RESPONSIVENESS OF THE ROLAND MORRIS
DISABILITY QUESTIONNAIRE:

CONSEQUENCES OF USING DIFFERENT EXTERNAL CRITERIA

CHAPTER 3

Wietske Kuijer
Sandra Brouwer
Pieter U. Dijkstra
Wim Jorritsma
Johan W. Groothoff
Jan H.B. Geertzen

Submitted
ABSTRACT

Objective: To determine consequences of using different external criteria on responsiveness of the RMDQ in patients with Chronic Low Back Pain.

Design: Questionnaire measures before and after rehabilitation treatment, to determine responsiveness.


Main measures: The Dutch language version of the Roland Morris Disability Questionnaire (RMDQ-Dv) was used to assess self-reported functional status. The used external criteria were: 1. Global perceived effect of change in complaints. 2. Global perceived effect of change in ability to take care of oneself. 3. Change in rating of pain intensity. 4. Smallest Real Difference (SRD). Standardised response means (SRMs), pooled effect sizes and Receiver Operating Curves were calculated to determine responsiveness and to enable comparison of effect sizes with the thresholds of Cohen.

Results: SRMs ranged from 1.33 to 3.45, pooled effect sizes ranged from 1.50 to 2.81, and areas under curves ranged from 0.76 to 1.00, dependent on the used external criterion.

Conclusion: Considerable differences were found in responsiveness, when using different external criteria. However, all pooled effect sizes were well above 0.80, and all other statistics were high, indicating good responsiveness of the RMDQ-Dv. Therefore, the RMDQ-Dv is able to distinguish between improved and non-improved patients independent of the used external criterion.
INTRODUCTION

The Roland Morris Disability Questionnaire (RMDQ) is often used as an evaluative outcome measure in patients with chronic low back pain (CLBP),\textsuperscript{1-3} to assess change in self-reported functional status after treatment. Evaluative outcome measures should be reliable and responsive to be able to assess change.\textsuperscript{4-6} Both the English and Dutch language versions of the RMDQ show good reliability when using a time interval < three weeks (Pearson’s \( r = 0.83 \) and Intra Class Correlations range from 0.79 to 0.91).\textsuperscript{7-12} Despite the good reliability, smallest real differences of 5.4 and 5.9 show that substantial variation must be taken into account when the RMDQ is used in a clinical setting.\textsuperscript{7,13} The responsiveness of the English and Dutch language versions of the RMDQ has also been investigated in several studies. However, different outcomes were found for both versions. Responsiveness statistics, such as areas under curves, ranged from 0.68 to 0.84 for the English version,\textsuperscript{14-17} and from 0.68 to 0.93 for the Dutch version.\textsuperscript{1,13,18} Effect sizes ranged from 0.50 to 1.60 for the English version\textsuperscript{14,19-21} and from 0.58 to 2.02 for the Dutch version.\textsuperscript{18,22}

No clarity exists in literature about responsiveness.\textsuperscript{23,24} Different external criteria are used to determine whether a patient has achieved a clinically important change in functional status, due to the absence of a gold standard to measure clinically important change. Furthermore, the terminology and calculation of responsiveness statistics as effect sizes and standardised response means may vary considerably,\textsuperscript{12} depending on the type of change that is intended to be measured,\textsuperscript{10,24} change in general, clinically important change, or the ability to detect changes in the construct being measured.\textsuperscript{25} Previous studies evaluating the responsiveness of the RMDQ, used different external criteria, such as a global rating scale of magnitude in or importance of change (15-points),\textsuperscript{17,21} a global rating scale of change in low back pain (7-points),\textsuperscript{1,13} and a global rating scale of change in complaints (5-points).\textsuperscript{18} Additionally, the terminology and calculation of responsiveness statistics varied in previous studies. Therefore, the previously found responsiveness statistics for the RMDQ can not be interpreted similarly. Additionally, whether the previously found effect sizes are trivial (\( \geq 0.20 \)), small (\( \geq 0.20, < 0.50 \)), moderate (\( \geq 0.50, < 0.80 \)) or large (\( \geq 0.80 \)), is not clear, because not all studies used the pooled effect size, as is needed for a direct comparison with these thresholds of Cohen.\textsuperscript{8}

