Qual Life Res (2007) 16:981–990
DOI 10.1007/s11136-007-9200-x

Differential item functioning impact in a modified version of the Roland–Morris Disability Questionnaire

Paul K. Crane · Karynsa Cetin · Karon F. Cook · Kurt Johnson · Richard Deyo · Dagmar Amtmann

Received: 13 December 2006 / Accepted: 10 February 2007 / Published online: 19 April 2007

Abstract
Objective To evaluate a modified version of the Roland–Morris Disability Questionnaire for differential item functioning (DIF) related to several covariates.
Background DIF occurs in an item when, after controlling for the underlying trait measured by the test, the probability of endorsing the item varies across groups.
Methods Secondary data analysis of two studies of participants with back pain (total $n = 875$). We used a hybrid item response theory/ logistic regression approach for detecting DIF. We obtained scores that accounted for DIF. We evaluated the impact of DIF on individual and group scores, and compared scores that ignored or accounted for DIF in terms of the strength of association with SF-36 subscale scores.
Results DIF was found in 18/23 items. Salient scale-level differential functioning was found related to age, education, and employment. Overall 24 participants (3%) had salient scale-level differential functioning. Mean scores across demographic groups differed minimally when accounting for DIF. The strength of association of scores with SF-36 scores was similar for scores that ignored and scores that accounted for DIF.
Conclusions The modified version of the Roland–Morris Disability Questionnaire appears to have largely negligible DIF related to the covariates assessed here.

P. K. Crane (✉) · R. Deyo
Department of Medicine, University of Washington, Harborview Medical Center, 325 Ninth Avenue, Box 359780, Seattle, WA 98104, USA
e-mail: pcrane@u.washington.edu

K. Cetin · K. F. Cook · K. Johnson · D. Amtmann
Department of Rehabilitation Medicine, University of Washington, Seattle, WA 98195, USA

Keywords Differential item functioning · Item response theory · Logistic regression · Test bias

Abbreviations

2PL	2-parameter logistic model. In this parametric item response theory model, two parameters are modeled for each item: item difficulty and item discrimination
DIF	Differential item functioning. DIF occurs when an item has different statistical properties in different groups when controlling for the underlying trait or ability measured by the test
IRT	Item response theory. This is a technique for analyzing item-level test data based on the premise that item responses are a function of the relationship between an underlying latent trait and characteristics of the item
SIP	Sickness Impact Profile. This is a patient-reported outcome measure of the impact of illnesses
SLIP	Seattle Lumbar Imaging Project, one of the two datasets of low back pain subjects analyzed in this study

Introduction

Back pain is a major clinical and societal problem. Measurement of back pain-related disability is important for clinical research that evaluates treatment outcomes and may also be useful in clinical practice. It is important that outcomes are measured consistently across demographic groups. Confident comparison of scores across groups requires an assessment of scale bias. If a scale has bias,

observed scores for a group may be artificially high or artificially low, which may minimize or exaggerate actual differences between groups. In educational testing statistical evaluation of test bias has focused on evaluating items for differential item functioning (DIF) [1–3]. DIF occurs in a test item if different responses to that item are expected from different demographic groups after controlling for the underlying trait or ability measured by the test.

The original Roland–Morris Disability Questionnaire [4] was developed by a primary care physician selecting 24 items from the larger Sickness Impact Profile (SIP) [5] most relevant to low back pain. Patrick and colleagues [6] proposed a modified version after examining SIP data from two clinical trials, selecting items most likely to change over time. Nineteen items are in common between the two versions, but five items from the original are replaced by four new items in the modified version. Evidence regarding validity, reliability, and responsiveness to change of the modified version of the Roland–Morris Disability Questionnaire have been published [6].

Two previous studies have evaluated the Roland–Morris Disability Questionnaire for DIF. One study analyzed the Turkish version for DIF related to age, gender, duration of low back pain, severity of pain, and whether the assessment was at baseline or follow-up; they found DIF related to gender in two items and DIF related to timing of assessment in an additional two items [7]. A second study found three different items had DIF related to gender; removing these items reduced apparent gender differences in back pain-related disability [8].

In this study we examined data from two studies of participants with back pain to determine whether items from the modified Roland–Morris Disability Questionnaire exhibit DIF related to several covariates. Our specific goals were to determine whether DIF caused salient differences in scores, whether apparent differences in scores between demographic groups are due to DIF, and whether IRT scores that account for DIF have stronger relationships with SF-36 scores than scores that ignore DIF.

