Qual Life Res (2015) 24:1809–1822
DOI 10.1007/s11136-015-0916-8

Testing item response theory invariance of the standardized Quality-of-life Disease Impact Scale (QDIS®) in acute coronary syndrome patients: differential functioning of items and test

Nina Deng · Milena D. Anatchkova ·
Molly E. Waring · Kyung T. Han · John E. Ware Jr.

Accepted: 5 January 2015 / Published online: 20 January 2015

Abstract

Purpose The Quality-of-life (QOL) Disease Impact Scale (QDIS®) standardizes the content and scoring of QOL impact attributed to different diseases using item response theory (IRT). This study examined the IRT invariance of the QDIS-standardized IRT parameters in an independent sample.

Method The differential functioning of items and test (DFIT) of a static short-form (QDIS-7) was examined across two independent sources: patients hospitalized for acute coronary syndrome (ACS) in the TRACE-CORE study ($N = 1{,}544$) and chronically ill US adults in the QDIS standardization sample. "ACS-specific" IRT item parameters were calibrated and linearly transformed to compare to "standardized" IRT item parameters. Differences in IRT model-expected item, scale and theta scores were examined. The DFIT results were also compared in a standard logistic regression differential item functioning analysis.

Results Item parameters estimated in the ACS sample showed lower discrimination parameters than the standardized discrimination parameters, but only small differences were found for thresholds parameters. In DFIT, results on the non-compensatory differential item functioning index (range 0.005–0.074) were all below the threshold of 0.096. Item differences were further canceled out at the scale level. IRT-based theta scores for ACS patients using standardized and ACS-specific item parameters were highly correlated ($r = 0.995$, root-mean-square difference = 0.09). Using standardized item parameters, ACS patients scored one-half standard deviation higher (indicating greater QOL impact) compared to chronically ill adults in the standardization sample.

Conclusion The study showed sufficient IRT invariance to warrant the use of standardized IRT scoring of QDIS-7 for studies comparing the QOL impact attributed to acute coronary disease and other chronic conditions.

Keywords Item response theory invariance · Differential item functioning · Differential test (scale) functioning · Measurement invariance · Disease-specific quality-of-life measures

Abbreviations

ACS	Acute coronary syndrome
ACS-LT	ACS-specific linearly transformed
CAT	Computerized adaptive testing
CDIF	Compensatory differential item functioning
CFA	Confirmatory factor analysis
DFIT	Differential functioning of items and tests
DICAT	The computerized adaptive Assessment of disease impact project

N. Deng (✉) · M. D. Anatchkova · M. E. Waring ·
J. E. Ware Jr.
Department of Quantitative Health Sciences, University of
Massachusetts Medical School, Worcester, MA 01655, USA
e-mail: nndeng@gmail.com

N. Deng
Measured Progress, Inc., Dover, NH 03820, USA

M. D. Anatchkova
Evidera, Lexington, MA 02420, USA

K. T. Han
Graduate Management Admission Council, Reston, VA 20195,
USA

J. E. Ware Jr.
John Ware Research Group, Incorporated, Worcester,
MA 01655, USA

DIF	Differential item functioning
DTF	Differential test (scale) functioning
GPCM	Generalized partial credit model
ICC	Item characteristic curve
IPD	Item parameter drift
IRT	Item response theory
MLHFQ	Minnesota Living with Heart Failure Questionnaire
NCDIF	Non-compensatory differential item functioning
PRO	Patient-reported outcome
PROMIS	Patient Reported Outcomes Measurement Information System
QDIS®	Quality-of-life Disease Impact Scale
QDIS-7	7-item short-form of QDIS®
QOL	Quality-of-life
RMSD	Root-mean-square difference
SAQ	Seattle Angina Questionnaire
TCC	Test characteristic curve
TRACE-CORE	The Transitions, Risks, and Actions in Coronary Events-Center for Outcomes Research and Education project

Introduction

Disease-specific patient-reported outcome (PRO) measures have been widely used to assess impacts of specific diseases on health-related quality of life (QOL) [1, 2]. These measures, however, have largely been developed independently of each other, greatly limiting comparisons of how different diseases impact QOL. For example, two of the heart disease-specific PRO measures, the Seattle Angina Questionnaire (SAQ) [3] and the Minnesota Living with Heart Failure Questionnaire (MLHFQ) [4], both represent only four common item content areas and they differ in scoring, limiting comparisons of QOL impact attributed to angina versus congestive heart failure. It is even more challenging to compare disease-specific QOL impact across more distinct conditions (e.g., heart diseases versus arthritis). Comparisons of general profiles show that some diseases affect the same areas of life but may have different "signature" profiles of impact on QOL depending on their varying symptoms, stage of treatment, progression, and overall severity [5, 42]. It is generally believed that disease-specific measures of QOL impact are more responsive than general QOL measures [43]. In theory, increased standardization of the content and scoring of QOL impact attributed to specific diseases would add to the advantages of specificity, the advantage of a better understanding of their relative overall impact on QOL.

The Quality-of-life Disease Impact Scale (QDIS®) was developed to standardize the content and scoring of QOL impact attributed to specific diseases using item response theory (IRT), in order to enable more meaningful comparisons of QOL impact across diseases [6–8]. QDIS items are standardized to have identical content (questions and response categories) and to differ only in their attribution to a specific disease. The 49 items form an item bank that allows researchers to administer the test in different ways, including: the full bank, a fixed short form such as the QDIS-7, or a computerized adaptive assessment. QDIS items were self-administered and IRT-calibrated across five independent US household samples differing in their chronic diseases. Tests in these development samples provided the basis for QDIS cross-disease standardized IRT item parameters. QDIS scores based upon the standardized IRT item parameters were highly consistent with those based on disease-specific IRT item parameters [44].

In this study, we examined the invariance (equivalence) of standardized IRT item parameters provided by the QDIS developer in comparison with parameters estimated for an independent clinical sample hospitalized with acute coronary syndrome (ACS). ACS is a major cause of morbidity and has been shown to have substantial impact on patients' QOL [12, 13]. These tests of IRT invariance are stringent since results for patients interviewed after an acute coronary event were compared with results from Internet-based self-administered surveys for household samples.

Methods

Participants

Standardization sample

The QDIS was developed during the computerized adaptive assessment of disease impact test (DICAT) Project, a three-year NIH-sponsored effort to develop and improve disease-specific PRO measures [6]. DICAT participants were recruited from an online research panel (www.knowledgepanel.com) in order to achieve a probability sample representing the US adult population and to achieve quotas for pre-identified samples with each of five chronic conditions: arthritis (including osteoarthritis and rheumatoid arthritis), cardiovascular disease (including angina, congestive heart failure, myocardial infarction), chronic kidney disease, diabetes, and respiratory disease (including asthma and chronic obstructive pulmonary disease). It should be noted that ACS was among the cardiovascular diseases sampled in the DICAT project. The pre-identified chronically ill samples ($N = 5{,}451$) were used to calibrate the QDIS-standardized IRT item parameters (QDIS standardization sample). Responses were collected online because of the practical advantages demonstrated in other recent studies (e.g., PROMIS). The latter has also added to the substantial evidence that Internet-based surveys produce both IRT parameters and scale scores that are equivalent to those from other self-administered modes [50]. The DICAT study was approved by the New England Institutional Review Board (NEIRB 09-062).

