a mile (PF7)	0.5	1.14	0.44			
Walk several blocks (PF8)	0.2	1.30	0.60			
Walk one block (PF9)	0.2	1.71	0.79			
Bathe or dress (PF10)	0	2.30	0.72			
Role-physical (RP)		23.73	33.66	0–100	55.5	11.0
Cut down time spent on work (RP1)	0.3	1.41	0.49			
Accomplish less (RP2)	0.3	1.17	0.37			
Limited in kind of work (RP3)	0.1	1.16	0.36			
Difficulty performing work (RP4)	0.2	1.22	0.41			
Bodily pain (BP)		49.25	26.51	0–100	1.7	10.3
Pain severity (BP1)	0	3.67	1.34			
Pain interfered with work (BP2)	0.1	2.66	1.34			
General health (GH)		45.57	21.10	0–100	0.9	0.4
Rating of general health (GH1)	0	3.58	0.93			
Get sick easier (GH2)	1.2	3.73	1.20			
As healthy as others (GH3)	1.4	3.33	1.34			
Health to get worse (GH4)	2.6	3.01	1.22			
Health excellent (GH5)	0.3	3.88	1.22			
Vitality (VT)		37.67	17.59	10–90	0	0
Full of pep (VT1)	0.4	3.94	1.07			
Lot of energy (VT2)	0.1	3.94	1.07			
Worn out (VT3)	0.1	2.86	1.19			
Tired (VT4)	0.3	2.55	1.10			
Social functioning (SF)		61.08	32.31	0–100	5.0	26.8
Social—extent (SF1)	0.6	2.53	1.38			
Social—frequency (SF2)	0.5	3.42	1.43			
Role-emotional (RE)		71.60	40.43	0–100	19.3	62.7
Cut down time (RE1)	0.6	1.74	0.44			
Accomplish less (RE2)	0.5	1.66	0.47			
Work not done as careful (RE3)	1.0	1.74	0.44			
Mental health (MH)		62.87	16.08	8–88	0	0
Nervous (MH1)	0.3	3.74	1.21			
Down in dumps (MH2)	0.2	4.19	1.06			
Calm and peaceful (MH3)	0.2	2.66	1.00			
Downhearted and blue (MH4)	0	3.84	1.06			
Happy (MH5)	0.1	2.40	0.97			

Hypothesized patterns of differences in item means were observed (Table 2). Within the PF scale, the most difficult item (vigorous activities) had the lowest mean (1.09) and the easiest item (bathe, dress) had the highest mean (2.30). Item means increased as item difficulty decreased across groups of PF items (PF4-PF5, PF7-PF9). As expected, MH items measuring positive affect (MH3 and MH5) had lower mean values than items measuring negative affect (MH1, MH2, and MH4). However, VT items that measured energy (VT1-VT2) had higher mean values than items measuring fatigue (VT3-VT4). Item standard deviations were roughly equivalent within scales, except for PF1

(vigorous activities), PF7 (walk more than a mile), GH1 (rating of general health), MH5 (happy). As hypothesized, floor and ceiling effects (Table 2) were generally low for the three bipolar scales (GH, VT, MH), ranging from 0 to 0.9 % in the total sample. Floor and ceiling effects were minimum to modest for the PF and BP scales (floor = 10.8 and 1.7 %, ceiling = 0 and 10.3 %, respectively). The RP and RE scales had substantial floor and ceiling effects (floor = 55.5 and 19.3 %, ceiling = 11.0 and 62.7 %, respectively). The SF scale had a small floor effect (5.0 %) and medium ceiling effect (26.8 %).

Scaling properties and score distributions

Figure 1 illustrates results from the scaling assumption tests. Standard deviations of items within a scale were similar for RP, BP, SF, and RE. All item-scale correlations were greater than 0.40 (horizontal line) except for PF1 (0.17), GH2 (0.36), and GH4 (0.34), as shown in the figure by three red letters (P, G, G) below the line at 0.4. The distances between hypothesized item-scales correlations and the non-hypothesized item-scales correlations were 0.02 for the PF1-RP pair and 0.04 for the GH2-MH pair, indicating scaling failure on PF1 and GH2 (Fig. 1).