Materials and methods

Participant selection

Data were analyzed from two studies; detailed methods of both studies have been published [9, 10]. The first study was a multi-center prospective cohort study of 495 patients with presumed discogenic back pain (''the Discogenic study'') [9]. Participants included those who had one- or two-level disc degeneration confirmed by imaging and normal neurological evaluations. The second study was a trial of 380 participants with low back pain randomly assigned to rapid magnetic resonance imaging or standard radiographs (the Seattle Lumbar Imaging Project, ''SLIP'') [10]. Data analyzed here are from the baseline assessment of the two studies. Both studies included identical modifications of the Roland–Morris Disability Questionnaire as the primary outcome, and also included the SF-36 [11].

Statistical analyses and item response theory (IRT) calibration

Demographic characteristics of participants in the two studies were compared using χ^2 statistics for categorical analyses and unpaired 2-sample t tests for continuous variables; all statistical analyses were performed using Stata 8.0 [12]. Modified version of the Roland–Morris Disability Questionnaire scale items were analyzed with the 2-parameter logistic model (2PL) using Parscale version 4.1 [13]. In the 2PL, both item difficulty (the amount of back pain disability associated with a 50% probability of endorsing an item) and discrimination (ability of the item to discriminate back pain disability levels immediately above and below the item difficulty level) may vary. We employed expected a posteriori scoring; this form of scoring permits a finite score to be estimated for participants who endorsed all of (or none of) the items.

Covariates examined for DIF

We evaluated each modified version of the Roland–Morris Disability Questionnaire scale item for DIF related to 8 covariates: gender, age (dichotomized both at age 65 and 60 years), education (categorized into five groups: less than high school graduate; high school graduate; some college; completed college; some graduate school), marital status (categorized as living alone vs. living with someone else), study (the Discogenic study vs. SLIP), employment status (categorized as those who were on leave or disability vs. those who were not on leave or disability), history of back surgery, and self-rated overall health (categorized as excellent or very good vs. good, fair, or poor). We evaluated DIF related to study to determine whether it was appropriate to combine the data across the two studies to investigate other sources of DIF.

DIF analyses

We have developed a hybrid item response theory / logistic regression approach to detecting DIF. Detailed methods of this approach have been published [14], as have comparisons of this method to other methods of detecting DIF [15, 16]. Complete description of these methods is provided in Appendix 1. Briefly, unadjusted IRT scores are used ini-

tially to evaluate items for DIF. Item responses are analyzed using logistic regression, with terms for back pain disability, demographic group, and the interaction between back pain disability and demographic group. Two types of DIF are identified in the literature. Items with non-uniform DIF are identified by examining the statistical significance of the interaction term; we used a criterion of $\alpha = 0.05$ in this study. Uniform DIF is assessed by examining the proportional change in the regression coefficient for back pain disability from models that include and exclude the group term. Items with regression coefficients that changed by at least 5% were identified as having uniform DIF in this study. We chose relatively sensitive cutoff values for identification of items with DIF; for a discussion and comparison of different criteria for DIF see Refs. [14, 17, 18].

Following initial identification of items with DIF related to a covariate, we generated IRT scores that accounted for DIF by using demographic-specific item parameters (see Fig. 1). We handled spurious false-positive and false-negative DIF by using an iterative approach for each covariate. We re-examined each item for DIF using the IRT scores that used demographic-specific item parameters for items found with DIF on the previous round. If different items are identified with DIF, we repeated the process, determining demographic-specific item parameters for items most recently found to have DIF. We repeated these steps until the same items are identified with DIF on successive rounds. We have modified this approach for covariates with more than two categories; see Appendix 1 for details. We analyzed items for DIF related to each covariate separately, and for DIF related to all covariates simultaneously as described in detail elsewhere [14, 16].

The median standard error of measurement served as an indicator of measurement noise. We calculated differences between IRT scores that accounted for DIF and IRT scores that ignored DIF. Differences larger than the median standard error of measurement were judged to have ''salient scale-level differential functioning'' [16]. We determined whether each covariate was associated with

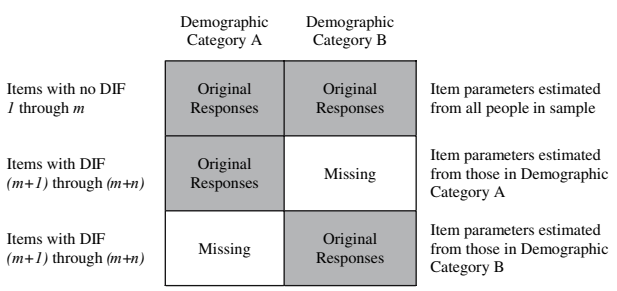

Fig. 1 Handling of m items found without DIF and n items found with DIF

salient scale-level differential functioning, and determined whether the sources of DIF were additive by evaluating IRT scores that accounted for all sources of DIF simultaneously.

