J Rehabil Med 2013; 45: 670–677

ORIGINAL REPORT

WHICH ROLAND-MORRIS DISABILITY QUESTIONNAIRE? RASCH ANALYSIS OF FOUR DIFFERENT VERSIONS TESTED IN A NORWEGIAN POPULATION

Margreth Grotle, PhD[1,2], Philip Wilkens, MChiro, PhD[3], Andrew M. Garratt, PhD[4], Inger Scheel, PhD[5] and Kjersti Storheim, PhD[1,3]

From the [1]FORMI, Clinic for Surgery and Neurology (C1), Oslo University Hospital, [2]Department of Physiotherapy, Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, [3]Department of Orthopaedics, Oslo University Hospital and University of Oslo, [4]Department of Quality Measurement and Patient Safety and [5]Global Health Unit, Norwegian Knowledge Centre for the Health Services, Oslo, Norway

Background: **The Roland-Morris Disability Questionnaire (RDQ) is one of the most frequently used and recommended outcome measures for patients with low back pain.**

Objective: **To examine the fit of data from 4 different versions of the RDQ to a Rasch model in a Norwegian sample of patients with chronic low back pain and degenerative lumbar osteoarthritis.**

Methods: **Patients with chronic low back pain and degenerative lumbar osteoarthritis completed the RDQ prior to treatment in a secondary healthcare clinic. Data were analysed using a dichotomous Rasch model.**

Results: **Of 250 included patients, 243 patients with a mean age of 48.5 years completed all 24 items of the RDQ. None of the 4 RDQ versions (the original 24-item, the 18-item versions of Williams and Stratford, and the 11-item of Stroud) were a unidimensional measure of disability due to low back pain. Items 3 and 23 were redundant and items 13 and 18 did not fit the Rasch model. Several items showed differential item functioning, indicating that the items performed differently in subgroups of the sample.**

Conclusion: **In the absence of consistent findings across studies that have evaluated the RDQ by Rasch analysis, caution should be exercised in the development and application of alternative versions of the RDQ.**

Key words: low back pain; outcome; Roland-Morris Disability Questionnaire; item response theory; Rasch analysis.

J Rehabil Med 2013; 45: 670–677

Correspondence address: Margreth Grotle, FORMI (Communication Unit for Musculoskeletal Disorders), Clinic for Surgery and Neurology (C1), Oslo University Hospital, Oslo University Hospital, Kirkeveien 166, NO-0407 Oslo, Norway. E-mail: margreth.grotle@medisin.uio.no

Accepted Jan 17, 2013; Epub ahead of print Jun 24, 2013

INTRODUCTION

Several questionnaires have been developed to evaluate disability in people with low back pain (LBP) (1, 2). The Roland-Morris Disability Questionnaire (RDQ) is one of the most frequently used back-specific measures. The RDQ has been translated into several languages. Many versions have been evaluated (1) and it has been recommended as a core outcome measure for this patient group (3). The RDQ assesses disability in daily living among patients with LBP, with 24 items in the original version (4), 18 items in two different versions (5, 6), and 11 items in a fourth version (7). In addition, there is a modified 23-item version developed for sciatica patients (8), with briefer sciatica versions comprising 11 (9) and 12 items (10).

The traditional metric properties, such as validity, reliability and responsiveness of the RDQ, have been described extensively in the literature, and are generally acceptable (1, 2, 11, 12). However, construct validity has mostly been examined using classic test theory. Item response theory and Rasch analysis have been increasingly applied in the field of patient-reported outcomes, and is considered a more appropriate method when assessing construct validity as it provides specific analyses of the unidimensionality (the extent to which items measure a single construct, e.g. disability due to LBP), item difficulty (the relative difficulty of the items when compared with each another), and person separation (the extent to which items distinguish between distinct levels of disability) (13). The few studies that have assessed the RDQ using item response theory and Rasch analysis found that there are misfitting items in the original 24-item version (7, 14–16). However, the misfitting items were not similar across the different studies, for example Garratt (14) found that items 1, 2, 15 and 19 did not fit a unidimensional construct, whereas a recent study of Davidson (16) found that items 9 and 17 did not fit the Rasch model. The lack of consistent findings might be due to the fact that the studies were carried out in different countries including Australia, Canada, Turkey and UK. If so, then this has important implications for cross-national comparisons and generalizability of RDQ scores. Therefore, it is important to further evaluate existing RDQ versions by using similar and appropriate methods, such as Rasch analysis.

The aim of this paper was to examine the fit of data from 4 different RDQ versions to a Rasch model when used in a Norwegian sample with chronic LBP and degenerative lumbar osteoarthritis: the original 24-item version (4), the 2 18-item versions (5, 6), and the 11-item version of the RDQ (7).

