Eur Spine J (2011) 20:458–463
DOI 10.1007/s00586-010-1608-2

ORIGINAL ARTICLE

Responsiveness of the 24-, 18- and 11-item versions of the Roland Morris Disability Questionnaire

Luciana Gazzi Macedo · Chris G. Maher ·
Jane Latimer · Mark J. Hancock ·
Luciana A. C. Machado · James H. McAuley

Received: 4 January 2010 / Revised: 17 September 2010 / Accepted: 17 October 2010 / Published online: 31 October 2010

Abstract Several versions of the 24-item Roland Morris Disability Questionnaire (RMDQ) have been proposed; however, their responsiveness has not been extensively explored. The objective of this study was to compare the responsiveness of four versions of the RMDQ. Perceived disability was measured using the 24-item, two 18-item and an 11-item RMDQ on 1,069 low back pain patients from six randomised controlled trials. Responsiveness was calculated using effect size, Guyatt's responsiveness index (GRI) and receiver operating characteristics (ROC) curves. Effect size analyses showed that both 18-item versions of the RMDQ were superior to the 24- and 11-item versions of the RMDQ. GRI showed that the 24- and 18-item versions of the RMDQ were similar but more responsive than the 11-item. ROC curves revealed that the 11-item was less responsive than the other three versions, which had similar responsiveness. The results of this study demonstrate that the 24-item and both 18-item versions of the RMDQ have similar responsiveness with all having superior responsiveness to the 11-item.

Keywords Low back pain · Questionnaires · ROC curve · Validity

L. G. Macedo (✉) · C. G. Maher · J. Latimer
The George Institute for Global Health,
The University of Sydney, PO Box M201,
Missenden Rd, Sydney, NSW 2050, Australia
e-mail: lucianagazzi@hotmail.com

M. J. Hancock
The Faculty of Health Sciences, The University of Sydney,
PO Box 170, Lidcombe, NSW 1825, Australia

L. A. C. Machado
Escola de Educação Física, Fisioterapia e Terapia Ocupacional,
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

J. H. McAuley
Neuroscience Research Australia, The University of New South
Wales, PO Box 1165, Randwick, Sydney, NSW 2031, Australia

Introduction

Questionnaires are routinely used to measure disability associated with low back pain (LBP). In clinical practice and research, the Roland Morris Disability Questionnaire (RMDQ) is one of the most frequently used LBP-specific disability measures [1, 2]. This questionnaire, developed by two health practitioners with experience in treating LBP patients, uses 24 items selected from the 136-item Sickness Impact Profile with addition of the phrase "because of my back problem" to the selected items [1].

For questionnaires to provide useful measures of clinical outcome they need to demonstrate adequate clinimetric properties. One of the most important clinimetric properties of a questionnaire is responsiveness, which is defined as the ability of an outcome measure to detect true change over time [2]. There are two main approaches to quantifying responsiveness: anchor-based (external responsiveness) and distribution-based methods (internal responsiveness) [3]. The anchor-based approach relies upon use of a "gold standard" of true change such as a single item global perceived effect (GPE) scale. The distribution-based approach quantifies responsiveness as a ratio of group mean change to variability of the measure. Jordan et al. [4] suggest that to fully understand the responsiveness of a questionnaire, a combination of the two methods is necessary.

Several versions of the original 24-item RMDQ have been proposed with the aim of improving the questionnaire's clinimetric properties [5]. Two different 18-item questionnaires proposed by Stratford and Binkley [6] and by Williams and Myers [7] were developed using classic test theory (frequency of endorsement and correlation between the items) and an 11-item questionnaire [5] was developed using item response theory (mathematical function to describe the relationship between examinee performance on a test and the unobservable latent trait underlying test performance). Studies have provided estimates for the reliability of the RMDQ questionnaires; reliability for the 24-item ranges from 0.42 to 0.91 [8–10], the 18-item Stratford ranges from (ICC) 0.68 to 0.75 [9, 11] and the reliability of the 11-item, given using correlation coefficient, is 0.89. No reliability data were found for the Williams and Myers 18-item scale. There is some evidence to show that the 18-item versions have similar responsiveness to the original 24-item RMDQ [9, 10] but at present there is no evidence for the 11-item version. Additionally, the available studies have some limitations in generalising to primary care as both used small sample sizes and one [9] studied patients in secondary care. Lauridsen [3] showed that the responsiveness between patients in the primary and secondary sectors of care are significantly different.