Figure 2 shows the scores of the eight subscales as well as the two summary measures for the study sample as compared to the US norm. The results were based on a linear transformation such that a value of 50 is the US general population mean (in 1998) and 10 is the standard deviation. Norm-based scoring of all eight scales and the summary measures have the advantage of easily facilitating the interpretation of results across measures [29]. The study

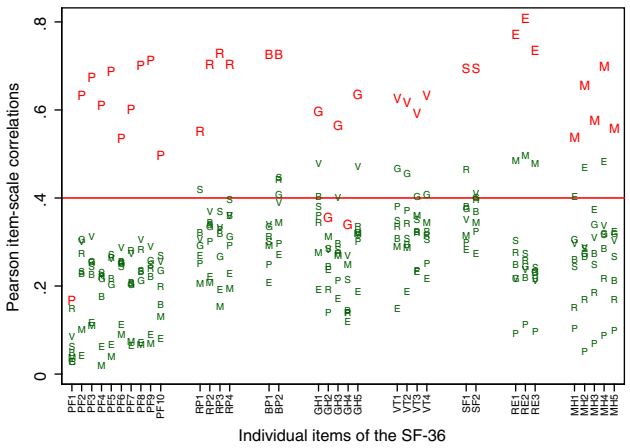

Fig. 1 SF-36 item-scale correlations. *Note* The *horizontal* axis shows the individual items; the *vertical* axis shows item-scale correlations. Correlations are labeled with *letters* indicating different scales (*P* = PF, *R* = RP, *B* = BP, *G* = GH, *V* = VT, *S* = SF, *E* = RE, *M* = MH). Pearson correlation coefficients are plotted in *large font in red* for hypothesized scales and in *small font in black* for non-hypothesized scales

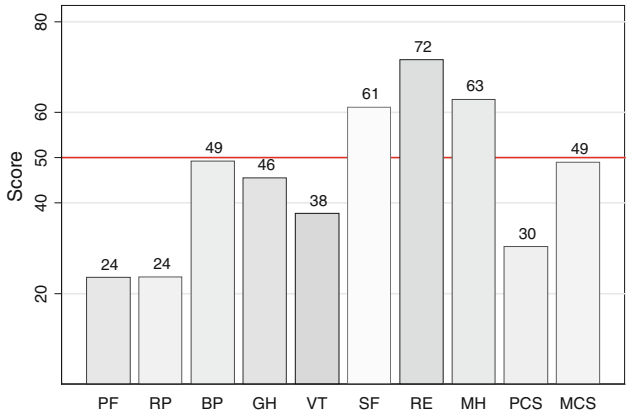

Fig. 2 SF-36 scores for elderly medicare beneficiaries with comorbid chronic conditions. *PF* physical function, *RP* role-physical, *BP* bodily pain, *GH* general health, *VT* vitality, *SF* social function, *RE* role-emotional, *MH* mental health, *PCS* physical component summary, *MCS* mental component summary

Table 3 Inter-scale correlations, internal consistency reliability, ceiling and flooring effects

	PF	RP	BP	GH	VT	SF	RE	MH
PF	(0.88)							
RP	0.33	(0.84)						
BP	0.29	0.40	(0.85)					
GH	0.32	0.39	0.41	(0.74)				
VT	0.39	0.41	0.39	0.54	(0.79)			
SF	0.33	0.46	0.37	0.43	0.44	(0.85)		
RE	0.10	0.32	0.26	0.28	0.28	0.35	(0.88)	
MH	0.12	0.27	0.36	0.42	0.42	0.40	0.56	(0.83)

Cronbach's alpha coefficients on the diagonal

sample scored well below national norms on physical health measures (e.g., PF, RP, and PCS), but at or near the norms on mental health measures (e.g., MH, RE).