MATERIAL AND METHODS

Measurement

The RDQ was developed in the UK in the early 1980s (4), and was based on the items of greatest relevance for patients with LBP from the Sickness Impact Profile (17). The items cover a range of aspects of daily living and were made specific to LBP by adding the words "because of my back" (or similar rephrasing) to the statements. RDQ items have a time frame of "today", a yes/no response format, and summed scores range from 0 (no disability) to 24 (severe disability).

The Norwegian version of the original RDQ has been cross-culturally adapted for patients in primary and secondary healthcare and tested for measurement properties, including reliability, validity and responsiveness (18, 19). Each of the RDQ versions evaluated in this study was based on patient responses to the full 24-item version.

Data collection

A total of 250 patients aged between 25 and 75 years with non-specific chronic LBP taking part in a double-blinded randomized, placebo-controlled trial comparing glucosamine sulphate with placebo were asked to complete the RDQ prior to treatment (20). Inclusion requirements were primary complaint of LBP longer than 6 months, more LBP than leg pain, no influential comorbidity, an RDQ score of 3 or more at baseline, no previous spinal fracture or surgery, no symptomatic disc herniation or spinal stenosis. Subjects were recruited from general practitioners, chiropractors, physiotherapists and a newspaper advertisement. Patients were given the self-completed questionnaire that included the RDQ after giving informed consent to take part in the trial. They were asked to complete the questionnaire at home and return it in a reply paid envelope. The study was approved by the Regional Ethical Committee for Medical Research (Regional komite for medisinsk forskningsetikk, reference number 53-06028 1.2006.40) in Norway.

Statistical analysis

Data were analysed in SPSS (version 14) and RUMM2020. The frequency of responses, including missing data, for each item was assessed. A Rasch analysis was used to test the RDQ scores against a mathematical measurement model, which is a probabilistic model that tests the extent to which the observed pattern of person and item responses fits the pattern expected by the model (23, 24). The Rasch analysis comprised a series of fit statistics in RUMM2020, which were used to indicate if the data met model expectations. Since RDQ has a dichotomous response category the dichotomous model was used. This model assumes that as a patient's disability increase, it is more likely that the item will be approved by the patient. Patients and item scores are used to "calibrate" items on a logit scale. Items at one end of the scale are "easier", while items at the other end are more "difficult". The difficulty of individual items is determined by the frequency of endorsement.

Fit to the Rasch model was examined for the original 24-item version (4), the two 18-item versions (5, 6), and the 11-item version of the RDQ (7), and a number of statistical analyses were carried out (21, 22):

First, a summary statistic of overall fit of data to the model was given by a Bonferroni-adjusted χ^2 Item-Trait Interaction statistic. In RUMM2020 the χ^2 statistics compares the difference in observed values with expected values across groups representing different ability levels (called class intervals) across the trait to be measured, which, in this case, is disability (21). In the present study 3 class intervals were used. A *non-significant* probability value indicated that there was no substantial deviation from the model and that the hierarchical ordering of the items was consistent across all levels of the underlying trait (21, 22).

The person separation index, which is equivalent to Cronbach's alpha, provides an indication of how many groups of ability the test can discriminate amongst (21, 22). The higher the person separation index, the more groups the test is able to detect; values of 0.8 and 0.9 indicate that the scale can statistically discriminate between at least 2 and 3 groups, respectively.

The individual person fit and item fit were assessed by inspecting the mean and standard deviation (SD) of the fit residuals. A mean value of approximately 0 and SD of 1 were expected. Misfitting items were identified by fit residuals of greater than plus or minus 2.5 or a significant χ^2 probability value (21, 22). High negative residuals are normally interpreted to indicate the redundancy of an item, suggesting that the item is not adding any new information to the scale.

To assess potential bias across groups of respondents, differential item functioning (DIF) was assessed in relation to gender, age, work status (in work or not), and use of pain medication (yes/no). Continuous variables must be split into categories in DIF analyses, therefore age was split according to the mean value of 48 years (median value was also 48 years). In the DIF analyses the responses of the different subgroups (gender, age, etc.) were analysed across the 3 class intervals mentioned above, which represents low, moderate and high ability. Two types of DIF can be identified: uniform and non-uniform DIF. A uniform DIF occurs when there is a difference between the subgroups across all the class intervals; for example, one subgroup is displaying a consistently greater ability to confirm an item than the other subgroup (analysis of variance (ANOVA) main effect). A non-uniform DIF indicates that the ability differences are inconsistent across the subgroups (ANOVA interaction effect).