Therefore, the objective of this study was to compare the responsiveness of four versions of the RMDQ for patients with non-specific LBP presenting for treatment in primary care. We conducted these analyses on a large dataset ($n = 1{,}069$) formed by pooling data from six trials completed by our group.

Methods

Study population

One thousand and sixty-nine people with non-specific LBP who had participated in six different primary care randomised controlled trials [12–17] and had full RMDQ data for baseline and after treatment scores were included in this study. The participants had been recruited via referral from general practitioners, rehabilitation clinics of public hospitals and advertisement. The common inclusion criteria were lower back pain with or without leg pain and age between 18 and 80 years. The common exclusion criteria included: spinal surgery, specific pathology (e.g. nerve root compression, fracture, and malignancy), contraindication to exercise and insufficient English language ability to complete questionnaires.

Outcome measures

Disability was assessed using the original 24-item RMDQ [1] at baseline and at completion of treatment in all six studies (1-week follow-up for the acute studies, 6 weeks for the subacute study and 8–12 weeks for the chronic studies). We used the responses to the 24-item RMDQ to calculate the two 18-item versions [6, 7] and the 11-item [5] version of the RMDQ. An 11 point GPE scale was also administered at the end of treatment in all included studies. This scale ranged from −5 (vastly worse) to 5 (completely recovered/much better), with 0 being unchanged. Although the randomised controlled trials included long-term follow-up, we restricted the analyses to the follow-up closest to the end of treatment since this is when the largest effect is expected [18].

Statistical analysis

Data analyses were performed using SPSS 17.0 and the Hong Kong ROC program [19]. Descriptive statistics were calculated for the participant characteristics (e.g. age, sex and duration of LBP) and for the RMDQ baseline and follow-up scores.

Internal responsiveness was assessed using effect sizes (ES). ES were defined as the mean difference divided by the standard deviation of the baseline score [20] and were calculated for each version of the RMDQ questionnaire. We calculated 84% confidence intervals because non-overlapping 84% confidence intervals are equivalent to a *Z* test of mean at the 0.05 level [21]. Internal responsiveness was also assessed using Guyatt's responsiveness index (GRI). This index takes into consideration true changes that are assessed by external criteria [2]. It is calculated by dividing the mean change of patients who have improved by the standard deviation of change of patients reporting no improvement. A cut-off of three units on the GPE was used to identify patients that improved and did not improve.

To assess external responsiveness, we correlated the RMDQ change scores with the GPE scores and used Cohen's test for paired correlation to compare the correlation coefficients [22]. The second approach to external responsiveness used Receiver Operating Characteristics (ROC) curve analyses to compare the ability of the various versions of the RMDQ to discriminate those people who had and had not improved. For the ROC curve analyses patients were classified as having improved if they scored greater than or equal to three on the GPE. A sensitivity analysis was also performed using a more strict (greater than or equal to 4) and less strict criterion for improvement (greater than or equal to 2). Separate ROC curves were calculated for each version of the RMDQ and the DeLong et al. [23] statistic was used to determine if AUC values were statistically significantly different.

We have used 84% confidence intervals for GRI and ES because no inferential statistics were available to compare these measures between scales. However, since inferential tests were available to compare correlation coefficients and ROC curves, we chose to maintain the commonly used 95% confidence intervals.