Internal consistency

Table 3 shows the inter-scale correlations and internal consistency reliability. Most inter-scale correlation coefficients were low to medium (0.10–0.56). Higher correlation coefficients were found between scales with similar constructs (MH and VT) and low coefficients between scales with different constructs (PF and RE). All eight internal consistency reliability estimates exceeded the 0.70 level recommended for group comparisons. However, none met the criterion of 0.90 for person-level comparisons.

Factor structure

Two measurement models were fitted to the data. Model 1, with two latent variables (PHYS and MENT HRQoL), did

not fit the data well (see Table 4), and standardized residuals were not helpful in identifying possible modifications to the model. Model 2, on the other hand, had better overall model fit, though still did not fit the data well. Based on the size of the standardized residuals, we removed the cross-loading of GH on PHYS HRQoL and the cross-loading of VT on MENT HRQoL, and included a cross-loading of SF on PHYS (HRQoL) and a covariance between RP and RE [24, 30]. This resulted in a satisfactorily fitting measurement model (Model 2.F, Table 4).

To test invariance, we split the covariance matrix into low and high levels of comorbidity and applied the same factor structure identified in Model 2.F. As the fit of this model was satisfactory (Model 3, Table 4), all factor loadings and intercepts were constrained to equality across groups (Model 4, Table 4). This led to a significant deterioration in model fit ($\chi^2\Delta(12) = 30.29$, $p = 0.002$). When investigating the tenability of the constraints using global Chi-squared difference and OPCs, we found that the removal of the equality constraints associated with SF on PHYS HRQoL (Model 5, Table 4) and BP on PHYS HRQoL (Model 6, Table 4) was not tenable. These results suggest that PHYS HRQoL directly affected respondents' perception of SF and BP more strongly in respondents with high level of comorbidity than those with low level of comorbidities. After accounting for measurement bias, we assessed whether the difference in latent variable means for the low and high levels of comorbidity was different. As demonstrated in Table 5, the means and effect sizes for MENT HRQoL and GEN WB remained the same in Model 4 and Model 6; however, we would have overestimated the difference between respondents with low and high comorbidity had we not accounted for measurement bias (Fig. 3). After accounting for measurement bias, we can conclude that respondents with higher levels of comorbidity had significantly lower PHYS HRQoL, MENT HRQoL, and GEN WB.

Discussion

During the past two decades, SF-36 has been widely used in the general adult population across a variety of settings to measure HRQoL. While SF-36 generally performed well in the general adult population, evidence regarding its psychometric property when applied in program evaluation studies with older adults remains mixed [10, 31]. Our finding of good data quality suggests that administering SF-36 face-to-face by trained interviewers is appropriate and can be effective in addressing issues related to non-response among cognitively intact older adults with comorbidities. However, issues related to ceiling and/or flooring effects among certain scales (primarily in RP, RE, and SF) should be addressed by either using a modified version of the SF-36 or using one of the more recently developed instruments. Consistent with findings of earlier studies, our finding of modest internal consistency reliability suggests that the SF-36 should not be used for studies in which the individual-level change in HRQoL is the primary focus. In addition, our findings regarding the