We compared mean scores across demographic groups to determine whether accounting for DIF led to differences observed between groups. We performed a number of analyses to compare the following three scores to each other: the standard modified version of the Roland–Morris Disability Questionnaire score (obtained by totaling the number of items endorsed), the unadjusted IRT score (without accounting for DIF), and the IRT score that accounted for all sources of DIF. To compare the relative validity of the three different scoring strategies, we calculated the amount of variance in each of set of scores explained by the SF-36 composite scores: physical functioning, role performance, bodily pain, general health, vitality, social function, role-emotional, and mental health. We regressed each score against demographics alone, and then against demographics and the SF-36 subscale scores, to determine the amount of variance explained by the SF-36 subscale. This value provides an estimate of the strength of relationship between the modified version of the Roland–Morris Disability Questionnaire and each SF-36 subscale; greater strength of relationship would imply higher levels of concurrent validity.

Results

Across the two studies, 740 (85%) were white, 71 (9%) were African-American or black, 18 (2%) were Asian, and 24 (3%) were of Hispanic origin. Because of small numbers in ethnic groups other than whites, we were not able to evaluate items for DIF related to ethnicity. Other demographic characteristics of the participants in the two studies are shown in Table 1. Participants in the Discogenic study were more likely to be younger, to be married, to be on leave or disabled, and to have had a history of surgery on their back, but less likely to have had graduate school education. Gender and self-rated health status were not significantly different between participants in the two studies.

Results from the DIF analyses are shown in Table 2. None of the items had DIF related to gender. For each of the other covariates, at least one item was found to have DIF. Nine of the items had DIF related to employment status, six items had DIF related to age, six items had DIF related to study source, and five items had DIF related to education.

The scale-level impact of accounting for DIF on individuals' scores is shown in Fig. 2. We obtained slightly different results when we dichotomized age at 65 vs. 60 years; results for both cutpoints are shown. The vertical lines in Fig. 2 indicate the median standard error of

Table 1 Demographic characteristics of participants in the two studies

Characteristics	Discogenic (n = 495)		SLIP (n = 380)		p value
	n	%[a]	n	%[a]	
Gender—Female	259	52.3	236	47.7	0.35
Age[b]					<0.001
18–39	175	35.4	65	17.1	
40–49	212	42.9	91	24.0	
50–59	93	18.8	100	26.3	
60+	14	2.8	124	32.6	
Education					<0.001
Less than high school	35	7.2	25	6.6	
High school	98	20.2	71	18.7	
Some college	198	40.8	126	33.2	
College graduate	115	23.7	83	21.9	
Graduate school	39	8.0	74	19.5	
Marital status					<0.001
Single	72	14.9	203	53.7	
Living with significant other	32	6.6	18	4.8	
Married	270	55.8	90	23.8	
Separated/divorced	101	20.9	22	5.8	
Widowed	9	1.9	45	11.9	
Employment status					<0.001
Full time	188	39.6	169	46.3	
Part time	49	10.3	34	9.3	
Leave	87	18.3	8	2.2	
Disability	75	15.8	40	11.0	
Unemployed	50	10.5	8	2.2	
Homemaker, student, retired	26	5.5	106	29.0	
Surgical status	112	22.9	23	6.2	<0.001
Self-rated general health					0.08
Excellent	57	11.6	42	11.1	
Very good	136	27.7	103	27.2	
Good	187	38.1	118	31.1	
Fair	82	16.7	85	22.4	
Poor	29	5.9	31	8.2	

[a] Sums may not total 100% due to rounding

[b] In the Discogenic study, age was missing for one participant, education was missing for 10 participants, marital status was missing for 11 participants, employment status was missing for 20 participants, surgical history was missing for seven participants, and general health was missing for four participants. In the SLIP, education was missing for one participant, marital status was missing for two participants, employment status was missing for 10 participants, surgical history was missing for 10 participants, and general health was missing for one participant

measurement found in this study (0.313). When the difference between IRT scores accounting for DIF and scores ignoring DIF are greater than this value, they are said to have salient scale-level differential functioning. We found salient scale-level differential functioning related to age (dichotomized at 65 years only), education, and employment status. The scale-level impact of accounting for all sources of DIF on IRT scores for individual participants is shown at the bottom of Fig. 2. For the Modified Rolland scale, the effects of DIF related to the different covariates did not cancel out and, instead, appear to be additive, resulting in many more participants with salient scale-level differential functioning. In all, 24 participants (3%) had salient scale-level differential functioning when accounting for all sources of DIF.