Because the lack of clarity about the definition of responsiveness, different external criteria\textsuperscript{1,8,21,22,25} and unambiguous responsiveness statistics should be used in the same study to determine responsiveness of an evaluative outcome instrument. The aim of this study is to determine consequences of using different external criteria on the responsiveness of the RMDQ-Dv, in patients with CLBP.
Chapter 3

METHODS

Subjects
An existing clinical database was used. Data were gathered before and after rehabilitation treatments of patients with non-specific CLBP in rehabilitation centre Beatrixoord in Haren, the Netherlands. In total, 83 patients (44 male and 39 female), with a mean age of 38.5 years (SD 9.7), participated, and filled out the RMDQ-Dv before and after treatment. The mean RMDQ score before treatment was 10.9 (SD 4.7). The mean duration of treatment was 28 weeks (SD 14.5), including one to two treatment sessions per week.

Outcome measure
The Dutch language version of the Roland Morris Disability Questionnaire (RMDQ-Dv) was used to assess self-reported functional status before and after treatment. The RMDQ is derived from the Sickness Impact Profile, a general health questionnaire, and assesses, dichotomously, perceived limitations due to low back pain in 24 activities of daily living. The time frame used in this study was ‘the past few days’. The sum score is calculated by summing the ‘yes’ answers. The scale ranges from zero (no disability) to 24 (severe disability).

External criteria
Different external criteria were used to analyse the responsiveness of the RMDQ-Dv. First, the global perceived effect of change in complaints due to CLBP (7-points scale, ranging from 'completely recovered' to 'worse than ever'). Patients were improved if they scored ‘completely recovered’ or ‘much recovered’. Second, the global perceived effect of change in ability to take care of oneself (4-points scale, ranging from 'much improved' to 'not improved'). Patients were improved if they scored ‘much improved’. Third, change in rating of pain intensity. Three 10-point pain intensity scales were used before and after treatment, representing ‘pain when as worst’, ‘pain when as least’ and ‘pain right now’. All scales were used as a separate external criterion. Additionally, the mean of the three scales represented the fourth pain intensity criterion. According to patients' ratings of pain intensity, patients were improved if the change score was 2 units or more on the pain scales. Finally, smallest real difference (SRD) of the RMDQ-Dv was used as external criterion. Patients were improved if their change score exceeded the SRD of the RMDQ-Dv; so if they changed 6 points or more.

Responsiveness statistics
Standardised response means (SRM) were calculated as the ratio of the mean difference of the improved group and the standard deviation of this mean difference. The higher the SRM, the better the responsiveness. Pooled effect sizes were calculated as the ratio of mean difference of the improved group and the
Responsiveness of the RMDQ

The pooled standard deviation of the improved group \( \text{SD}_{\text{pooled improved}} \), in which \( \text{SD}_{\text{pooled improved}} = \Delta \left( \text{SD}_{\text{before treatment}}^2 + \text{SD}_{\text{after treatment}}^2 \right)/2 \). Effect sizes are large when exceeding 0.80.\(^8\) Receiver operating curves (ROC) were calculated. The test whose ROC encloses (below and to the right) the largest area (area under curves (AUC) approximates to 1.00) is the most accurate one.\(^28\) AUC could range from 0.50 (no accuracy in discriminating improved from non-improved patients) to 1.00 (perfect accuracy in discriminating improved from non-improved patients). The four different external criteria were used to determine the improved group, as is necessary in the calculations of these statistics.