Table 4 Overall goodness-of-fit and chi-square difference test

	Model	χ^2 (*df*)	OPC	RMSEA (90 % CI)	Comparison models	χ^2 difference (*df*)	*p* value	ECVI (90 % CI)
1	Measurement model 1; 2 latent variables	314.79 (18)	NA	0.129 (0.116; 0.141)	NA	NA	NA	0.352 (0.298; 0.415)
2	Measurement model 2; 3 latent variables	266.29 (15)	NA	0.131 (0.117; 0.145)	NA	NA	NA	0.315 (0.264; 0.373)
2.F	Measurement model 2; 3 latent variables	71.41 (15)	NA	0.062 (0.048; 0.077)	NA	NA	NA	0.117 (0.093; 0.148)
3	Model 2; low and high comorbidity	61.07 (28)	NA	0.049 (0.032; 0.066)	NA	NA	NA	0.189 (0.169; 0.216)
4	Low and high comorbidity; equality constraints	91.36 (40)	NA	0.053 (0.037; 0.065)	3 vs. 4	30.29 (12)	0.002	0.195 (0.170; 0.228)
5	SF factor loadings and intercepts—PHYS HRQoL	74.03 (38)	λ_1 0.09 λ_2 −0.18 τ_1 −0.10 τ_2 0.04	0.044 (0.029; 0.059)	4 vs. 5	17.34 (2)	<0.001	0.181 (0.160; 0.211)
6	BP factor loadings and intercepts—PHYS HRQoL	64.87 (36)	λ_1 0.08 λ_2 −0.10 τ_1 −0.16 τ_2 0.05	0.041 (0.024; 0.056)	5 vs. 6	9.16 (2)	0.010	0.176 (0.157; 0.204)

PHYS HRQoL physical health-related quality of life, *OPC* observed parameter change, *RMSEA* root mean square error of approximation, *ECVI* Expected Cross-Validation Index, *CI* confidence interval, *df* degrees of freedom

Table 5 Final model latent means and covariance estimates (Model 6)

	PHYS HRQoL Low comorbidity group	MENT HRQoL Low comorbidity group	GEN WB Low comorbidity group	PHYS HRQoL High comorbidity group	MENT HRQoL High comorbidity group	GEN WB High comorbidity group
Common factor means						
Means (Model 4)	0	0	0	−0.55	−0.22	−0.49
95 % CI				−0.73; −0.38	−0.38; −0.07	−0.66; −0.33
Effect size D				−0.50	−0.19	−0.44
Means (Model 6)	0	0	0	−0.45	−0.22	−0.49
95 % CI				−0.63; −0.29	−0.38; −0.07	−0.66; −0.33
Effect size D				−0.45	−0.19	−0.44
Common factor covariance/variance matrix						
PHYS $HRQoL_{Low\ comorbidity}$	1.00					
MENT $HRQoL_{Low\ comorbidity}$	0.38	1.00				
GEN $WB_{Low\ comorbidity}$	0.81	0.72	1.00			
PHYS $HRQoL_{High\ comorbidity}$	NA	NA	NA	0.49		
MENT $HRQoL_{High\ comorbidity}$	NA	NA	NA	0.22	1.05	
GEN $WB_{High\ comorbidity}$	NA	NA	NA	0.51	0.51	0.84

PHYS HRQoL physical health-related quality of life, *MENT HRQoL* mental health-related quality of life, *GEN HRQoL* general health-related quality of life, *CI* confidence interval