Finally, tests for potential multidimensionality in terms of both a paired and independent *t*-test procedure of the person estimates derived from subsets of positively and negatively loaded items (21, 22) were carried out. Two item subsets were created from items loading positively and negatively on the first residual factor in the principal components analysis. First, the component loadings of the two subsets were compared and a paired *t*-test was used to determine if the associated person estimates were significantly different from that for all items. If the person estimate was different between the subset and the full item scale, this would indicate a breach of the assumption of local independence and unidimensionality. Secondly, the independent *t*-test procedure compared the proportion of persons with significantly different person estimates at a 5% significance level.

RESULTS

Of the 250 patients included in the trial, 243 completed all 24 items of the RDQ. Their mean age was 48.5 years (SD 11.2) and 48.4% were women. Most of the patients (74.4%) were working. The mean total score of the 24-item RDQ was 9.5 (SD 4.2) and the median score was 9 (interquartile range 6) on the 0–24 scale. Table I shows the endorsement frequencies, which reflect the difficulty order of the 24 items, as indicated by the logit measure. Item 2 of the RDQ "I change positions frequently to try to get my back comfortable", had the highest endorsement of 90.8%, whereas item 24 "I stay in bed most of the time because of my back", had the lowest endorsement of 1.2%. Items at the top reflect low disability and items at the bottom reflect high disability due to LBP.

Table I also shows that 4 items had fit statistics outside the acceptable level of ±2.5 in the original 24-item version and in the 2 18-item versions; item 3 "I walk more slowly than usual because of my back" and 23 "Because of my back, I go upstairs more slowly than usual" were redundant, whereas item 13 "My back is painful almost all of the time" and 18 "I sleep less well because of my back" did not fit the model. Item 23 was also redundant in the 11-item Stroud version, and item 10 "I only stand up for short periods of time because of my back" did not fit the model.

Table I. *Response frequencies, missing data, logit measures and fit statistics for the Roland-Morris Disability Questionnaire versions (n=250) in the order of difficulty. Item order and mean location (SE) from most to least difficult*

		Original 24-item version				Williams 18-item		Stratford 18-item		Stroud 11-item	
Original item number	Item description Items listed according to location order	Frequency, % Mean (SE)	Missing, %	Logit measure[a] (error) Mean (SE)	Fit residual[b]	Logit measure[a] (error) Mean (SE)	Fit residual[b]	Logit measure[a] (error) Mean (SE)	Fit residual[b]	Logit measure[a] (error) Mean (SE)	Fit residual[b]
2	I change positions frequently to try to get my back comfortable	227 (90.8)	0	–3.36 (0.22)	1.39	–	–	–	–	–	–
13	My back is painful almost all of the time	184 (73.6)	0	–1.87 (0.15)	3.12*	–1.34 (0.15)	4.28*	–1.26 (0.15)	3.59*	–	–
16	I have trouble putting on my socks because of the pain in my back	159 (63.6)	0	–1.58 (0.15)	0.75	–1.04 (0.15)	0.63	–0.97 (0.15)	0.61	–0.97 (0.15)	0.73
21	I avoid heavy jobs around the house because of my back	152 (60.8)	2	–1.50 (0.15)	–0.88	–0.96 (0.15)	–0.81	–0.87 (0.15)	–0.87	–0.83 (0.15)	0.81
9	I get dressed more slowly than usual because of my back	149 (59.6)	0	–1.38 (0.15)	–0.69	–0.85 (0.15)	–1.10	–0.77 (0.15)	–1.24	–0.75 (0.15)	–0.73
18	I sleep less well because of my back	147 (58.8)	0	–1.23 (0.14)	2.70	–0.70 (0.15)	3.40	–0.63 (0.14)	2.94	–	–
6	Because of my back, I lie down to rest more often	143 (57.2)	0	–1.20 (0.14)	0.45	–0.66 (0.14)	0.79	–0.58 (0.14)	0.59	–	–
7	Because of my back, I have to hold on to something to get out of my easy chair	134 (53.6)	0	–1.10 (0.14)	–1.02	–0.56 (0.14)	–0.92	–0.47 (0.14)	–1.25	–0.46 (0.15)	–0.74
14	I find it difficult to turn over in bed because of my back pain	134 (53.6)	0	–1.06 (0.14)	0.60	–0.53 (0.14)	0.80	–0.45 (0.14)	0.63	–	–
11	Because of my back, I try nto to bend or kneel down	125 (50.0)	1	–0.90 (0.14)	0.62	–0.36 (0.14)	0.48	–0.27 (0.14)	0.62	–0.25 (0.15)	0.80
3	I walk more slowly than usual because of my back	114 (45.6)	0	–0.74 (0.14)	–2.68	–0.20 (0.15)	–2.85	–0.09 (0.14)	–2.62	–0.07 (0.15)	–1.58
22	Because of my back, I am more irritable and bad tempered with people than usual	110 (44.0)	0	–0.60 (0.15)	1.82	–	–	0.02 (0.15)	2.47	–	–
23	Because of my back, I go upstairs more slowly than usual	105 (42.0)	0	–0.56 (0.15)	–4.20*	–0.02 (0.15)	–4.20*	0.09 (0.15)	–4.05*	0.12 (0.15)	–3.75*
12	I find it difficult to turn over in bed because of my back pain	102 (40.8)	0	–0.38 (0.15)	–0.22	0.15 (0.15)	0.01	0.23 (0.15)	–0.30	0.31 (0.15)	–0.42
5	Because of my back, I use a handrail to get upstairs.	85 (34.0)	0	–0.11 (0.15)	–1.43	0.44 (0.15)	–1.37	0.53 (0.15)	–1.44	0.56 (0.16)	–0.30
8	Because of my back, I try to get other people to do things for me	79 (31.6)	0	0.01 (0.15)	–0.20	0.56 (0.15)	0.11	0.64 (0.15)	–0.05	–	–
10	I only stand up for short periods of time because of my back	71 (28.4)	1	0.13 (0.16)	1.27	0.67 (0.16)	1.46	0.75 (0.16)	1.37	0.76 (0.16)	3.56*
17	I only walk short distances because of my back pain	49 (19.6)	0	0.85 (0.18)	–0.75	1.42 (0.18)	–0.31	–	–	1.58 (0.19)	0.22
1	I stay at home most of the time because of my back	36 (14.6)	3	1.34 (0.20)	–1.56	1.93 (0.21)	–1.23	2.00 (0.21)	–1.23	–	–
4	Because of my back, I am not doing any of the jobs that I usually do around the house	32 (12.8)	1	1.43 (0.21)	–0.97	2.02 (0.21)	–0.57	2.08 (0.21)	–0.70	–	–
20	I sit down for most of the day because of my back	14 (5.6)	0	2.60 (0.30)	–0.80	–	–	–	–	–	–