ACS clinical sample

An independent sample was analyzed from the Transitions, Risks, and Actions in Coronary Events—Center for Outcomes Research and Education (TRACE-CORE) project, a prospective cohort study of patients hospitalized for ACS at six hospitals in Massachusetts and Georgia. The design and methods of TRACE-CORE are described elsewhere [14]. Briefly, one month following hospital discharge, ACS patients ($N = 1{,}544$) completed a 7-item short-form of the QDIS during a telephone interview. The Committee for the Protection of Human Subjects at the University of Massachusetts Medical School and Institutional Review Boards at each participating site approved this study.

Measures

The full-length QDIS consists of 49 items assessing disease-specific impacts on patients' physical, emotional, social functioning, and well-being. For each item, content is the same across diseases except for the disease attribution. For example, in DICAT, the pre-identified cardiovascular patients were asked "how often did your angina/CHF/MI limit your everyday activities during the past 4 weeks," and the pre-identified arthritis patients were asked "how often did your OA/RA limit your everyday activities during the past 4 weeks." In TRACE-CORE, the ACS patients were interviewed using a more general attribution, such as "how often did your heart problem limit your everyday activities during the past 4 weeks." Each item included the same five-choice categorical rating scale, higher score indicating worse QOL. The QDIS-standardized IRT item parameters were calibrated using the entire 49 items fitted with the generalized partial credit model (GPCM) across samples of the five chronic conditions pre-identified in the QDIS standardization sample [44]. IRT was used to enable administrations and comparisons across different (adaptive and static) QDIS forms. The 7-item QDIS-7, which is a static short-form constructed and recommended by QDIS developers using various criteria [7], was administered in TRACE-CORE as required by its study protocol including paper–pencil interviewer-administered surveys (as opposed to computer adaptive testing). The QDIS-7 covers seven distinct content areas and performed very well for cardiovascular and other conditions in the DICAT study [7] (see content area in Table 2). Standardized IRT parameters for the seven items had already been calibrated along with the rest of the 49 items.

Analyses

Descriptive statistics

Descriptive statistics for QDIS-7 were compared across the samples, including the item-level mean, standard deviation (SD), and item-total correlations (corrected for overlap as recommended for short surveys [15]). Scale internal consistency (the Cronbach's alpha [16]) and ceiling and floor percentages were also compared.

Differential functioning of items and tests (DFIT)

To test the IRT invariance of QDIS-7 across samples, differential functioning was compared for the items and the scale. For this purpose, IRT item parameters were calibrated independently within the QDIS standardization sample and the ACS sample. Results were compared in terms of item-, scale-, and person-level score differences. This method, referred to as differential functioning of items and tests (DFIT) [17, 18], involves three general steps as illustrated in Fig. 1.

Step 1: calibrating sample-specific IRT item parameters Since the standardized IRT item parameters were provided by the QDIS developer, we only needed to calibrate IRT parameters for the new ACS sample. The same IRT model (GPCM) as used in the standardization sample was applied. Items were freely calibrated without imposing

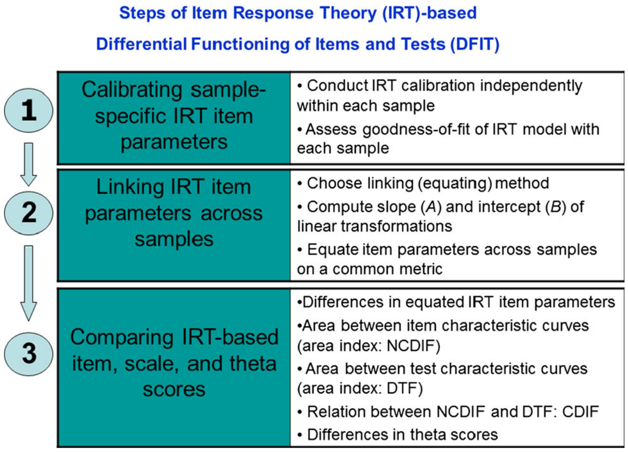

Steps of Item Response Theory (IRT)-based
Differential Functioning of Items and Tests (DFIT)
1
Calibrating sample-specific IRT item parameters
• Conduct IRT calibration independently within each sample
• Assess goodness-of-fit of IRT model with each sample
2
Linking IRT item parameters across samples
• Choose linking (equating) method
• Compute slope (A) and intercept (B) of linear transformations
• Equate item parameters across samples on a common metric
3
Comparing IRT-based item, scale, and theta scores
•Differences in equated IRT item parameters
•Area between item characteristic curves (area index: NCDIF)
• Area between test characteristic curves (area index: DTF)
• Relation between NCDIF and DTF: CDIF
• Differences in theta scores

Fig. 1 Illustration of steps in IRT-based differential functioning of items and tests (DFIT) analyses

any prior knowledge of the standardized IRT item parameters. Results in the independent clinical sample are referred to as "ACS-specific" IRT item parameters. As a validity check, we evaluated the goodness of fit of the GPCM model in the new ACS sample, based on the Chi-square item fit statistics [19] and IRT graphical residual plots [20, 21]. The IRT calibration was completed using the software Parscale [22], and the IRT graphical residual analyses were completed using ResidPlots [23].

Step 2: linking IRT item parameters across samples Despite the desirable feature of IRT parameter invariance across samples, the ACS-specific and standardized IRT item parameters were not readily comparable due to IRT scale indeterminacy [9, 24]. IRT scale indeterminacy indicates that the solution to IRT item parameter estimation is mathematically unidentifiable (illustration in 'Item response theory indeterminacy' in Appendix 1). Due to the nature of IRT indeterminacy, item parameters from separate IRT calibrations are invariant (consistent) only up to a set of linear transformations rather than being numerically equivalent.

The linking (equating) process aims to compute the slope and intercept of the linear transformation, in order to put IRT item parameters from separate calibrations back onto a common metric for meaningful comparisons [25]. There are various linking methods available, such as using item parameters, test characteristic functions, or theta scores, and the choice often depends on factors such as study design, item quality, and software availability [26]. Because only 7 items were used, any substantial DIF could greatly contaminate a linking process based on item parameters. Therefore, we chose to link item parameters based on theta scores. To begin, we scored the QDIS-7 for ACS patients twice using the ACS-specific and standardized IRT parameters, respectively. Then we computed the slope (A) and the intercept (B) of the linear transformation based on the two sets of theta scores (calculations in 'Slope (A) and intercept (B) of linear transformation using IRT theta scores' in Appendix 1). Lastly, we applied the slope and intercept to link the ACS-specific with the standardized IRT item parameters. The ACS-specific item parameters after the linear transformation were labeled "ACS-LT." After linking, the ACS-LT and standardized IRT item parameters were meaningfully comparable.