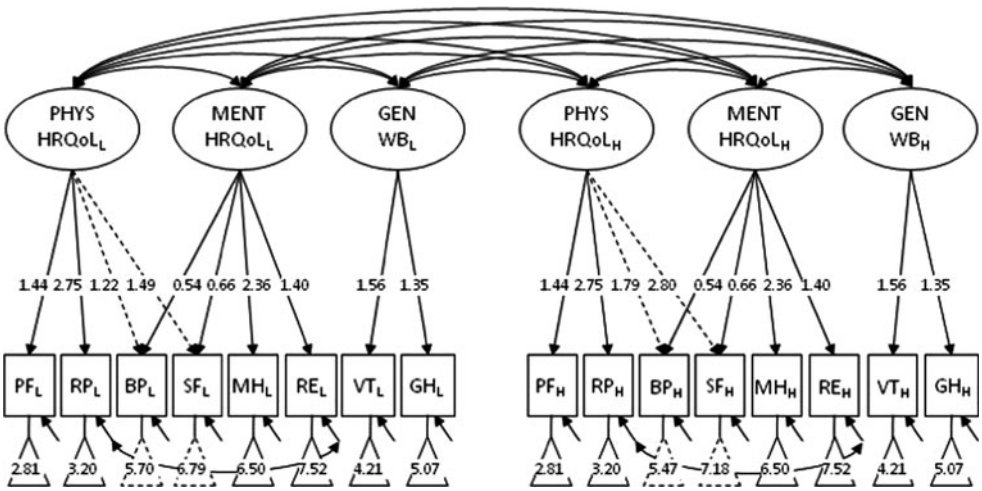

Fig. 3 SF-36 confirmatory factor analysis model: unstandardized parameter estimates from Model 6. *PF* physical functioning, *RP* role-physical, *BP* body pain, *SF* social functioning, *MH* mental health, *RE* role-emotional, *VT* vitality, *GH* general health

possible measurement bias with respect to the level of comorbidity suggest that clinicians and researchers should be cautious in instrument selection when considering the use of SF-36 among community-living older adults with high levels of comorbidities. The SF-36's measurement properties and its limitations are largely due to the theoretical foundation of the measurement of generic construct of well-being. As such, its use in older adults with comorbidities should take into consideration the impact of measurement bias.

Our findings on the missing data rates were very similar to those reported by a Canadian study of SF-36 among a group of frail older adults [15]. However, the missing data rates were lower than those reported by Gandek et al. [32] among a national sample of Medicare Managed Care (MMC) enrollees when a mail survey was supplemented by

a follow-up telephone survey. While the item-scale correlations were generally acceptable, serious issues existed in PF1, GH2, and GH4. The patterns of these correlations were similar to those found in the US MMC enrollee population [32], with even lower values found in the present study due to lower health status and higher disease burden. It is likely that these items present similar challenges for older adults with comorbidities. For example, PF1 asks: "Does your health limit you in vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?" This arguably is less relevant for the average older adult with comorbidities because the likelihood of them participating in these activities is low. Hayes et al. [4] suggested an alternative format in which the question is reworded as: "Does your health limit you in strenuous activities such as work around the house, making a bed, moving a table, and gardening?" However, because the SF-36 was designed as a generic measure of well-being in the general population, modifications like this may reduce the sensitivity of the instrument to differentiate among medium-to-higher functioning individuals. Therefore, future studies on the measurement of well-being among older adults should examine the trade-off between having a more general instrument and a more suitable instrument for those with comorbidities. In addition, future studies should examine whether modified questions lead to improved measurement properties in this population.

Another area of concern is significant floor and ceiling effects on many subscales. The effect of flooring may be especially problematic in measuring patient-reported outcomes over time because it is impossible to measure a decline once the floor of the subscale is reached. Our findings of substantial floor and/or ceiling effects on RP, BP, SF, and RE closely mirrored findings from previous study in the United States (MMC enrollees) and abroad (frail older adult sample) [15, 32]. This suggests that using the SF-36 among the chronically ill older population may result in lack of variability on these domains and subsequently poor discriminative ability of the subscale scores. The developers of SF-36 have since revised the response choices for the role functioning (RP, RE) items and evidence suggests that these modifications resulted in reduced floor and ceiling effects [33]. Nevertheless, modifications or development of specialized instruments may be needed to better measure these domains among older sicker populations.

The internal consistency reliability estimates for the subscales were all between 0.74 (GH) and 0.88 (PF, RE), suggesting that using SF-36 in this older adult population with comorbidities yielded reliable measures for group survey and program evaluation. However, our results echoed the findings of similar studies in that the reliability was insufficient for clinical applications among frail elderly persons [34]. The lack of homogeneity of the measures may be partially explained by the phenomenon of response shift, in which individuals with chronic conditions make adaptive adjustments to their internal standards of quality of life [17].