Table I. *Contd.*

Original item number	Item description Items listed according to location order	Original 24-item version			Williams 18-item		Stratford 18-item		Stroud 11-item	
		Frequency, % Mean (SE)	Missing, %	Logit measure[a] (error) Mean (SE)	Fit residual[b]	Logit measure[a] (error) Mean (SE)	Fit residual[b]	Logit measure[a] (error) Mean (SE)	Fit residual[b]	Logit measure[a] (error) Mean (SE)	Fit residual[b]
19	Because of my back pain, I get dressed with help from someone else	6 (2.4)	0	3.20 (0.38)	0.71	–	–	–	–	–	–
15	My appetite is not very good because of my back pain	7 (2.8)	0	3.42 (0.41)	–0.58	–	–	–	–	–	–
24	I stay in bed most of the time because of my back	3 (1.2)	0	4.57 (0.67)	–0.71	–	–	–	–	–	–

[a]Logits of a greater magnitude represent increasingly difficult items or increasing disability due to low back pain.
[b]Fit residuals with greater than plus or minus 2.5 are marked; negative value indicate redundant items, positive value misfitting items.
*χ^2 probability value (p-value) <0.002 (0.05/24) for the original version, <0.0028 for the two 18-item versions, and <0.0045 for the 11-item version.
SE: standard error.

Fig. 1. Person-item location distributions for the 4 versions of the Roland-Morris Disability Questionnaire; 24-item original version at the top, 18-item William and Stratford in the middle, and 11-item Stroud at the bottom.

Table II. *Fit to the Rasch model of the 4 Roland-Morris Disability Questionnaire versions*

	Original 24-item version	Williams 18-item version	Stratford 18-item version	Stroud 11-item version
Total item χ^2 Item-Trait Interaction statistic	161.23	155.35	138.18	82.44
χ^2 probability	$p<0.001$	$p<0.001$	$p<0.001$	$p<0.001$
Separation index	0.77	0.76	0.75	0.73
Item fit residual <–2.5 (redundant items)	Item 3 (walk more slowly) Item 23 (go upstairs more slowly)	Item 3 (walk more slowly) Item 23 (go upstairs more slowly)	Item 3 (walk more slowly) Item 23 (go upstairs more slowly)	Item 23 (go upstairs more slowly)
Item fit residual >+2.5 (misfit items)	Item 13 (painful all the time) Item 18 (sleep less well)	Item 13 (painful all the time) Item 18 (sleep less well)	Item 13 (painful all the time) Item 18 (sleep less well)	Item 10 (stand for short periods)
Item fit χ^2 <Bonferroni adjusted *p*-value	Item 13 (painful all the time) Item 23 (go upstairs more slowly)	Item 3 (walk more slowly) Item 13 (painful all the time) Item 23 (go upstairs more slowly)	Item 13 (painful all the time) Item 23 (go upstairs more slowly)	Item 10 (stand for short periods)