Step 3: comparing IRT-based item, scale, and theta scores First, we evaluated scatter plots of the ACS-LT and standardized item parameters, for IRT a- and b-parameters separately, in terms of the magnitude of their correlation and differences (distance off the diagonal line). Any data point with a difference larger than 2 times the SD of the differences across the items, roughly equivalent to a significance level of 0.05, was flagged [27]. Next, the IRT item characteristic curve (ICC), encompassing both a- and b-parameters, was used to evaluate their joint impact on item scores. Two ICCs were plotted for each item, one using the ACS-LT and the other using the standardized IRT item parameters. Smaller areas between two ICCs indicate smaller impact on item scores. The non-compensatory differential item functioning (NCDIF) index was computed to objectively compare the area across the items (calculation in 'Non-compensatory differential item functioning (NCDIF)' in Appendix 1) [11, 28]. The NCDIF index was compared to the cutoff calculated as 0.006 $(k - 1)^2$. The cutoff is 0.096 when $k = 5$, which is for five-choice categorical rating items [29–31].

The IRT test characteristic curve (TCC), which aggregates ICCs across the items, was used to evaluate the impact of items on scale scores. Likewise, two TCCs were plotted, one using the ACS-LT and the other using the standardized item parameters. In this plot, the more overlapping the two curves, the smaller the impact on the scale scores. The differential test functioning (DTF) index was computed to quantify the area (calculation in 'Test differential functioning (DTF)' in Appendix 1). Additionally, a third index, the compensatory differential item functioning (CDIF) was computed to evaluate the relative contribution of each item score difference to the scale score difference [32], thus relating the NCDIF with the DTF (calculation in 'Compensatory differential item functioning' in Appendix 1). Specifically, an item with a large positive CDIF contributes a lot to the DTF; on the contrary, an item with a small or negative CDIF contributes little or helps to eliminate the DTF. The statistical language R [45] was used to compute these DFIT indices based on the formulae illustrated in the appendix.

Finally, we examined the difference between the ACS-specific and the standardized IRT-based theta scores. To make them comparable, we performed the linear transformation using the slope (A) and intercept (B) derived in Step 2. After the linear transformation, we constructed a scatter plot of the standardized and the linearly transformed ACS-specific (ACS-LT) theta scores and evaluated their correlation and differences.

Logistic regression DIF

The widely used logistic regression DIF analysis was performed for validation and to enable comparisons with the previous DFIT analyses [10]. In this regression, the item response was treated as the dependent variable and the scale score was treated as the independent variable. We conducted the analyses using both the summed scale and IRT-based theta scores as the matching independent variable, respectively. A binary grouping variable (dummy coded for the ACS and the standardization sample) and an interaction term (group by scale score) were added to the model. An item was considered to exhibit DIF if the model fit was significantly improved as a consequence. We detected DIF items based on the criteria of change in the proportion of explained variance ($\Delta R^2 > 0.02$) and change in the regression beta coefficient ($\Delta\beta > 10$ %) [33, 34]. The Chi-square statistic was also computed but was considered only as a secondary criterion due to its over-sensitivity to large sample sizes. The R package lordif [51] was used to perform the logistic regression.

Results

Participant demographics

The two samples were statistically different on age, gender, race/ethnicity, educational level, and marital status; the ACS sample was older, more likely to be male, non-Hispanic black, less educated, and divorced or separated (Table 1).

Descriptive statistics

Table 2 summarizes item-level descriptive statistics and item content. Compared to the standardization sample, ACS patients had higher means indicating greater disease burden, and larger SD indicating greater variability (a more heterogeneous group). Lower item-total correlations indicating weaker discrimination across the items were also observed. The floor and ceiling percentages were 12.6 and 0.86 % with the ACS patients (mean = 15.6, SD = 7.1 of summed scores), versus 24.8 and 0.35 % for the standardization sample (mean = 12.8, SD = 6.5 of summed scores), respectively, indicating that a larger proportion of the ACS sample reported at least some degree of disease burden. The scale internal consistency estimates were high and similar (Cronbach's $\alpha = 0.91$ and 0.94, for the ACS and standardization sample, respectively).

Table 1 Demographic characteristics of the QDIS standardization sample in the DICAT and the ACS sample in the TRACE-CORE, *N* (%) or M ± SD

	QDIS standardization sample ($N = 5{,}451^a$)	ACS sample ($N = 1{,}544^b$)	*P* value
Male	2,344 (43.0)	1,026 (66.5)	<0.001
Age	59.56 ± 13.73	61.85 ± 11.05	<0.001
Median	61	62	
Range	18–97	28–92	
Race/ethnicity			<0.001
White, non-Hispanic	4,358 (79.9)	1,194 (77.3)	
Black, non-Hispanic	431 (7.9)	204 (13.2)	
Hispanic/Latino	318 (5.8)	48 (3.1)	
Other race/ethnicity or multi-racial[c]	344 (6.3)	86 (5.6)	
Education			<0.001
Less than high school	179 (3.3)	204 (13.2)	
High school	1,000 (18.3)	458 (29.7)	
Some college	2,071 (38.0)	466 (30.2)	
Bachelor's degree or higher	2,201 (40.4)	415 (26.9)	
Marital status			<0.001
Married, or living with partner	3,576 (65.6)	936 (60.6)	
Divorced or separated	483 (8.9)	261 (16.9)	
Widowed	809 (14.8)	178 (11.5)	
Never married	583 (10.7)	168 (10.9)	

[a] $N = 33$ in the QDIS standardization sample were excluded from IRT invariance analyses since their pre-identified chronic conditions were not consistently confirmed

[b] Missing data in the ACS sample on race/ethnicity ($N = 8$), education ($N = 1$), and marital status ($N = 1$)

[c] Other race includes Asian, Pacific Islander, Native American

QDIS Quality-of-life Disease Impact Scale, *DICAT* computerized adaptive assessment of disease impact project, *ACS* acute coronary syndrome, *TRACE-CORE* Transitions, Risks, and Actions in Coronary Events—Center for Outcomes Research and Education

Table 2 Item description, statistics, and IRT invariance analyses across the ACS and QDIS standardization samples for the QDIS-7

Item #	Abbreviated text (content domain[a])	Sample[b]	Descriptive statistics		IRT parameters			DFIT[d]		LR-DIF[e]	
			Mean (SD)	R^c	Slop (SE)	Minimum threshold (SE)	Maximum threshold (SE)	NCDIF	CDIF	ΔR^2	Δß
1	Everyday activity QOL (QL)	ACS-specific	2.24 (1.22)	0.77	2.10 (0.09)	−0.24 (0.04)	1.83 (0.07)	0.005	0.038	0.000	0.001
		ACS-LT	–	–	2.14	0.25	2.27				
		Standardization	1.80 (1.04)	0.84	3.55	0.18	2.11				
2	Everyday work school chores (RL)	ACS-specific	2.45 (1.39)	0.75	1.96 (0.08)	−0.14 (0.04)	1.00 (0.06)	0.074	−0.032	0.009	0.009
		ACS-LT	–	–	2.01	0.35	1.47				
		Standardization	1.77 (1.06)	0.88	4.67	0.26	1.96				
3	Social family friends (SA)	ACS-specific	2.00 (1.23)	0.76	2.08 (0.09)	0.38 (0.04)	1.41 (0.08)	0.009	0.030	0.001	0.002
		ACS-LT	–	–	2.12	0.86	1.87				
		Standardization	1.56 (0.96)	0.84	4.56	0.57	2.04				
4	Physical function walking (PF)	ACS-specific	2.24 (1.33)	0.74	1.58 (0.06)	0.30 (0.05)	1.18 (0.08)	0.042	0.062	0.005	0.024
		ACS-LT	–	–	1.61	0.78	1.64				
		Standardization	1.95 (1.20)	0.80	2.20	0.25	1.64				
5	Bother emotionally (EM)	ACS-specific	2.01 (1.19)	0.66	1.03 (0.05)	0.58 (0.07)	1.41 (0.15)	0.027	0.100	0.000	0.001
		ACS-LT	–	–	1.05	1.05	1.86				
		Standardization	1.67 (1.00)	0.76	2.05	0.53	2.00				
6	General health, worry (GH)	ACS-specific	2.33 (1.22)	0.62	0.86 (0.04)	−0.09 (0.08)	1.33 (0.15)	0.063	0.108	0.003	0.016
		ACS-LT	–	–	0.88	0.40	1.79				
		Standardization	2.11 (1.13)	0.74	1.49	−0.10	1.86				
7	Worn out, tired (FT)	ACS-specific	2.36 (1.27)	0.77	1.85 (0.07)	−0.15 (0.04)	1.35 (0.07)	0.012	0.065	0.001	0.004
		ACS-LT	–	–	1.89	0.33	1.80				
		Standardization	1.94 (1.13)	0.86	3.42	0.06	1.81				