Results

All patients included in this study had non-specific LBP: 376 had acute LBP (<6 weeks duration) [13, 17], 232 had subacute LBP (>6 and <12 weeks) [12] and 461 had chronic LBP (>12 weeks) [14–16]. Demographic characteristics for participants in each trial are presented on Table 1. Mean and standard deviation, mean difference (baseline minus post-intervention scores) and ES (mean difference divided by baseline standard deviation) for each version of the RMDQ are presented on Table 2.

The results for internal responsiveness using ES and 84% confidence intervals showed that the 18-item versions were superior to both the 24- and 11-item versions of the RMDQ. The 24- and 11-item versions had similar responsiveness. The results of the GRI showed that the responsiveness of the 24- and 18-item questionnaires was similar but they were all significantly more responsive than the 11-item questionnaire (Table 1).

Correlation between the GPE and the four versions of the RMDQ was similar in magnitude and not statistically significantly different ($P > 0.05$ for all comparisons) (Table 3). The results for external responsiveness evaluated by ROC curve analyses are shown in Fig. 1 and Table 3. Delong's test of paired ROC curves revealed that the 11-item RMDQ was less responsive than the other three versions ($P = 0.002$ between 24- and 11-item versions,

Table 1 Demographic characteristics of patients included in each randomized controlled trial

	Acute		Subacute	Chronic			Overall
	Hancock et al. [13]	Machado et al. [17]	Pengel et al. [12]	Ferreira et al. [15]	Costa et al. [16]	Macedo et al. [14]	
Number of subjects	237	139	232	220	152	89	1,069
Age (years) Mean (SD)	40.7 (15.7)	46.6 (14.7)	49.9 (15.8)	53.7 (14.9)	53.7 (12.8)	50.2 (14.5)	50.5 (15.6)
Gender (%)	44 F, 56 M	50 F, 50 M	48 F, 52 M	69 F, 39 M	60 F, 40 M	54 F, 46 M	55 F, 45 M
Pain duration Mean (SD)	Pain less than 6 weeks	2.6 (1.1) months	6–8 weeks (47%) 9–11 weeks (38%) 12 weeks (15%)	103.1 (120.7) months	83 (98) months	102.2 (119.6) months	N/A
Timing of follow-ups (weeks)	1	1	6	8	8	8–12	1–12
Baseline VAS Mean (SD)	6.5 (1.7)	6.7 (1.9)	5.4 (1.2)	6.3 (2.0)	6.7 (2.0)	6.5 (1.9)	6.4 (1.8)
Baseline RMDQ 24 items	13.1 (5.4)	13.7 (5.3)	8.3 (4.8)	13.2 (5.4)	13.2 (5.0)	12.1 (5.0)	12.1 (5.6)
Baseline RMDQ 18 items (Stratford)	11.0 (4.4)	11.6 (4.2)	6.9 (4.1)	11.1 (4.5)	10.8 (4.0)	10.1 (4.3)	10.2 (4.6)
Baseline RMDQ 18 items (Williams)	11.1 (4.5)	11.6 (4.4)	6.9 (4.2)	11.2 (4.6)	10.9 (4.1)	10.2 (4.4)	10.1 (4.6)
Baseline RMDQ 11 items	7.1 (3.0)	7.6 (2.9)	4.3 (3.0)	7.3 (3.0)	7.0 (2.8)	7.6 (2.9)	6.6 (3.2)

Table 2 Mean, standard deviation, mean difference and effect sizes for the RMDQ scales

	Mean (SD)		Mean difference (SD)	ES (84% CI)	GRI (84% CI)
	Baseline	Post-treatment			
24-item version	12.1 (5.6)	8.4 (5.9)	3.7 (4.7)	0.67 (0.63–0.71)	1.55 (1.48–1.62)
18-item version[a]	10.2 (4.7)	6.6 (5.2)	3.6 (4.7)	0.75 (0.71–0.79)	1.49 (1.42–1.57)
18-item version[b]	10.1 (4.6)	6.5 (5.1)	3.6 (4.6)	0.78 (0.73–0.82)	1.52 (1.45–1.59)
11-item version	6.6 (3.2)	4.5 (3.5)	2.1 (3.1)	0.65 (0.61–0.69)	1.30 (1.23–1.38)