Items 2, 15, 19, 20 and 24, which are not included in the other versions of the RDQ (Table I), were all localized at the lower end of the difficulty order, the only exception being item 2. Fig. 1 displays the person-item distributions for the 4 RDQ versions, showing person ability (upper part of the graph) mapped towards the item difficulty (lower part of the graph). When comparing the person-item distributions, only the original RDQ version has items covering the whole person distribution. For the Stroud 11-item version a large proportion of the lower and higher person distributions was not covered by the items.

The fit statistics in Table II shows that all the RDQ versions had probability values under 0.001, and hence none reflected a unidimensional instrument of disability. The Person Separation Reliability Index varied from 0.77 to 0.73, indicating an acceptable ability to discriminate amongst respondents with two different levels of disability. The individual person and item fit values of the 24-item version indicated a lower ability level of the persons (–0.83, SD 1.13) than difficulty level of the RDQ (0.00, SD 1.90).

Analysis of DIF for gender, age, work status, and use of pain medication showed that several of the items performed differently across these groups, and that the pattern was similar across the original RDQ, the Willams and Stratford versions (Table III). The 11-item Stroud version showed fewer DIFs than the other 3 (Table III). The most consistent findings across all the 4 versions were the DIF for gender in item 9 and the DIF for use of pain medication in item 11. Men consistently had a higher probability to affirm item 9 (get dressed more slowly), whereas people who did not use pain medication were more likely to affirm item 11 (try not to bend or kneel). The DIF for gender showing that females were more likely to affirm item 14 (difficult to turn over in bed) was consistent in 3 of the 4 RDQ versions. Also, the findings regarding DIF according to age groups were rather consistent across most of the RDQ versions. For example, people in the higher age group (≥48 years) had a higher probability to affirm items 5 and 23, whereas people younger than 48 years were more likely to affirm item 22. Furthermore, we found two uniform DIFs for work status. The most frequent non-uniform DIF was found for work status in items 13 and 14.

The paired *t*-tests were highly significant for all the 4 versions ($p<0.001$), whereas the proportion of persons with significantly different persons estimates based on the two-item subsets was 5.08% for the original RDQ, 12.2% for the William, 4.5% for the Stratford, and 0% for the Stroud version, respectively.

DISCUSSION

The main finding of this study is that none of the 4 RDQ versions (the original 24-item, the 18-item versions of Williams and Stratford, and the 11-item of Stroud) are a unidimensional measure of disability due to LBP when used in this Norwegian sample of patients with chronic LBP and degenerative lumbar osteoarthritis. Items 13 and 18 did not fit the Rasch model, whereas 3 and 23 were redundant.

The current study found that items 13 and 18 were consistently misfitting to the Rasch model in the 3 RDQ versions in which they appeared. According to the ICF classification (23), these two items represent "impairments" (pain and sleep, respectively) whereas the other items represent aspects of activity limitations. Previous studies have also found evidence of misfit for item 18 (15, 16). Item 10 of the 11-item Stroud version also showed a misfit to the model, which also was reported for the Turkish RDQ (15). However, there is only limited consistency in misfitting items across studies that have used Rasch analysis. Table IV shows that the greatest consistency was that items 15 and 19 were identified as misfitting in 3 of 5 studies (7, 14, 15).

It may be that differences in culture, sample characteristics and sample sizes have contributed to this lack of consistency in the fit of RDQ data in back pain populations from Australia, Norway, Turkey, USA and the UK. Although all the previous studies recruited patients from the primary care there are variations across these samples (Table IV). For example, the study in USA included patients with chronic pain in different body areas, of which LBP was reported by 36% of the total sample. The mixed population might explain why many items were misfitting in this study, which limits the comparison of this study with the others in Table IV. Furthermore, cultural differences might explain why patients in Turkey scored much higher on the original RDQ (mean of 15) than patients from Norway and the UK (mean of 9). In order to address the is-

Table III. *Differential Item Functioning (DIF) for gender and age in the 4 Roland-Morris Disability Questionnaire versions*