[a] *QL* Quality of life, *RL* role functioning, *SA* social activity, *PF* physical functioning, *EM* emotional, *GH* general health, *FT* fatigue

[b] ACS-specific = clinical sample of acute coronary syndromes (ACS) patients in the TRACE-CORE (N = 1,544), ACS-LT = linearly transformed ACS-specific IRT item parameters using slope = 0.979 and intercept = 0.485, Standardization = QDIS standardization sample of household adults pre-identified with one of five chronic conditions in the DICAT (N = 5,451)

[c] The item-total correlation corrected for overlap in short surveys

[d] *DFIT* differential functioning of items and tests analyses, *NCDIF* non-compensatory differential item functioning index, *CDIF* compensatory differential item functioning index

[e] LR-DIF = logistic regression differential item functioning analyses, ΔR^2 = change in the proportion of explained variance, $\Delta\beta$ = change in the regression beta coefficient

Differential functioning of items and tests (DFIT)

Step 1. ACS-specific IRT calibration

While the Chi-square item fit statistics were all large and significant, the IRT graphical residual analyses (‘Appendix 2’) suggest a reasonable fit between the IRT model and the new ACS sample: The observed values fell within the 95 % confidence bands of IRT model-expected values. Item 6 presented a poorer fit with a few values falling outside the confidence bands. The standardized residuals of all items were approximately normally distributed (mean = −0.05, SD = 1.63), suggesting an acceptable model fit at the scale level. The ACS-specific IRT a-parameter ranged 0.86–2.10, and b-parameters ranged −0.24–1.83 (Table 2).

Step 2. IRT parameter linking

The slope and the intercept of the linear transformation were computed: $A = 0.979$ and $B = 0.485$, based on the theta score distribution of ACS patients using the ACS-specific (mean = −0.010, SD = 0.937) and the standardized (mean = 0.476, SD = 0.917) IRT item parameters. After linking, there were three sets of IRT item parameters for each item: the standardized (derived from the standardization sample provided by the QDIS developer), ACS-specific (before linear transformation), and ACS-LT (after linear transformation) item parameters (Table 2).

Step 3. Differences in item, scale, and theta scores

The standardized and ACS-LT IRT item parameters were highly correlated, $r = 0.89$ and 0.90, for a-, and *b*-parameters, respectively (Fig. 2). For *a*-parameters, all items had a significantly smaller value in the ACS-LT than the standardized version; Item 2 (“Everyday work school chores”) and Item 3 (“Social family friends”) had the largest drift (standardized distances of 3.24 and 2.97); The *b*-parameters appear to be far more stable across the standardized and ACS-LT versions, although Item 6 (“General health, worry”) had a greater drift in one threshold (b_3, the threshold between the response categories of “Sometimes” and “Often,” standardized distance of 2.36).

The areas between the ICC of the standardized and ACS-LT IRT item parameters were, however, small across the items (Fig. 3). The area index NCDIF (Table 2) was smaller than the suggested threshold of 0.096 for all items; Items 2 and 6, again, showed the largest values (NCDIF = 0.074 and 0.063, respectively). The TCC (bottom right corner, Fig. 3) had only small discrepancies. The area index DTF is 0.37. The CDIF index (Table 2) showed that Item 6 (CDIF = 0.108) contributed the most to, while Item 2 (CDIF = −0.032) actually eliminated the scale score difference. This is consistent with Fig. 2, where Item 2, despite a relatively large area between the curves, was offset by differences in an opposite direction for the other items—the curve of ACS-LT on the top of that of the standardized item parameter.

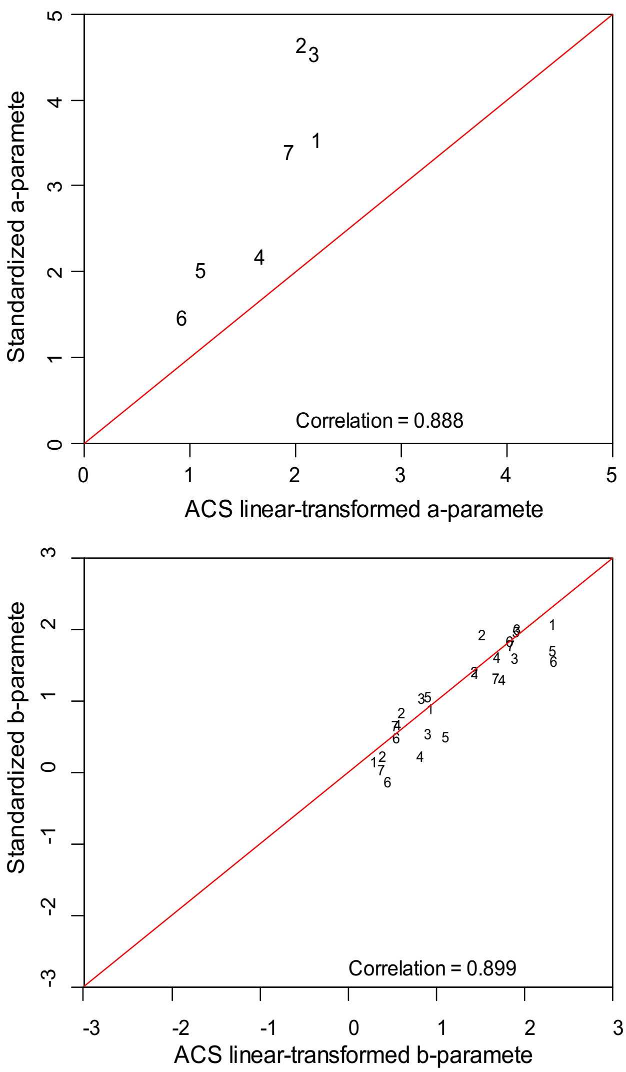


Fig. 2 *Scatter plot* of standardized and linearly transformed ACS-specific (ACS-LT) IRT item slope (*top*) and item threshold (*bottom*) parameters for the QDIS-7. The numbers in the plots (1–7) are the indices representing item 1–7