ES were calculated using mean differences divided by the standard deviation of the baseline scores

SD standard deviation, *ES* effect sizes, *GRI* Guyatt responsiveness index

[a] 18-item RMDQ version proposed by Williams and Myers [6]

[b] 18-item RMDQ version proposed by Stratford and Binkley [7]

Table 3 Pearson correlation coefficients between the RMDQ questionnaires and the GPE and AUC for the RMDQ scales using the cut offs of 3, 2 and 4 on the GPE

	Correlation with GPE scale	AUC (cut off 3 or greater on GPE)	AUC (cut off 2 of greater on GPE)	AUC (cut off 4 or greater on GPE)
24-item version	0.49 (0.45–0.54)	0.78 (0.76–0.81)	0.79 (0.76–0.82)	0.74 (0.71–0.77)
18-item version[a]	0.49 (0.44–0.53)	0.78 (0.75–0.81)	0.79 (0.76–0.82)	0.74 (0.71–0.77)
18-item version[b]	0.49 (0.44–0.53)	0.78 (0.75–0.81)	0.79 (0.76–0.82)	0.74 (0.71–0.77)
11-item version	0.44 (0.39–0.49)	0.75 (0.72–0.78)	0.76 (0.73–0.79)	0.73 (0.70–0.77)

Numbers in brackets are 95% confidence intervals

[a] 18-item RMDQ version proposed by Williams and Myers [6]

[b] 18-item RMDQ version proposed by Stratford and Binkley [7]

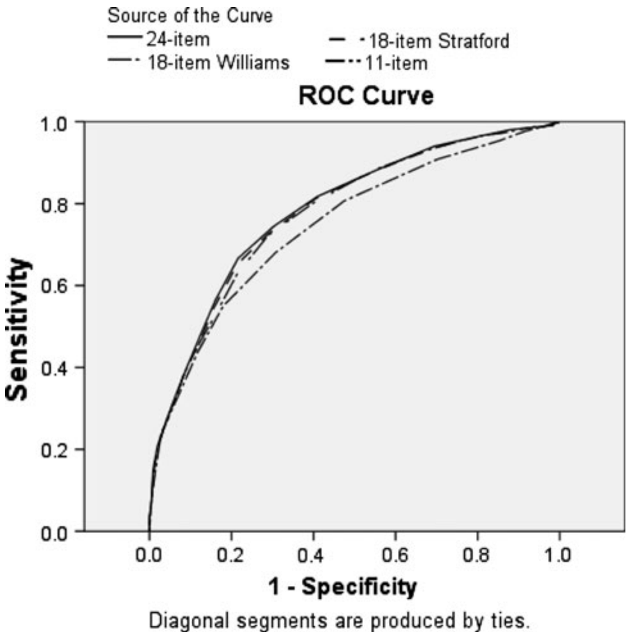

Fig. 1 ROC curves for the four versions of the RMDQ using three as the cut off for the GPE to identify who improved and who did not improve

$P = 0.002$ between 18-item Stratford and 11-item version and $P = 0.0008$ between 18-item Williams and 11-item version), which had similar responsiveness ($P > 0.05$ for all comparisons between 24- and both 18-item versions). The sensitivity analyses revealed a similar pattern of results for the ROC curve analyses.

Discussion

The results of this study demonstrate that the 24- and both 18-item versions of the RMDQ have similar responsiveness with all three versions superior to the 11-item version. The poor performance of the 11-item RMDQ is unequivocal with all internal and external responsiveness analyses demonstrating that the 11-item RMDQ had significantly lower responsiveness than the other versions of the RMDQ. While some analyses favoured the 18-item versions of the RMDQ over the 24-item RMDQ, the difference in responsiveness was small and not evident across all analyses.