We were interested in whether accounting for DIF would make a substantive difference in mean scores across demographic groups. The greatest impacts of accounting for DIF were in comparisons between participants who were disabled or on leave compared to those who were not

Table 2 Differential item functioning findings for items in the Modified Roland Scale*

Item	Content	Gender		Age		Education		Marital status		Study		Employment		Surgical status		General health	
		NU	U	NU	U	NU	U	NU	U	NU	U	NU	U	NU	U	NU	U
1	Stay at home most of time	0.84	0.00	0.64	–0.01	**<0.01**	–0.02	0.92	0.00	0.35	–0.02	0.53	**–0.05**	0.67	–0.01	0.71	–0.01
2	Change positions frequently	0.96	0.04	0.11	–0.01	0.65	0.01	**<0.01**	0.00	0.09	–0.04	0.89	–0.01	0.88	0.01	0.97	0.03
3	Walk more slowly than usual	0.68	0.00	0.25	0.03	0.26	–0.02	0.90	0.02	0.63	0.11	0.62	0.02	0.72	0.01	0.80	0.00
4	Not doing that jobs that usually do around the house	0.72	0.00	0.31	0.01	0.65	0.03	0.82	0.00	0.26	0.02	0.61	0.01	0.47	–0.01	0.43	0.02
5	Use a handrail to get up stairs	0.61	0.02	**0.03**	**0.07**	0.55	0.00	**0.01**	0.01	0.57	**0.05**	0.56	**0.09**	0.71	0.02	0.92	0.00
7	Hold onto something to get out of easy chair	0.18	0.00	0.69	0.02	**<0.01**	**0.08**	0.73	0.01	0.43	0.04	0.34	0.04	0.29	0.02	0.93	0.00
9	Get dressed more slowly	0.81	–0.01	0.46	0.00	0.81	**0.06**	0.51	0.00	0.80	0.04	**0.05**	**0.10**	0.16	0.01	0.27	**0.08**
10	Stand for short periods of time	0.07	0.02	0.70	0.02	0.35	–0.04	0.42	0.01	0.29	0.03	**<0.01**	**–0.05**	0.26	0.00	0.98	–0.04
11	Try not to bend or kneel	0.32	0.00	0.00	0.00	0.61	0.04	0.81	0.00	0.83	–0.02	0.80	0.01	0.48	0.02	0.52	0.00
12	Difficult to get out of chair	0.57	0.00	0.40	0.01	0.59	0.01	0.81	0.00	0.61	**0.05**	**<0.01**	**0.08**	0.88	0.03	0.41	0.00
13	Painful almost all the time	0.88	0.01	**0.01**	–0.01	0.47	–0.01	0.58	0.00	0.39	**–0.05**	0.11	0.03	0.47	–0.02	0.85	0.03
14	Difficult to turn over in bed	0.84	0.01	0.27	0.00	0.64	0.00	0.59	0.00	0.75	0.02	0.11	**0.09**	0.68	0.02	0.22	0.01
16	Trouble putting on socks	0.84	0.00	0.43	–0.01	0.13	0.04	0.89	0.00	0.35	0.00	0.21	**0.07**	0.09	0.01	0.52	**0.06**
17	Walk short distances	0.16	0.01	**0.03**	0.03	0.11	–0.03	0.19	0.01	**<0.01**	**0.07**	0.13	–0.01	**<0.01**	0.01	0.61	–0.03
18	Sleep less well	0.50	0.02	0.77	–0.01	0.91	–0.01	**<0.01**	0.00	0.14	**–0.06**	**0.02**	–0.02	**0.04**	0.00	0.40	0.03
21	Avoid heavy jobs around the house	0.44	0.02	0.17	0.01	0.76	0.04	0.10	0.00	0.14	–0.02	0.90	0.02	0.23	–0.01	0.46	0.02
22	More irritable and bad-tempered	0.11	0.01	0.07	–0.02	**<0.01**	0.02	0.70	0.00	0.57	**–0.06**	0.39	0.02	0.61	0.00	0.93	0.01
23	Go upstairs more slowly	0.46	0.01	0.64	**0.11**	0.38	0.01	0.33	0.01	0.32	**0.12**	0.84	0.02	0.70	0.00	0.87	–0.01
24	Stay in bed most of the time	0.42	0.01	0.32	–0.01	0.89	0.01	0.75	0.00	0.68	0.00	0.28	0.01	0.39	0.00	0.26	0.02
M1	Sexual activity has decreased	0.33	0.00	0.55	–0.02	0.60	0.02	0.80	0.01	0.67	–0.04	0.45	–0.04	0.85	–0.01	0.53	0.03
M2	Keep rubbing or holding areas	0.73	0.04	0.84	0.01	0.34	0.02	0.12	0.00	0.85	–0.01	0.38	**0.05**	0.65	–0.01	0.54	0.00
M3	Doing less of the daily work around the house	0.57	0.03	0.44	–0.02	0.64	**0.11**	0.19	0.00	0.69	–0.01	0.13	0.02	0.77	–0.01	0.94	–0.01
M4	Express concern to other people	0.49	0.00	0.64	–0.01	0.79	–0.04	0.07	0.00	0.30	0.02	0.17	–0.03	0.27	–0.01	**0.04**	**–0.08**