Results from CFA showed that the perception of SF and BP is stronger in respondents with high level of comorbidity than those with low level of comorbidity (measurement bias present). This suggests that respondents with high level of comorbidity place more emphasis on SF and BP in relation to PHYS HRQoL than those with low level of comorbidity group. These findings suggest that the high comorbidity group experienced an unobserved response shift in that the frames of reference used by the respondents with regard to SF and BP may have changed. However, future research using longitudinal data is needed to ascertain the source of the measurement bias. After we accounted for measurement bias, we found that the latent mean difference for PHYS HRQoL was higher prior to accounting for bias. Future studies should investigate the temporal stability of the response shift, as well as its impact on group comparisons regarding treatment effect in randomized controlled studies.

A number of limitations should be considered when interpreting the findings. First, the cross-sectional design of this study precludes drawing conclusion about why the measurement bias exists based on the levels of comorbidity. Second, the definition of high versus low levels of comorbidity may not be able to capture the effect of specific medical conditions. For example, studies have identified that certain chronic conditions (arthritis, chronic lung disease, and congestive heart failure) had larger impact on physical HRQoL and that the impact of chronic conditions on HRQoL was similar across countries [35, 36]. Future research is needed to examine whether different chronic conditions and/or different combinations of chronic conditions affect measurement bias in HRQoL differently.

In summary, despite the wide adoption of the SF-36 in program evaluation studies with older adults, our findings suggest that clinicians and researchers should be cautious in instrument selection when considering the use of SF-36 among community-living older adults with comorbidities. The SF-36's measurement properties and its limitations are largely due to the theoretical foundation of the measurement of generic construct of well-being. As such, its use in older adults with comorbidities should take into consideration the impact of measurement bias. In addition, SF36 may also be supplemented with other measures such as activities of daily living (ADL), and instrumental activities of daily living (IADL) in this population to alleviate issues such as flooring and ceiling effects. Moreover, recent advances in the uniform measurement of patient-reported outcomes should be considered as an alternative to the SF-

36 and tested for measurement bias. Finally, future research should examine the measurement properties of SF-36 in longitudinal studies to ensure it can capture changes in quality of life at the group level for older adults with high levels of comorbidities.

Acknowledgments We would like to thank Dr. Lynette L-Y Lim for providing us the programming code which forms the basis for generating Fig. 1.