	Original 24-item version	Willliams 18-item version	Stratfords 18-item version	Strouds 11-item version
Gender DIF				
Uniform DIF	Item 9 (get dressed more slowly)[a] Item 14 (difficult to turn over in bed)[b]	Item 9 (get dressed more slowly)[a] Item 14 (difficult to turn over in bed)[b]	Item 9 (get dressed more slowly)[a] Item 14 (difficult to turn over in bed)[b]	Item 9 (get dressed more slowly)[a]
Non-uniform DIF	Item 16 (trouble putting on my socks)[a]	None	None	None
Age DIF				
Uniform DIF	Item 5 (use handrail)[c] Item 22 (more irritable/bad tempered)[d] Item 23 (go upstairs more slowly)[c]	Item 5 (use handrail)[c] Item 23 (go upstairs more slowly)[c]	Item 5 (use handrail)[c] Item 22 (more irritable/bad tempered)[d] Item 23 (go upstairs more slowly)[c]	None
Non-uniform DIF	Item 24 (stay in bed)[d]	Item 13 (painful all the time)[d] Item 14 (difficult to turn over in bed)[d] Item 18 (sleep less well)[d]	None	Item 23 (go upstairs more slowly)[c]
Work DIF				
Uniform DIF	Item 1 (stay at home)[e] Item 18 (sleep less well)[f]	Item 1 (stay at home)[e] Item 18 (sleep less well)[f]	Item 1 (stay at home)[e] Item 18 (sleep less well)[f]	None
Non-uniform DIF	Item 13 (painful all the time)[f] Item 14 (difficult to turn over in bed)[f]	Item 13 (painful all the time)[f]	Item 13 (painful all the time)[f] Item 14 (difficult to turn over in bed)[f] Item 23 (go upstairs more slowly)[e]	None
Use of pain medication DIF				
Uniform DIF	Item 11 (try not to bend or kneel)[h]	Item 1 (stay at home)[i] Item 11 (try not to bend or kneel)[h]	Item 11 (try not to bend or kneel)[h]	Item 11 (try not to bend or kneel)[h]
Non-uniform DIF	None	None	Item 3 (walk slowly)[i]	None

[a]Males have a higher probability to affirm this item; [b]Females have a higher probability to affirm this item; [c]People who are more than 48 years have a higher probability to affirm this item; [d]Younger people (<48) have a higher probability to affirm this item; [e]People who are out of work have a higher probability to affirm this item; [f]People in work have a higher probability to affirm this item (for item 18; only in the lower and upper end of the construct); [g]People who are out of work have a higher probability to affirm this item, except in the upper end of the construct; [h]People who do not use painkillers have a higher probability to affirm this item; [i]People who use painkillers have a higher probability to affirm this item.

sue of cross-cultural equivalence, one could explore DIF by country on merged data-sets from several countries (21, 24).

Although there was little overlap in misfitting items when compared with previous studies, the difficulty level of the items was very similar (14–16). The items 15, 19, 20 and 24, which were excluded in all the adapted versions (5–7), were all "easy" items reflecting increasing disability in the present study. Also, the excluded item 2 was a "hard" item reflecting little disability in the current study. In the current and comparable studies (14–16) the RDQ items tended to cluster around the middle of the scale of difficulty, with relatively few items at the extremes. Hence, it is ill advised to remove items at the extreme ends of the scale hierarchy, where few of the 24 items contribute in terms of describing the disability of patients. However, this is exactly what all the shortened versions of the RDQ have done. In particular, the exclusion of item 2 reflecting low disability seems inappropriate and as shown by both the current and previous studies (14–16), item 2 has high endorsement and hence is important for determining patients with very low levels of disability. Only the original 24-item version had items that covered the very low difficulty level. Fig. 1 indicates that there is poor targeting of the items in both the lower and upper end of the scale in all the RDQ versions. This finding is very similar to what Davidson et al. (16) found, and suggests that more items are needed in order to assess lower and higher levels of disability more appropriately than what is possible with today's RDQ versions.

We found that several items performed differently across subgroups of gender, age group, work status and use of pain medication. Both item 9 and item 14 showed DIF for gender and items 5, 22, and 23 for age. Only two of the previous studies have reported on DIF in their Rasch analyses. Similar to the present study, Davidson found a DIF for age in item 5 ("Because of my back, I use a handrail to get upstairs"), which showed that older persons were more likely to affirm this item than younger persons (16). Kucukdeveci et al. (15) found a DIF for gender, but that was in item 5, and not in items 9 or 14 as in the present study. Again, these findings might be explained by differences in culture and back pain populations across Australia, Norway, Turkey, USA and the UK. Furthermore, in the Turkish study they found no DIF by duration or severity of pain. As far as we know, none have explored DIF by work status and use of pain medication as we did in the present study.