*Numbers in the uniform DIF columns (''U'') represent proportional change in β coefficients from models 2 and 3. Gray shading indicates a proportional change as large or larger than 0.05. Numbers in the non-uniform DIF columns (''NU'') represent p values for interaction terms. Gray shading indicates a p value <0.05. See ''Materials and methods'' section for details

disabled or on leave (Tables 3, 4). Accounting for DIF decreased the difference between groups by 0.07 points; ignoring DIF related to employment status led to overestimating the difference between groups by approximately 10%. The p value of the difference between groups was not affected by accounting for DIF (<0.0001 for the unadjusted score, the IRT score accounting for DIF related to employment status alone, and the IRT score accounting for all sources of DIF). For other demographic covariates, ignoring DIF had even less of an impact on the observed differences between demographic groups (data not shown).

The scores from the three different strategies (total score, unadjusted IRT score, and IRT score accounting for all sources of DIF) were highly correlated. The correlation of the standard score with the unadjusted IRT score was 0.98, the correlation of the standard score with the IRT score accounting for all sources of DIF was 0.97, and the correlation of the unadjusted IRT score and the IRT score accounting for all sources of DIF was 0.99.

We compared the amount of variance explained for each of the three scores by each of the eight SF-36 composite scores. These results are shown in Table 5. There were few differences between the three scores in the amount of variance explained by the SF-36 subscales. The standard score had 1–2% more of its variance explained by the physical functioning, role-physical, bodily pain, vitality, and role-emotional domain scores than either of the IRT scores.

Discussion

We found that 18 of the 23 items (78%) in the modified version of the Roland–Morris Disability Questionnaire had

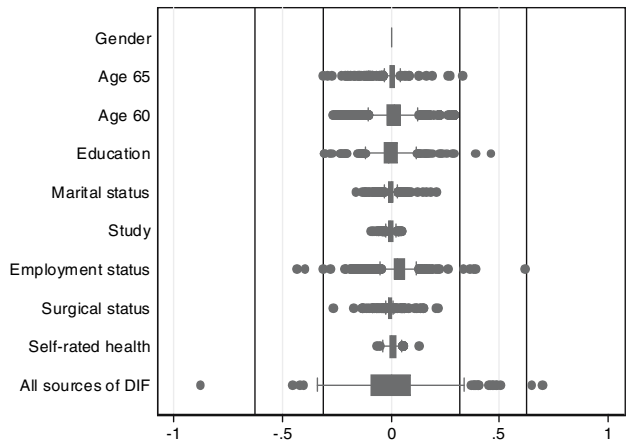

Fig. 2 Impact of DIF related to eight covariates on estimated modified Roland IRT score. In this box and whisker plot, the box indicates the 25th and 75th percentiles, while the whiskers indicate 1½ times the interquartile range. Observations more extreme are indicated by dots. The graph shows the difference between scores accounting for DIF for each covariate and unadjusted scores. If there were no impact of DIF for an individual that observation would be at 0. Vertical lines placed at multiples of 0.313 and –0.313 indicate the median standard error of measurement found in this study. The three covariates for which participants have differences greater than the median standard error of measurement (age, education, and employment) are said to be associated with salient scale-level differential functioning

DIF related to at least one covariate, and all of the covariates except gender were associated with DIF in at least one item. The covariates associated with the largest number of items with DIF were employment status, study source, age, and education. Education and employment status were associated with salient scale-level differential functioning. In contrast, despite having five items with DIF, study source (SLIP vs. the Discogenic study) was not associated with salient scale-level differential functioning, supporting our decision to combine data from the two studies for analyses. The numbers of participants with IRT score changes when accounting for DIF increased when we accounted for all sources of DIF, but still only 3% of participants had salient scale-level differential functioning when accounting for all sources of DIF. Accounting for DIF had negligible impacts on mean scores across demographic groups. The greatest differences were for employment status, but accounting for DIF did not appreciably change the magnitude of differences found between groups. We found little difference between IRT scores that accounted for DIF and IRT scores that ignored DIF in terms of the strength of association with SF-36 domain scores. In all, we found that while most of the modified version of the Roland–Morris Disability Questionnaire items had DIF, the impact of this DIF was negligible. It may be useful to determine not only whether items in a scale have DIF but also the impact of that DIF.