The theta scores of ACS patients using the standardized and ACS-LT IRT item parameters were very highly correlated ($r = 0.995$) and presented a strong linear relationship (Fig. 4, top). The mean difference between the standardized and ACS-LT theta scores was plotted as a function of the ACS-LT theta score (Fig. 4, bottom). The standardized theta estimates were greater than the ACS-LT estimate on the lower end of theta scale, and lower on the higher end of the theta scale. Larger residuals (up to 0.23

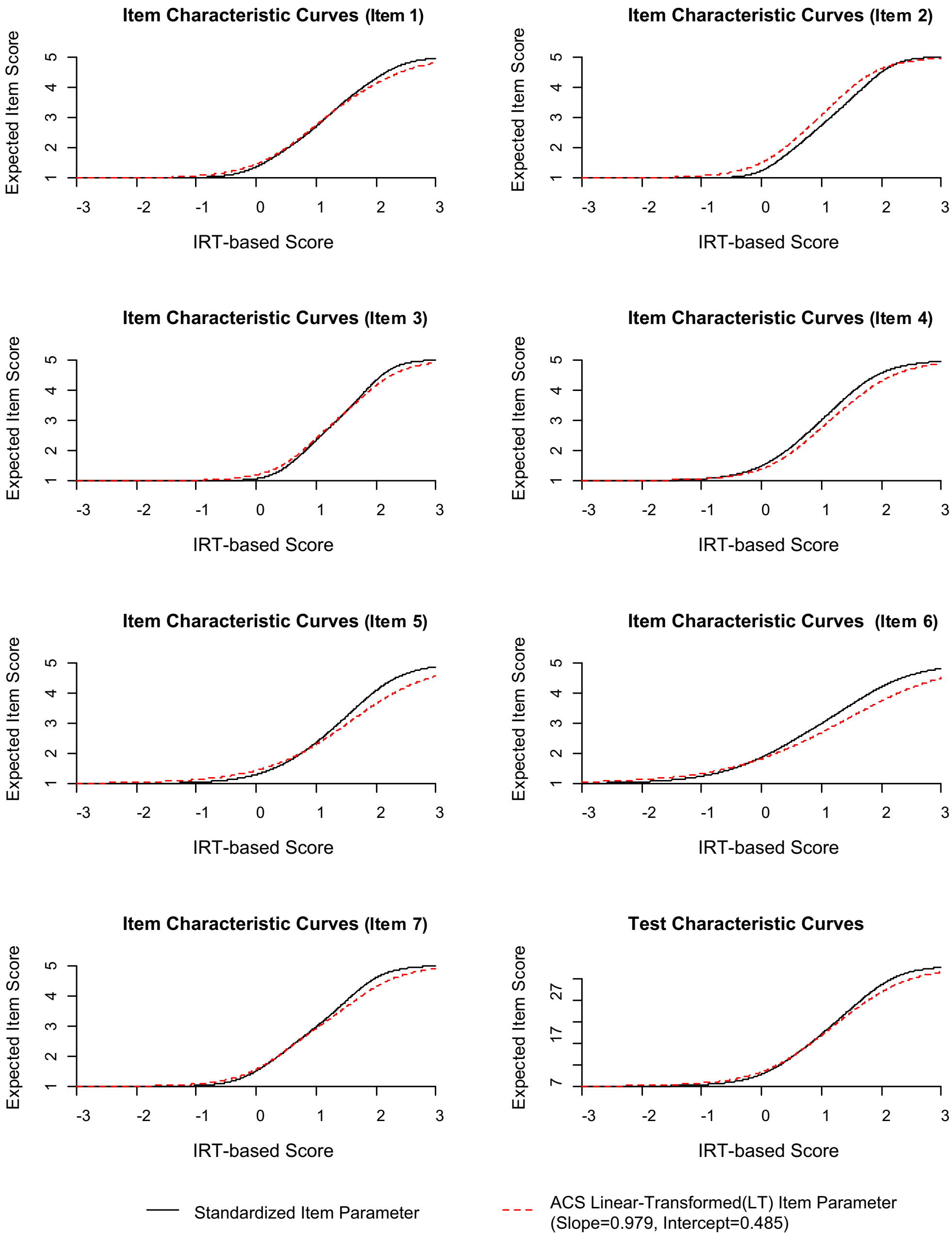


Fig. 3 Item and test characteristic curves of standardized and linearly transformed ACS-specific (ACS-LT) IRT item parameters for the QDIS-7. There are seven plots of item characteristic curves for item 1–7, respectively, and one plot of test characteristic curve for the total scale (*bottom right corner*)

SD) were observed at the two extreme ends. The root-mean-squared difference (RMSD) was 0.09. When estimating using the standardized item parameters, the ACS patients' standardized theta scores had mean = 0.476, SD = 0.917, one-half SD higher (greater disease burden) than the standardization sample that had mean = 0 and SD = 1.

Logistic regression DIF

While the Chi-square statistics were all large and significant as expected for a large sample, no items were detected with meaningful DIF using the criterion of $\Delta R^2 > 0.02$ or $\Delta\beta > 10$ %. The results based on the summed scores and IRT theta scores were highly similar (differences in ΔR^2 or $\Delta\beta$ only on the third decimal or smaller). Table 2 displays the latter, where Item 2 had the largest ΔR^2 (0.009) and Item 4 ("Physical function walking") had the largest $\Delta\beta$ (0.024).

Discussion and conclusions

Our study evaluated a fundamental psychometric assumption of IRT invariance for a short-form of the IRT-based Quality-of-Life Disease Impact Scale, the QDIS-7, across

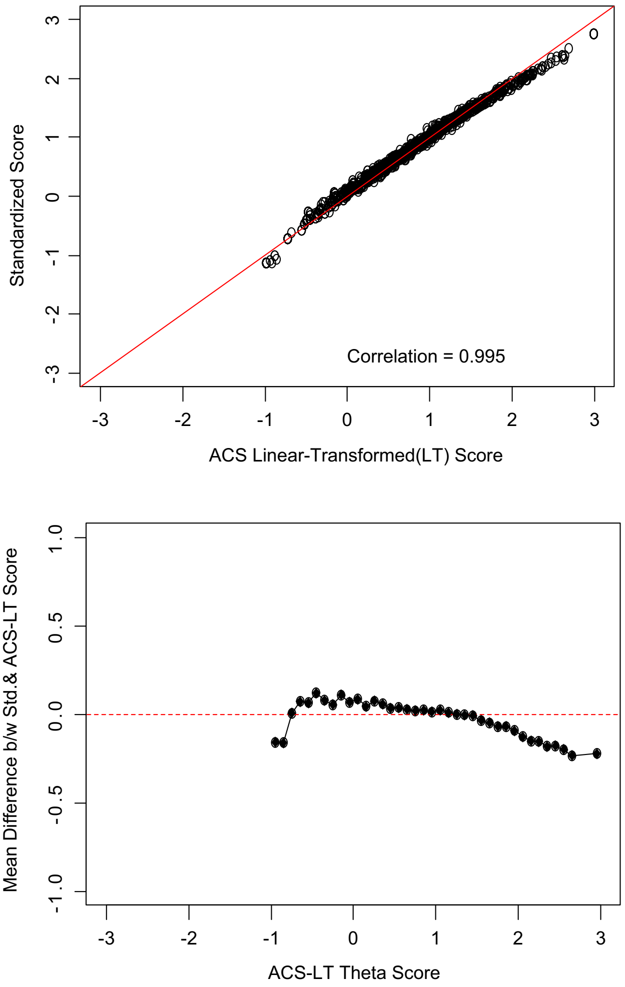


Fig. 4 *Scatter plot* of IRT-based theta scores of ACS patients using standardized and linearly transformed ACS-specific (ACS-LT) IRT item parameters for the QDIS-7 (*top*). Mean difference between theta scores using standardized versus ACS-LT IRT item parameters as a function of ACS-LT theta scores (*bottom*)