It is interesting to consider why the additional six items in the 24-item RMDQ do not translate into greater responsiveness. The mean difference over time was very similar between these questionnaires (Table 1) despite the additional six items of the 24-item questionnaire. This result suggests that there may be items on the 24-item questionnaire that are insensitive to change (or time) and so do not add to the responsiveness of the questionnaire. Davidson [24] performed a Rasch analysis on the 24-item questionnaire and showed that four of the items from the 24-item questionnaire that are not present in the 18-item versions were the least difficult items (my appetite is not very good because of my back, because of my back I get dressed with help from someone else, I sit down for most of the day because of my back and I stay in bed most of the time because of my back) and one item from the 24-item questionnaire that is not on the 18-item versions was the second most difficult item to respond (I change positions frequently to try and get my back comfortable). Although items with different levels of difficulty need to be present on a questionnaire to guarantee an adequate level of discrimination, we believe that in this questionnaire these items may not be adding to the responsiveness of the 24-item questionnaire and may be irrelevant.

The only study to our knowledge that used a statistical procedure to compare ES between RMDQ questionnaires is the one by Pengel et al. [20]. Their results showed that internal responsiveness was similar between the 24- and both 18-item versions, a result which concurs with ours. The results of Ostelo et al. [9] comparing external responsiveness of the 24- and the 18-item versions proposed by Stratford and Binkley [6] showed higher AUC for both of the questionnaires (0.95 and 0.95) when compared to AUC found in this study (0.78 and 0.78). A potential explanation for this difference is that the external criterion used by Ostelo et al. [9] had only 3 points whereas the GPE scale used in our study had 11 points.

The important strengths of our study are the large sample size with a wide range of patients with acute, subacute and chronic LBP and the fact that patients were undergoing different treatment regimes. Another important feature of this study is the use of preferred statistical methods with the use of inferential statistics to investigate if any observed differences in responsiveness indices were greater than those expected by chance. The limitations of this study are the inclusion of a population from only one country and the fact that results for the 18- and 11-item questionnaires were drawn from the 24-item questionnaire. We do not know how much the presentation of questionnaires with a smaller number of questions could influence the patient's responses.

Most studies published to date comparing different versions of the questionnaires [9, 20, 24] concluded that there were no differences, or only slight differences, in responsiveness between the 18- and 24-item versions of the RMDQ and therefore, the use of the 24-item was preferred in order to guarantee homogeneity between trials and allow for further comparisons. While we agree that the responsiveness of the 18- and 24-item versions are similar, we can imagine scenarios where the 18-item version would be preferred. An additional six items is probably of no consequence when looking at a pen and paper administration of this questionnaire but the six extra items may be of concern when administering the questionnaire over the phone and/or when the RMDQ is part of a larger battery of questionnaires. Also in clinical situations where language barriers exist and interpreters may be needed, the length of a questionnaire may be an important consideration.

Acknowledgments Luciana Macedo holds a PhD scholarship jointly funded by the University of Sydney and the Australian Government. Chris Maher's research fellowship is funded by Australia's National Health and Medical Research Council. Jane Latimer's research fellowship is funded by the Australian Research Council. Luciana Machado's postdoctoral fellowship is funded by Fundação de Amparo à Pesquisa do Estado de Minas Gerais, FAPEMIG, Brazil.

Conflict of interest None.