Two publications have assessed the Roland–Morris Disability Questionnaire for DIF: one study evaluated DIF related to age, gender, duration of low back pain, severity of pain, timing of assessment [7], while the other focused on DIF related to gender alone [8]. The analyses shown

Table 3 Demographic group summary scores of modified Roland scores[a]

Score type	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)	Difference (*p* value*)
	Males		Females		
Adjusted for all sources of DIF	–0.06 (0.95)	–0.02 (–0.69–0.65)	–0.02 (0.95)	–0.02 –0.69–0.67)	0.04 (0.53)
	Less than 60		60 and older		
Adjusted for all sources of DIF	0.01 (0.95)	0.01 (–0.64–0.70)	0.32 (0.91)	–0.30 (–0.96–0.36)	0.31 (0.0004)
	Less than 65		65 and older		
Adjusted for all sources of DIF	0.00 (0.94)	0.01 (–0.67–0.69)	–0.38 (0.93)	–0.42 (–0.98–0.33)	0.39 (0.0003)
	Living alone		Living with someone else		
Adjusted for all sources of DIF	–0.09 (0.98)	–0.09 (–0.74–0.65)	0.01 (0.91)	0.02 (–0.64–0.68)	0.10 (0.12)
	Discogenic study		SLIP		
Adjusted for all sources of DIF	0.18 (0.90)	0.20 (–0.44–0.85)	–0.32 (0.94)	–0.38 (–0.94–0.35)	0.50 (<0.0001)
	Not disabled or on leave		Disabled or on leave		
Unadjusted	–0.25 (0.92)	–0.22 (–0.86–0.38)	0.57 (0.85)	0.69 (0.10–1.27)	0.82 (<0.0001)
Adjusted for DIF related to employment	–0.21 (0.93)	–0.19 (–0.84–0.44)	0.54 (0.82)	0.65 (0.02–1.22)	0.75 (<0.0001)
Adjusted for all sources of DIF	–0.23 (0.92)	–0.23 (–0.86–0.40)	0.51 (0.82)	0.63 (0.01–1.11)	0.75 (<0.0001)
	No surgical history		Surgical history		
Adjusted for all sources of DIF	–0.12 (0.95)	–0.11 (–0.76–0.62)	0.38 (0.85)	0.40 (–0.06–1.03)	0.50 (<0.0001)
	Excellent or very good health		Good, fair, or poor health		
Adjusted for all sources of DIF	–0.35 (0.97)	–0.44 (–0.97–0.25)	0.16 (0.88)	0.17 (–0.45–0.84)	0.51 (<0.0001)

[a] 2-sample *t* tests

Table 4 Demographic group summary scores for education of modified Roland scores[a]

	Less than high school		High school		Some college		College graduate		Graduate school	
	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)
Adjusted for all sources of DIF	0.55 (0.84)	0.69 (0.07–1.07)	0.18 (0.86)	0.18 (–0.36–0.82)	0.03 (0.96)	0.04 (–0.63–0.80)	–0.30 (0.87)	–0.36 (–0.88–0.27)	–0.42 (0.99)	–0.44 (–1.12–0.35)

[a] 2-sample *t*-tests $p = 0.004$ for less than high school vs. high school; 0.08 for high school vs. some college; 0.0001 for some college vs. college graduate; and 0.26 for college graduate vs. graduate school

here are the most comprehensive to date, including evaluations of items for DIF with respect to eight different covariates as well as assessment of the scale-level impact of DIF. In addition to comparing means across groups when accounting for DIF to determine whether group differences may be due to DIF (as did Pietroban et al. [8]), we also compared the strength of association between back pain disability scores and an external criterion, the SF-36 subscales.

In contrast to the earlier studies, we found no items with DIF related to gender. When we modified our standard criterion for declaring an item to have uniform DIF from a change in parameter criterion to a statistical significance criterion, we found five items had DIF related to gender (results not shown). However, accounting for these five items led to no salient changes in scores—correlations between the scores were 0.9999, and the participants most affected by accounting for DIF related to gender had scores that changed by approximately 0.05 standard deviations, about 1/6 of the median standard error of measurement used to determine if changes are salient. The finding that items had DIF related to gender in the earlier studies but not in the present analysis may be related to use of a statistical significance criterion in large data sets [17]. It should also be recalled that different specific items were found with DIF related to gender in the two previous publications [7, 8].

When Pietrobon et al. found DIF in several items related to gender, they compared differences between males and females for scores that included all of the items and scores that removed the items found to have DIF [8]. In our study we performed similar analyses, but used demographic specific item parameters for items found to have DIF. Especially when dealing with DIF due to many sources simultaneously, the demographic-specific item parameter approach may be preferred, because eliminating all items found with any source of DIF could lead to eliminating a sizable proportion of the available items [19]. We found DIF in all but five items from the modified version of the Roland-Morris Disability Questionnaire. Limiting total scores to the five items found to have no DIF would result in markedly impaired measurement variability and precision.

We found little difference in the amount of variance explained by SF-36 subscale scores between scores that ignored DIF and IRT scores that accounted for DIF. Similarly, in a test of shoulder functioning, we found very few items with DIF, and those with DIF were not associated with salient scale-level differential functioning [14].