the original standardization sample that completed surveys via the Internet and an independent clinical sample that was interviewed after a month following hospitalization for ACS. We compared the consistency between the standardized (pre-calibrated in household) and the ACS-specific (freely calibrated in the independent sample) IRT item parameters, and evaluated the consequences of any differential functioning at various levels. While the ACS-specific item discrimination parameters were systematically lower than the standardized parameters, item threshold parameters and theta scores were consistent, supporting the future use of the standardized IRT item parameters in scoring the QDIS-7 for ACS patients.

PRO measures of QOL impact are increasingly relied upon in population health and clinical research. Current disease-specific QOL measures differ widely in item content and scoring and often do not represent content areas shown to differ between those with and without specific conditions. For example, the SAQ measure does not include items measuring limitations in role and social activities, commonly shown to be affected using general measures administered in studies of heart disease. Developed using advanced psychometric models, the QDIS enables more standardized comparisons of QOL impact attributed to a specific condition across conditions. If this standardization is justified, it could be a noteworthy advance with implications for disease-specific PRO measures in QOL research. For example, using the QDIS norming sample, which is a representative sample of the US chronically ill population ($N = 4{,}120$) in their naturally occurring proportions, it is possible to compare the QOL impact that ACS patients attributed to their condition to the impact for chronically ill adults in general [7]. When compared to QDIS norms, we found that, 1 month following hospital discharge, ACS patients reported, on average, worse QOL than about 80 % of the chronically ill US adults.

Our findings supported the invariance property of the pre-calibrated standardized IRT parameters in the independent ACS sample. Generally we observed very small differences between item and scale scores using the standardized vs. ACS-specific IRT item parameters. We did find consistently larger item slope (a) parameters in the standardization sample (Table 2), indicating that the items were more discriminating in the standardization sample than in ACS sample (also see the steeper curves in Fig. 2). This may be due to one or more different characteristics across the samples, for example, different wordings of disease attribution, number of items, and interviewer-versus self-administration. One implication of slope parameter differences is reflected in very high and low theta score estimates. For example, for a given ACS patient on the low end of the theta scale, the theta score estimate using the standardized item parameters was likely to be greater than that using the ACS-LT item parameters, but vice versa for patients on the high end of the theta scale (Fig. 4). The larger difference at the two ends is likely due to larger measurement errors for extreme theta values. Additionally, the standard errors of theta estimates of ACS patients calculated by test information function are likely to be underestimated if using the standardized item parameters. On the contrary, the item threshold (b) parameters were far more consistent across the samples, which explain the two sets of very highly correlated theta scores with a strong linear relationship.

We tested IRT parameter invariance via the DFIT and compared it with the logistic regression analyses. IRT invariance is increasingly important along with growing IRT and computerized adaptive testing (CAT) applications in PRO measures. One major advantage of IRT and CAT is allowing items administrated in different versions or lengths (e.g., long, short, adaptive, and static forms), maintaining score comparability. Different survey versions across samples present an additional DIF threat, making IRT invariance a more prominent question to investigate. In our study, the logistic regression method analyzed the same version of QDIS-7 across samples. In contrast, the DFIT evaluated IRT item parameters from two different versions (QDIS-49 vs. QDIS-7) across samples. Thus, the latter provided more stringent evidence for score comparability not only across samples but also across survey versions.

We found the DFIT provided additional information. First, it evaluates practical significance of IRT item parameters on score differences. DTF was computed as MSD between the summed scores using the two sets of IRT item parameters, which has the unit squared of the QDIS-7 summed score (mean = 15.6, SD = 7.1). In this study, DTF is 0.37 and its square root is 0.6, which is less than one score point as well as one-tenth of SD. Thus both the practical and statistical difference tended to be very small. We observed the same small differences for the theta scores. Secondly, the (non-)compensatory DIF indices provided by DFIT (namely NCDIF and CDIF) enable the evaluation of relative contribution of each item to scale score difference. Of note, we found an example of DIF offset; Item 2 actually eliminated scale-level score difference despite its relatively large item-level difference. Developers can utilize such information to cancel out DIF effects when it is not practical to develop a DIF-free scale [17]. Finally, since DFIT does not require item-level response data of both samples, it provides some flexibility for researchers who do not have item-level data of development sample, but can obtain their IRT item parameters from published literature (e.g., the PROMIS measures [35, 36]). It is noteworthy that DFIT has both advantages and limitations. One limitation is that DFIT requires item parameters from separate IRT calibrations, which, however, are not readily compared and need to be linked. In the case that an initial DFIT analysis identifies any item with DIF, it is recommended to have a two-stage linking, in which the scale is purified by a second-stage linking based on the remaining non-DIF items [46, 47]. Another limitation is the lack of generalizability of significance tests using the pre-determined cutoffs [48]. Newer methods were recently developed (e.g., the item parameter replication [IPR] method [49]), and they deserve more studies.

Our study utilized diverse samples recruited from two large NIH-funded projects. Given that there are other methods widely known for examining measurement invariance (e.g., the multi-group confirmatory factor analysis [37–39]), it would be desirable to compare the different methods. Nevertheless, we evaluated IRT item slope and threshold parameters, which are analogous to the factor loadings and intercepts in the CFA framework. Furthermore, we examined their joint impacts on item, scale, and person scores in the IRT framework, upon which the QDIS was developed and scored, which should provide a more direct evaluation than methods in other frameworks. Given that QDIS items were administered using different modes across the samples, it is noteworthy that although the literature reports little or no impact of different self-administered methods on item responses [40, 41], the interview mode has tended to yield more favorable responses, which may have important implications in interpreting the difference scores across samples and deserves future investigation.

In conclusion, the seven QDIS items administered in an independent clinical sample by telephone interviews showed lower IRT item discrimination parameters than in the self-administered standardization sample via the Internet. However, differences in their IRT threshold parameters and patient scores were small. Our findings support the use of QDIS-7 with the developer-provided standardized IRT item parameters for future studies comparing QOL impact attributed to coronary heart disease with the impact attributed to other chronic conditions.

Acknowledgments TRACE-CORE is supported by the National Institutes of Health National Heart, Lung, and Blood Institute (1U01HL105268). DICAT is supported by the National Institutes of Health National Institute of Aging (2R44AG025589). Partial salary support was provided by TRACE-CORE and PhRMA foundation Research Starter Grant (M.D.A.). Additional support was provided by NIH Grant KL2TR000160 (M.E.W.). The authors are very grateful for the editor and reviewers' comments and personal communication with Jakob Bjørner.