Table IV. *Comparing sample characteristics and results across five different studies using Rasch analysis for the Roland-Morris Disability Questionnaire*

	Garratt et al. (2003) (14)	Kucukdeveci et al (2001) (15)	Stroud et al (2004) (7)	Davidson et al (2009) (16)	Current study
Type of back pain sample and clinical setting	Subacute and chronic LBP recruited from primary care to a randomized, controlled trial – UK	LBP in an outpatient clinic – Ankara, Turkey	Chronic pain, screened for admission to a multidisciplinary pain management programme – Washington, USA	Subacute and chronic LBP, physiotherapy clinics – Melbourne, Australia	Chronic LBP and degenerative lumbar osteoarthritis recruited from primary care to a randomized, controlled trial – Norway
Sample size	1,008	81	993	140	250
Age, years, mean (SD)	42.9 (SD not reported)	37.0 (10.6)	43.5 (SD 12.6)	51 (SD 17.0)	48.5 (SD 11.2)
Gender, % females	55%	63%	57%	66%	48%
Duration LBP	All >4 weeks	All >4 months with average duration of 4.6 years (SD 3.7)	36.2% of the patients had chronic LBP with average duration of 6.5 years (SD 8.3)	43% <6 weeks	All >6 months
Work status (employed full- or part-time)	Not reported	Not reported	41%	41%	73%
Sum score, mean (SD), and/or median when available	9.0 (4.1)	Median 15.0 (interquartile range 8)	Not reported	Not reported	9.5 (4.2) Median 9 (interquartile range 6)
Item fit					
1	Poor outfit but frequently endorsed	×	Poor fit	×	×
2	Misfit but most frequently endorsed	×	Poor fit	×	×
3	×	×	×	×	Poor fit, redundant
4	×	×	Poor fit	×	×
5	×	×	×	×	×
6	×	×	Poor fit	×	×
7	×	Poor outfit	×	×	×
8	×	×	Poor fit	×	×
9	×	×	×	Poor fit	×
10	×	Poor outfit but frequently endorsed	×	×	×
11	×	×	×	×	×
12	×	×	×	×	×
13	×	×	Poor fit	×	Poor fit to construct
14	×	×	Poor fit	×	×
15	Poor outfit but frequently endorsed	Poor outfit	Poor fit	×	×
16	×	×	×	×	×
17	×	×	×	Poor fit	×
18	×	Poor outfit	Poor fit	×	Poor fit to construct
19	Poor outfit but frequently endorsed	Poor outfit and seldom endorsed	Poor fit	×	×
20	×	×	Poor fit	×	×
21	×	×	×	×	×
22	×	×	Poor fit	×	×
23	×	×	×	×	Poor fit, redundant
24	×	×	Poor fit	×	×

SD: standard deviation; ×: items meeting Rasch criteria.

The weaknesses of the RDQ have been reported in a number of studies (1, 14). It has been argued that patients should contribute to item selection for instruments designed to assess health and quality of life in back pain (25). In contrast to a large number of specific instruments that are now available for different health problems, the content of the RDQ was not developed following input from patients including interviews or focus groups. Hence, the RDQ may lack content validity as a patient-reported outcome, since it may not adequately reflect the concerns of patients. The RDQ assesses disability, but other aspects of health and quality of life are important to patients with back pain (26). Moreover, the aspects of disability assessed by the RDQ may not concord with those of back pain patients, the content of the RDQ being based on the generic Sickness Impact Profile (17). Other criticisms levelled at the RDQ include the use of dichotomous items, which generally have lower levels of data quality and reliability than categorical rating scales with more response alternatives (14). They may also be less responsive to change (27, 28).

This study has some limitations. First, the material in the present study was recruited to a randomized, controlled trial, and the patients needed a confirmed diagnosis of degenerative lumbar osteoarthritis to be included in the study. Therefore, this sample might represent a slightly different group of back pain patients than the typical patient seeking care in primary healthcare. Secondly, the patients completed the original 24-item RDQ version, and each of the RDQ versions were extracted from this full data-set. Whether administration of the actual shorter versions would yield equivalent data cannot be demonstrated from this study.

The lack of consistent findings across studies means that caution should be exercised in developing new versions of the RDQ. If researchers want to continue to use the RDQ despite its weaknesses it is more appropriate to use the 24-item version so that scores can be compared across studies. Furthermore, it is important that researchers and clinicians are aware that the RDQ cannot be considered a unidimensional measure of disability due to LBP. The application of Rasch analysis to merged data from different countries might further our understanding of the performance of the RDQ.