The SF-36 subscales may have DIF that was not accounted for in these analyses. Perkins and colleagues analyzed the subscales of the SF-36 for DIF related to gender, age, education, and race (African-American vs.

Table 5 Amount of variance explained by SF-36 domain scores of the total score, the unadjusted IRT score, and the IRT score accounting for all sources of DIF

	Standard score			Unadjusted IRT score			IRT score accounting for all sources of DIF		
	Full model	Demographics only	Difference	Full model	Demographics only	Difference	Full model	Demographics only	Difference
Physical functioning	0.51	0.24	**0.26**	0.49	0.25	**0.24**	0.49	0.24	**0.25**
Role-physical	0.38	0.24	**0.14**	0.37	0.25	**0.12**	0.36	0.24	**0.12**
Bodily pain	0.44	0.24	**0.20**	0.43	0.25	**0.18**	0.42	0.24	**0.18**
General health	0.27	0.24	**0.02**	0.27	0.25	**0.02**	0.26	0.24	**0.02**
Vitality	0.32	0.24	**0.08**	0.32	0.25	**0.07**	0.31	0.24	**0.07**
Social function	0.40	0.24	**0.15**	0.40	0.25	**0.15**	0.39	0.24	**0.15**
Role emotional	0.29	0.24	**0.05**	0.30	0.25	**0.04**	0.29	0.24	**0.04**
Mental health	0.30	0.24	**0.06**	0.31	0.25	**0.06**	0.30	0.24	**0.06**

white) [20]. They found several items in the SF-36 to have DIF, including several items in the Physical Functioning, the Role-Physical, and the Bodily Pain domains, the same domains in which we found the greatest differences between the amount of variability explained by the SF-36 domains and the modified version of the Roland–Morris Disability Questionnaire scores. It may be that the higher amount of variability in the standard score explained by the SF-36 subscales is precisely because of DIF in the SF-36. When we account for DIF in the Modified Roland scale, we may have decreased the strength of association with the SF-36 because of residual bias in the SF-36.

An important limitation of this study is that there was limited ethnic heterogeneity among the participants in these two studies. Methodology similar to that employed here should be performed in a more representative sample before generalizing results to ethnic minorities.

In conclusion, we found that most of the modified version of the Roland–Morris Disability Questionnaire items had DIF related to at least one of the eight covariates. We found salient scale-level differential functioning related to education and employment status. IRT scores that ignored DIF and IRT scores that accounted for DIF were highly correlated with one another. Accounting for DIF did not markedly change differences between demographic groups. Further, accounting for DIF made little difference in the strength of association with SF-36 scores. Our findings suggest that the modified version of the Roland–Morris Disability Questionnaire may have DIF, but for most practical purposes, this DIF may be ignored without threatening the validity of results.

Acknowledgements Data were collected under the auspices of grants P60 AR48093 from the National Institutes of Health, National Institute for Arthritis, Musculoskeletal, and Skin Diseases, and HS-09499 from the Agency for Healthcare Research and Quality. Data were analyzed under the auspices of U01AR52171-01 from the National Institutes of Health, National Institute of Arthritis and Musculoskeletal and Skin Diseases. Data collection and analyses were reviewed by the University of Washington's Institutional Review Board.

Appendix 1

Detailed methods of DIF detection

We have developed an approach to DIF assessment that combines ordinal logistic regression and IRT. Details of this approach are outlined in earlier publications [14, 17]. The modified version of the Roland-Morris Disability Questionnaire contains only dichotomous items, so logistic regression was used for all DIF analyses.

We use IRT scores to initially evaluate items for DIF. We examine three models for each item for each demographic category (labeled here as ''group'') selected for analysis:

$$\text{Logit } p(Y=1|\theta,\ \text{group}) = \beta_1 * \theta + \beta_2 * \text{group} + \beta_3 * \theta * \text{group} \quad \text{(model1)}$$

$$\text{Logit } p(Y=1|\theta,\ \text{group}) = \beta_1 * \theta + \beta_2 * \text{group} \quad \text{(model2)}$$

$$\text{Logit } p(Y=1|\theta) = \beta_1 * \theta. \quad \text{(model3)}$$

In these equations, $p(Y = 1)$ is the probability of endorsing an item, θ is the IRT estimate of back pain disability, and group is the demographic category.

Two types of DIF are identified in the literature. In items with *non-uniform DIF*, demographic interference between ability level and item responses differs at varying levels of back pain disability. In items with *uniform DIF*, this interference is the same across all levels of back pain disability.