Appendix 1: illustration of calculations

Item response theory indeterminacy

The logit function in an IRT model is defined by

$$Logit\left[P_i\left(\theta_j\right)\right] = a_i(\theta_j - b_{i)}$$

where $P_i(\theta_j)$ is the probability of endorsing item response category i for person j, a_i and b_i are the item discrimination and threshold parameters for response category i, respectively, and θ_j is the IRT-based theta score for person j. The logit will preserve the same value if a_i, θ_j , and $\mathrm{b_i}$ were replaced by $a_i^{'}$, $\theta_j^{'}$, and $b_i^{'}$, respectively, which follow a set of linear transformations of

$$b_i^{'} = A * b_i + B$$

$$\theta_j^{'} = A * \theta_j + B$$

$$a_i^{'} = a_i / A$$

where A is the slope and B is the intercept of the linear transformations.

Slope (A) and intercept (B) of linear transformation using IRT theta scores

$$A = \sigma(\theta_{\mathrm{ST}})/\sigma(\theta_{\mathrm{ACS}})$$

$$B = \mu(\theta_{\mathrm{ST}}) - A * \mu(\theta_{\mathrm{ACS}})$$

where σ (θ_{ST}) and σ (θ_{ACS}) are the SD of the IRT scores of ACS patients using the standardized (θ_{ST}) and the ACS-specific (θ_{ACS}) IRT item parameters, respectively. μ (θ_{ST}) and μ (θ_{ACS}) are the means of the IRT scores.

Non-compensatory differential item functioning (NCDIF)

$$\mathrm{NCDIF}_i = \frac{\sum_{j=1}^{N} \left(S_{\mathrm{ST},i,j} - S_{\mathrm{ACS},i,j}\right)^2}{N}$$

where $S_{\mathrm{ST},i,j}$ is the IRT model-expected item response score of item i for respondent j using the standardized IRT item parameters. $S_{\mathrm{ACS},i,j}$ is the IRT model-expected item response score of item i for respondent j using the ACS-LT IRT item parameters. N is the sample size of ACS patients.

Test differential functioning (DTF)

$$\mathrm{DTF} = \frac{\sum_{j=1}^{N} \left(\mathrm{TS}_{\mathrm{ST},j} - \mathrm{TS}_{\mathrm{ACS},j}\right)^2}{N}$$

where $\mathrm{TS}_{\mathrm{ST},j}$ is the IRT model-expected scale(test) score for respondent j using the standardized IRT item parameters. $\mathrm{TS}_{\mathrm{ACS},j}$ is the IRT model-expected scale(test) score for respondent j using the ACS-LT IRT item parameters. N is the sample size of ACS patients.

Compensatory differential item functioning (CDIF)

$$d_i(j) = S_{\mathrm{ST},i,j} - S_{\mathrm{ACS},i,j}$$

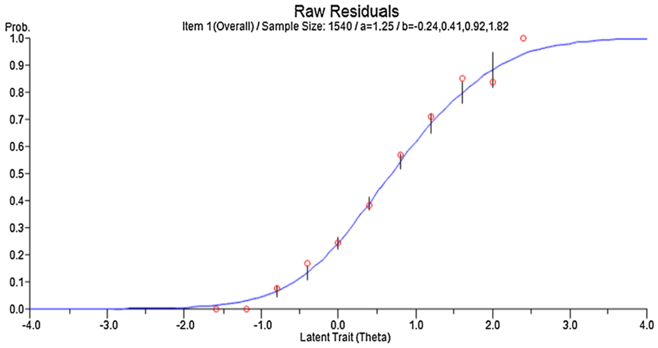
$$D(j) = \mathrm{TS}_{\mathrm{ST},j} - \mathrm{TS}_{\mathrm{ACS},j}$$

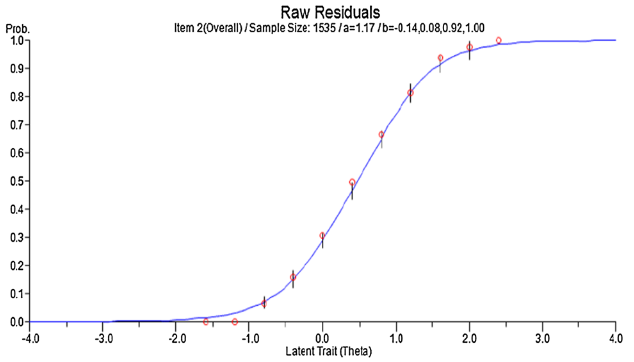
$$\mathrm{CDIF}_i = \mathrm{COV}(d_i, D) + \mu(d_i) * \mu(D)$$

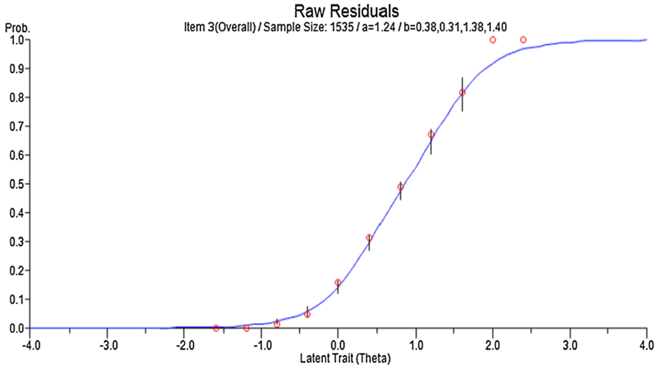
where d_i is the difference in the IRT model-expected item response score of item i between the standardized and the ACS-LT IRT item parameters. D is the difference in IRT model-expected scale score between the standardized and the ACS-LT IRT item parameters. COV(d_i, D) is the covariance of d_i and D. μ (d_i) and μ (D) are their means, respectively. Of note that DTF is equivalent to the sum of CDIF_i added across all the items: $\mathrm{DTF} = \sum_{i=1}^{I} \mathrm{CDIF}_i$.

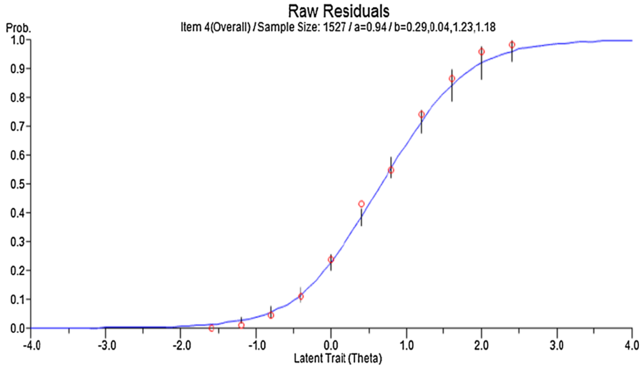
Appendix 2

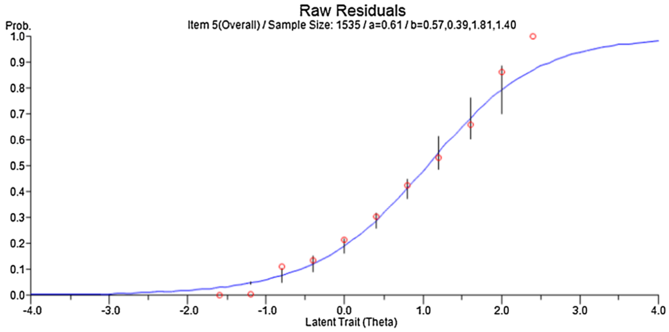
Raw residual plots of fitting the IRT model in the ACS patients for the QDIS-7 (ordered by row from Item 1 to Item 7)