In conclusion, in this sample of Norwegian patients, none of the 4 versions of the RDQ were found to be unidimensional according to the Rasch model. Several studies based on modern psychometric methods have identified problems with the instrument. There is considerable variation in misfitting and redundant items across different studies. In the absence of consistent findings across studies, caution should be exercised in the development and application of alternative versions of the RDQ.

ACKNOWLEDGEMENTS

The data for the main study was supported by grants from the EXTRA funds from the Norwegian Foundation for Health and Rehabilitation through the Norwegian Low Back Pain Association, Norwegian Chiropractic Associations Research Fund, and Wilhelmsens Research Fund.

REFERENCES

1. Grotle M, Brox JI, Vøllestad NK. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. Spine 2005; 30: 130–140.
2. Kopec JA. Measuring functional outcomes in persons with back pain. A review of back-specific questionnaires. Spine 2000; 25: 3110–3114.
3. Bombardier C. Outcome Assessments in the evaluation of treatment of spinal disorders. Summary and general recommendations. Spine 2000; 25: 3100–3103.
4. Roland M, Morris R. 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 1983; 8: 141–144.
5. Stratford PW, Binkley JM. Measurement properties of the RM-18. A modified version of the Roland-Morris Disability Scale. Spine 1997; 22: 2416–2421.
6. Williams RM, Myers AM. Support for a shortened Roland-Morris Disability Questionnaire for patients with acute low back pain. Physio Can 2001; 53: 60–66.
7. Stroud MW, McKnight PE, Jensen MP. Assessment of self-reported physical activity in patients with chronic pain: development of an abbreviated Roland-Morris disability scale. J Pain 2004; 5: 257–263.
8. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995; 20: 1899–1909.
9. Atlas SJ, Deyo RA, van den AM, Singer DE, Keller RB, Patrick DL. The Maine-Seattle back questionnaire: a 12-item disability questionnaire for evaluating patients with lumbar sciatica or stenosis: results of a derivation and validation cohort analysis. Spine 2003; 28: 1869–1876.
10. Cook KF, Choi SW, Crane PK, Deyo RA, Johnson KL, Amtmann D. Letting the CAT out of the bag: comparing computer adaptive tests and an 11-item short form of the Roland-Morris Disability Questionnaire. Spine 2008; 33: 1378–1383.
11. Roland M, Fairbank JC. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine 2000; 25: 3115–3124.
12. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002; 82: 8–24.
13. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004; 7 Suppl 1: S22–S26.
14. Garratt AM. Rasch analysis of the Roland disability questionnaire. Spine 2003; 28: 79–84.
15. Kucukdeveci AA, Tennant A, Elhan AH, Niyazoglu H. Validation of the Turkish version of the Roland-Morris Disability Questionnaire for use in low back pain. Spine 2001; 26: 2738–2743.
16. Davidson M. Rasch analysis of 24-, 18- and 11-item versions of the Roland-Morris Disability Questionnaire. Qual Life Res 2009; 18: 473–481.
17. Bergner M, Bobbitt RA, Carter WB, Gibson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981; 19: 787–805.
18. Grotle M, Brox JI, Vollestad NK. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. J Rehab Med 2003; 35: 241–247.
19. Grotle M, Brox JI, Vollestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine 20041; 29: E492–E501.
20. Wilkens P, Scheel IB, Grundnes O, Hellum C, Storheim K. Effect of glucosamine on pain-related disability in patients with chronic low back pain and degenerative lumbar osteoarthritis: a randomized controlled trial. JAMA 2010; 304: 45–52.
21. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis rheum 2007; 8: I358–1362.
22. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007: Pt 1: 1–18.
23. International Classification of Functioning, Disability and Health.: WHO (2001).
24. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronsen NK, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project. J Clin Epidemiol 1998; 51 (1203): 1214.
25. Wood-Dauphinee SL. Assessment of back-related quality of life: the continuing challenge. Spine 2001; 26: 857–861.
26. Hush JM, Refshauge KM, Sullivan G, De SL, McAuley JH. Do numerical rating scales and the Roland-Morris Disability Questionnaire capture changes that are meaningful to patients with persistent back pain? Clin Rehabil 2010; 24: 648–657.
27. Garratt AM, Klaber MJ, Farrin AJ. Responsiveness of generic and specific measures of health outcome in low back pain. Spine 2001; 26: 71–77.
28. Macedo LG, Maher CG, Latimer J, Hancock MJ, Machado LA, McAuley JH. Responsiveness of the 24-, 18- and 11-item versions of the Roland Morris Disability Questionnaire. Eur Spine J 2011; 20: 458–463.