To detect non-uniform DIF, we compare the log likelihoods of models 1 and 2 using a χ^2 test, $\alpha = 0.05$. To detect uniform DIF, we determine the relative difference between the parameters associated with θ (β_1 from models 2 and 3) using the formula $|(\beta_{1(\text{model 2})} - \beta_{1(\text{model 3})})/\beta_{1(\text{model 3})}|$. If the relative difference is large, group membership interferes with the expected relationship between back pain disability and item responses. There is little guidance from the literature regarding how large the relative difference should be. A simulation by Maldonado and Greenland on confounder selection strategies used a 10% change criterion in a very different context [21]. We have previously used 10% [17] and 5% [14] change criteria. In this data set, we compared results for each covariate using a 5 and 10% criterion. While there was little difference between results using a 5 and 10% criterion, we chose to show the results from the more sensitive 5% criterion.

We have developed an approach to generate scores that account for DIF [14]. When DIF is found, we create new datasets as summarized in Fig. 1. Items without DIF have item parameters estimated from the whole sample, while items with DIF have demographic-specific item parameters estimated.

Spurious false-positive and false-negative results may occur if the back pain disability score (θ) used for DIF detection includes many items with DIF [2]. We therefore use an iterative approach for each covariate. We generate IRT scores that account for DIF, and use these as the back pain disability score to detect DIF. If different items are identified with DIF, we repeat the process outlined in Fig. 1, modifying the assignments of items based on the most recent round of DIF detection. If the same items are identified with DIF on successive rounds, we are satisfied that we identified items with DIF (as opposed to spurious findings).

We have modified this approach for demographic categories with more than two groups (such as education in this data set). Indicator terms for each group are generated, and interaction terms are generated by multiplying θ by the indicator terms. All indicator terms and interaction terms are included in model 1; all indicator terms are included in model 2; and only the ability term θ is included in model 3. For the determination of non-uniform DIF, we compared the likelihoods of models 1 and 2 to a χ^2 distribution with degrees of freedom equal to the number of groups minus 1. The determination of uniform DIF is unchanged, except all the group terms are included in model 2.

References

1. Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage.
2. Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
3. Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.
4. Roland, M., & Morris, R. (1983). A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. *Spine, 8*, 141–144.
5. Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The sickness impact profile: Development and final revision of a health status measure. *Medical Care, 19*, 787–805.
6. Patrick, D. L., Deyo, R. A., Atlas, S. J., Singer, D. E., Chapin, A., & Keller, R. B. (1995). Assessing health-related quality of life in patients with sciatica. *Spine, 20*, 1899–1908; discussion 1909.
7. Kucukdeveci, A. A., Tennant, A., Elhan, A. H., & Niyazoglu, H. (2001). Validation of the Turkish version of the Roland–Morris disability questionnaire for use in low back pain. *Spine, 26*, 2738–2743.
8. Pietrobon, R., Taylor, M., Guller, U., Higgins, L. D., Jacobs, D. O., & Carey, T. (2004). Predicting gender differences as latent variables: Summed scores, and individual item responses: A methods case study. *Health and Quality of Life Outcomes, 2*, 59.
9. Deyo, R. A., Mirza, S. K., Heagerty, P. J., Turner, J. A., & Martin, B. I. (2005). A prospective cohort study of surgical treatment for back pain with degenerated discs; study protocol. *BMC Musculoskeletal Disorder, 6*, 24.
10. Jarvik, J. G., Hollingworth, W., Martin, B., Emerson, S. S., Gray, D. T, Overman, S., Robinson, D., Staiger, T., Wessbecher, F., Sullivan, S. D., Kreuter, W., & Deyo, R. A. (2003). Rapid magnetic resonance imaging vs radiographs for patients with low back pain: A randomized controlled trial. *JAMA, 289*, 2810–2818.
11. Ware, J. E. Jr. (2000). SF-36 health survey update. *Spine, 25*, 3130–3139.
12. StataCorp (2003). *Stata statistical software: Release 8.0*. College Station, TX: Stata Corporation.
13. Muraki, E., & Bock, D. (2003). *PARSCALE for Windows version 4.1*. Chicago: SSI.
14. Crane, P. K., Hart, D. L., Gibbons, L. E., & Cook, K. F. (2006). A 37-item shoulder functional status item pool had negligible differential item functioning. *Journal of Clinical Epidemiology, 59*, 478–484.
15. Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care, 44*, S115–S123.
16. Crane, P. K., Gibbons, L. E., Narasimhalu, K., Lai, J. S., & Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. *Quality of Life Research, 16*, 101–114.
17. 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*, 241–256.
18. Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R., & Teresi, J. (2007). A Comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research* (in press).

19. Crane, P. K. (2006). Commentary on comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. *Quality of Life Research, 15*, 1117–1118.
20. Perkins, A. J., Stump, T. E., Monahan, P. O., & McHorney, C. A. (2006). Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Quality of Life Research, 15*, 331–348.
21. Maldonado, G., & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology, 138*, 923–936.