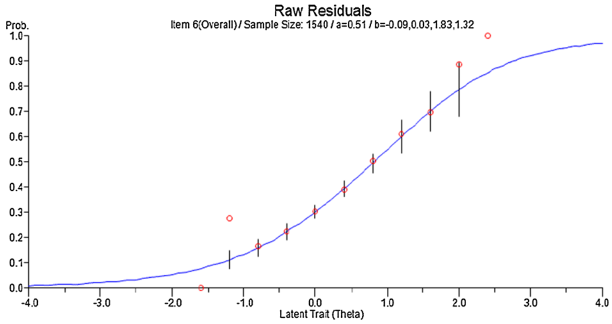


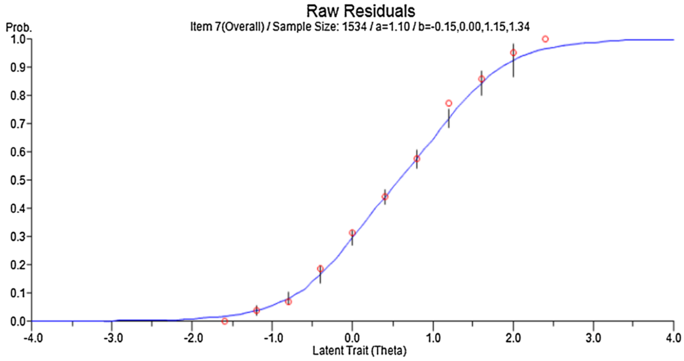












References

1. Ware, J. E, Jr, Kemp, J. P., Buchner, D. A., et al. (1998). The responsiveness of disease-specific and generic health measures to changes in the severity of asthma among adults. *Quality of Life Research, 7*(3), 235–244.
2. De Boer, A. G., Spruijt, R. J., Sprangers, M. A., & de Haes, J. C. (1998). Disease-specific quality of life: is it one construct? *Quality of Life Research, 7*(2), 135–142.
3. Spertus, J. A., Winder, J. A., Drewhurst, T. A., et al. (1995). Development and evaluation of the Seattle Angina Questionnaire: a new functional status measure for coronary artery disease. *Journal of the American College of Cardiology, 25*, 333–341.
4. Rector, T. S., Kubo, S. H., & Cohn, J. N. (1987). Patients' self-assessment of their congestive heart failure. Part 2: content, reliability and validity of a new measure, the Minnesota Living with Heart Failure questionnaire. *Heart Failure, 3*, 198–209.

5. Stewart, A. L., Greenfield, S., Hays, R. D., et al. (1989). Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. *JAMA, 262*, 907–913.
6. Ware, J. E, Jr, Harrington, M., Guyer, R., & Boulanger, R. (2012). A system for integrating generic and disease-specific patient-reported outcome (PRO) measures. *Patient Reported Outcomes Newsletter, 48*(Fall), 2–4.
7. Ware, J. E, Jr, Gandek, B., & Guyer, R. (2014). *Measuring disease-specific quality of life (QOL) impact: A manual for users of the QOL Disease Impact Scale (QDIS®)*. Worcester, MA: JWRG Incorporated.
8. Ware, J. E. Jr., Guyer, R., Gandek, B., Deng, N. *Standardizing disease-specific quality of life (QOL) impact measures: Development and initial evaluation of the QOL Disease Impact Scale (QDIS®)* (submitted).
9. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
10. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.
11. Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368.
12. Oldroyd, J. C., Cyril, S., Wijayatilaka, B. S., et al. (2013). Evaluating the impact of depression, anxiety & autonomic function on health related quality of life, vocational functioning and health care utilisation in acute coronary syndrome patients: the ADVENT study protocol. *BMC Cardiovascular Disorders, 13*, 103.
13. Kim, M. J., Jeon, D. S., Gwon, H. C., et al. (2013). Health-related quality-of-life after percutaneous coronary intervention in patients with UA/NSTEMI and STEMI: The Korean multicenter registry. *Journal of Korean Medical Science, 28*(6), 848–854.
14. Waring, M. E., McManus, R. H., Saczynski, J. S., et al. (2012). Transitions, risks, and actions in coronary events-center for outcomes research and education (TRACE-CORE): Design and rationale. *Circulation Cardiovascular Quality and Outcomes, 5*(5), e44–e50.
15. Howard, K. I., & Forehand, G. G. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement, 22*, 731–735.
16. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
17. Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27*, 43–50.
18. Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309–326.
19. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.
20. Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (pp. 683–718). London: Elsevier Publishing Co.
21. Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington: Degnon Associates.
22. Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating scale data [computer program]*. Chicago, IL: Scientific Software.
23. Liang, T., Han, K. T., & Hambleton, R. K. (2009). ResidPlots-2: Computer software for IRT graphical residual analyses. *Applied Psychological Measurement, 33*(5), 411–412.
24. De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.
25. Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.
26. Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking: Methods and practices* (2nd ed.). New York: Springer.
27. Sukin, Tia M. (2010). *Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods & accurate classification of examinees.* Doctoral dissertation. http://scholarworks.umass.edu/open_access_dissertations/301 Accessed 13 March 2014.
28. Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.
29. Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. Doctoral dissertation, Illinois Institute of Technology. *Dissertation Abstracts International, 54*, 2266.
30. Raju, N. (2000). *Notes accompanying the differential functioning of items and tests (DFIT) computer program*. Chicago: Illinois Institute of Technology.
31. Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., et al. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research, 16*, 43–68.
32. Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.
33. Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (Ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
34. Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241–256.
35. Reeve, B., Hays, R. D., Bjorner, J., et al., on behalf of the PROMIS cooperative group. (2007). Psychometric evaluation and calibration of health–related quality of life item banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care, 45*(5), S22–S31.
36. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supports the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology, 61*, 17–33.
37. Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.
38. Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–529.
39. Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(11 Suppl 3), S78–S94.
40. Wong, C. K., Lam, C. L., & Mulhern, B. (2013). Measurement invariance of the Functional Assessment of Cancer Therapy—Colorectal quality-of-life instrument among modes of administration. *Quality of Life Research, 22*, 1415–1426.

41. Bjørner, J. B., Rose, M., Gandek, B., et al. (2014). Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. *Journal of Clinical Epidemiology, 67*(1), 108–113.
42. 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.
43. Patrick, D. L., & Deyo, R. A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care, 27*(3 Suppl), S217–S232.
44. Ware, J. E, Jr, Guyer, R., Harrington, M., & Boulanger, R. (2012). Evaluation of a more comprehensive survey item bank for standardizing disease-specific impact comparisons across chronic conditions. *Quality of Life Research, 21*(1 Suppl), 27–28.
45. R Core Team (2014). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. http://www.R-project.org/.
46. Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253–260.
47. Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62*, 32–44.
48. Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 2*, 113–141.
49. Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1–17.
50. Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E, Jr. (2013). Difference in method of administration did not significantly impact item response: An IRT-based analysis from the Patient-Reported Outcomes Measurement Information System initiative. *Quality of Life Research, 23*, 217–227.
51. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R Package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of Statistical Software, 39*(8), 1–30.