References

1. Roland M, Morris R, 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
2. Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 40:171–178
3. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsoon N (2006) Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord 7:1–16
4. Jordan K, Dunn KM, Lewis M, Croft P (2006) A minimal clinically important difference was derived for the Roland–Morris disability questionnaire for low back pain. J Clin Epidemiol 59:45–52
5. Stroud MW, McKnight PE, Jensen MP (2004) Assessment of self-reported physical activity in patients wiht chronic pain: development of an abbreviated Roland–Morris disability scale. J Pain 5:257–263
6. Stratford PW, Binkley JM (1997) Measurement properties of the RM-18. A modified version of the Roland–Morris disability scale. Spine 22:2416–2421
7. Williams RM, Myers AM (2001) Support for a shortened Roland–Morris Disability Questionnaire for patietns with acute low back pain. Physiother Can 53:60–66
8. Brouwer S, Kuijer W, Dijkstra P, Goeken L, Groothoff J, Geertzen J (2004) Reliability and stability of the Roland Morris disability questionnaire: intra class correlation and limits of agreement. Disabil Rehabil 26:162–165
9. Ostelo RWJG, de Vet HCW, Knol DL, van den Brandt PA (2004) 24-item Roland Morris disability questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. J Clin Epidemiol 57:268–276
10. Riddle DL, Stratford PW (2002) Roland-Morris scale reliability. Phys Ther 82:512–515 (author reply 515–517)
11. Chansirinukor W, Maher CG, Latimer J, Hush J (2005) Comparison of the functional rating index and the 18-item Roland–Morris disability questionnaire: responsiveness and reliability. Spine 30:141–145
12. Pengel LHM, Refshauge KM, Maher CG, Nicholas MK, Herbert RD, McNair P (2007) Physiotherapist-directed exercise, advice, or both for subacute low back pain: a randomized trial. Ann Intern Med 146:787–796
13. Hancock MJ, Maher CG, Latimer J, McLachlan AJ, Cooper CW, Day RO, Spindler MF, McAuley JH (2007) Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: a randomised controlled trial. Lancet 370:1638–1643
14. Macedo LG, Latimer J, Maher CG, Hodges PW, Nicholas M, Tonkin L, McAuley JH, Stafford R (2008) Motor control or graded activity exercises for chronic low back pain? A randomised controlled trial. BMC Musculoskelet Disord 9 [Epub ahead of print]
15. Ferreira ML, Ferreira PH, Latimer J, Herbert RD, Hodges PW, Jennings MD, Maher CG, Refshauge KM, Ferreira ML, Ferreira PH, Latimer J, Herbert RD, Hodges PW, Jennings MD, Maher CG, Refshauge KM (2007) Comparison of general exercise, motor control exercise and spinal manipulative therapy for chronic low back pain: a randomized trial. Pain 131:31–37
16. Costa LOP, Maher CG, Latimer J, Hodges PW, Herbert RD, Refshauge KM, McAuley JH, Jennings MD (2009) Motor control exercises for chronic low back pain: a randomized placebo-controlled trial. Phys Ther 89:1275–1286
17. Machado LAC, Maher CG, Herbert RD, Clare H, McAuley JH (2010) The effectiveness of the McKenzie method in addition to first-line care for acute low back pain: a randomized controlled trial. BMC Med 8:10
18. Machado LAC, Kamper SJ, Herbert RD, Maher CG, Mcauley JH (2009) Analgesic effects of treatments for non-specific low back pain: a meta-analysis of placebo-controlled randomized trials. Rheumatology 48:520–527
19. Cheng A. The HONG KONG ROC program: Chinese University of Hong Kong, Department of Obstetrics and Gynecology. http://department.obg.cuhk.edu.hk/researchsupport/statmenu.asp

20. Pengel LHM, Refshauge KM, Maher CG (2004) Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. Spine 29:879–883
21. Tryon W (2001) Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. Psychol Methods 6:371–386
22. Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences. Erlbaum Associates, Hillsdale
23. DeLong E, DeLong D, Clarke-Pearson D (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
24. Davidson M (2009) Rasch analysis of 24- 18- and 11-item versions of the Roland–Morris disability questionnaire. Qual Life Res 18:473–481