References

1. Cella, D., Riley, W., Stone, A., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*, 1179–1194.
2. Ware, J. E., Jr. (1998). Gandek B: Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *Journal of Clinical Epidemiology, 51*, 903–912.
3. Ware, J. E., Jr. (1992). Sherbourne CD: The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*, 473–483.
4. Hayes, V., Morris, J., Wolfe, C., et al. (1995). The SF-36 health survey questionnaire: Is it suitable for use with older adults? *Age and Ageing, 24*, 120–125.
5. Parker, S. G., Peet, S. M., Jagger, C., et al. (1998). Measuring health status in older patients. The SF-36 in practice. *Age and Ageing, 27*, 13–18.
6. Rothrock, N. E., Hays, R. D., Spritzer, K., et al. (2010). Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology, 63*, 1195–1204.
7. Yu, Y. F., Yu, A. P., & Ahn, J. (2007). Investigating differential item functioning by chronic diseases in the SF-36 health survey: A latent trait analysis using MIMIC models. *Medical Care, 45*, 851–859.
8. Anderson, G. F. (2005). Medicare and chronic conditions. *New England Journal of Medicine, 353*, 305–309.
9. Hill, S., Harries, U., & Popay, J. (1996). Is the short form 36 (SF-36) suitable for routine health outcomes assessment in health care for older people? Evidence from preliminary work in community based health services in England. *Journal of Epidemiology and Community Health, 50*, 94–98.
10. Mallinson, S. (1998). The Short-Form 36 and older people: some problems encountered when using postal administration. *Journal of Epidemiology and Community Health, 52*, 324–328.
11. Reuben, D. B., Valle, L. A., Hays, R. D., et al. (1995). Measuring physical function in community-dwelling older persons: A comparison of self-administered, interviewer-administered, and performance-based measures. *Journal of the American Geriatrics Society, 43*, 17–23.
12. Lyons, R. A., Perry, H. M., & Littlepage, B. N. (1994). Evidence for the validity of the Short-form 36 Questionnaire (SF-36) in an elderly population. *Age and Ageing, 23*, 182–184.
13. Hobson, J. P., & Meara, R. J. (1997). Is the SF-36 health survey questionnaire suitable as a self-report measure of the health status of older adults with Parkinson's disease? *Quality of Life Research, 6*, 213–216.
14. O'Mahony, P. G., Rodgers, H., Thomson, R. G., et al. (1998). Is the SF-36 suitable for assessing health status of older stroke patients? *Age and Ageing, 27*, 19–22.
15. Stadnyk, K., Calder, J., & Rockwood, K. (1998). Testing the measurement properties of the Short Form-36 Health Survey in a frail elderly population. *Journal of Clinical Epidemiology, 51*, 827–835.
16. Walters, S. J., Munro, J. F., & Brazier, J. E. (2001). Using the SF-36 with older adults: A cross-sectional community-based survey. *Age and Ageing, 30*, 337–343.
17. Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science and Medicine, 48*, 1507–1515.
18. Ware, J. E., Jr. (1998). Gandek B: Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*, 945–952.
19. Meng, H., Wamsley, B. R., Friedman, B., et al. (2010). Impact of body mass index on the effectiveness of a disease management-health promotion intervention on disability status. *American Journal of Health Promotion, 24*, 214–222.
20. Morris, J. N., Fries, B. E., Mehr, D. R., et al. (1994). MDS Cognitive Performance Scale. *The Journal of Gerontology, 49*, M174–M182.
21. Cureton, E. E. (1966). Corrected item-test correlations. *Psychometrika, 31*, 93–96.
22. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
23. Lim, L. L., Seubsman, S. A., & Sleigh, A. (2008). Thai SF-36 health survey: tests of data quality, scaling assumptions, reliability and validity in healthy men and women. *Health and Quality of Life Outcomes, 6*, 52.
24. Keller, S. D., Ware, J. E., Jr, Bentler, P. M., et al. (1998). Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*, 1179–1188.
25. Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230–258.
26. Millsap, R. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.
27. King-Kallimanis, B. L., Oort, F. J., Nolte, S., et al. (2011). Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Quality of Life Research, 20*, 1527–1540.
28. Holms, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.
29. Ware, J., & Kosinski, M. (2001). *SF-36 Physical and Mental Health Summary Scales: A manual for users of version 1* (2nd ed.). Lincoln, RI: QualityMetric Inc.
30. Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research, 14*, 599–609.
31. McHorney, C. A., Ware, J. E., Jr, Lu, J. F., et al. (1994). The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*, 40–66.
32. Gandek, B., Sinclair, S. J., Kosinski, M., et al. (2004). Psychometric evaluation of the SF-36 health survey in Medicare managed care. *Health Care Financing Review, 25*, 5–25.
33. Ware, J. E., Kosinski, M., & Dewey, J. E. (2000). *How to score version 2 of the SF-36 Health Survey*. Lincoln, RI: QualityMetric Inc.
34. McHorney, C. A. (1996). Measuring and monitoring general health status in elderly persons: Practical and methodological issues in using the SF-36 Health Survey. *Gerontologist, 36*, 571–583.
35. Alonso, J., Ferrer, M., Gandek, B., et al. (2004). Health-related quality of life associated with chronic conditions in eight

countries: Results from the International Quality of Life Assessment (IQOLA) Project. *Quality of Life Research, 13*, 283–298.

36. Banks, P., Martin, C. R., & Petty, R. K. (2012). The factor structure of the SF-36 in adults with progressive neuromuscular disorders. *Journal of Evaluation Clinical Practice, 18*, 32–36.