Data analyses

An RMDQ-score defined as 'no limitations' can vary from 0 to 5 points, based upon the established SRD of the RMDQ-Dv. Therefore, it was decided to perform all analyses not only in the total improved group, but also in an improved group of patients with an initial score \( \geq 6 \). Only these patients are 'certain' of having a limitation in self-reported functional status due to CLBP (initial RMDQ-score > SRD), and can show improvement according to the SRD of the RMDQ.

To calculate the association between the different external criteria, Spearman rank correlation coefficients were calculated. To calculate the association between the different external criteria and RMDQ change scores, Pearson correlation coefficients were calculated. Missing data were excluded pair-wise from analyses. The SRD criterion is not compared with the other external criteria and with the RMDQ-change score, because this criterion is only a cut-off score. No correlation coefficients were calculated between mean pain intensity score and other pain scores, because the mean score is computed from the other pain scores. Therefore, these scores are not independent of each other. Means and standard deviations of the RMDQ before and after treatment, and mean difference and the standard deviation of the difference were calculated for the different groups of patients classified as improved by the different external criteria. Additionally, SRMs, pooled effect sizes, and ROCs were calculated. Finally, the number of improved patients classified by combinations of the different external criteria was calculated.
RESULTS

All results are shown for both improved patients and improved patients with an initial RMDQ score $\geq 6$, classified by the different external criteria. All improved patients classified by the SRD criterion have an initial RMDQ score $\geq 6$. Therefore, only one group of patients is shown according to this criterion.

Spearman rank correlation coefficients between the different external criteria and Pearson correlation coefficients between the different external criteria and the RMDQ change score range from 0.27 to 0.85 (table 1). Valid percentages of the improved patients are presented in table 2, as well as means, standard deviations of initial, post and RMDQ change scores, effect sizes (SRMs and pooled effect sizes), and AUC with confidence intervals. SRMs range from 1.33 to 3.45. Pooled effect sizes range from 1.50 to 2.81. AUC range from 0.76 to 1.00. In total, 16 patients are classified as improved if all external criteria are applied. Applying a single external criterion, 30 to 39 patients are classified as improved, dependent of the used external criterion.

Table 1. Relationships between change in self-reported functional status and different external criteria*  
<table>
<thead>
<tr>
<th></th>
<th>$\Delta$ RMDQ$^1$</th>
<th>Complaints</th>
<th>Taking care of oneself$^2$</th>
<th>$\Delta$ Pain least$^3$</th>
<th>$\Delta$ Pain worst$^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complaints$^2$</td>
<td>-0.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taking care of oneself$^2$</td>
<td>-0.51</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ Pain mean$^4$</td>
<td>0.85</td>
<td>-0.56</td>
<td>-0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ Pain least$^5$</td>
<td>0.66</td>
<td>-0.41</td>
<td>-0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ Pain worst$^6$</td>
<td>0.80</td>
<td>-0.50</td>
<td>-0.38</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>$\Delta$ Pain now$^7$</td>
<td>0.80</td>
<td>-0.60</td>
<td>-0.47</td>
<td>0.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>

*All coefficients were significant at $p \geq 0.001$

$^1$change in self-reported functional status, measured by the RMDQ-Dv

$^2$global rating of change in complaints

$^3$global rating of change in taking care of oneself

$^4$mean pain intensity difference

$^5$pain intensity difference ‘pain when as least’

$^6$pain intensity difference ‘pain when as worst’

$^7$pain intensity difference ‘pain right now’
DISCUSSION

The choice of the external criterion influences the size of the responsiveness statistic. In our study, comparing application of external criteria in the same study population, considerable differences in effect sizes were found. The differences in effect sizes amounted to 2.12 points. Additionally, the differences found in AUC amounted to 0.24. Refrained from these large differences, comparison of pooled effect sizes with the thresholds of Cohen showed that all effect sizes were ranging well above 0.80. Furthermore, all AUC were above 0.75. These results indicate good responsiveness of the RMDQ-Dv, independent of the used external criterion.

In order to compare our results with previous studies, we calculated a SRM. The effect sizes found in our study ranged from 1.33 to 3.45. This means that a large amount of variability in effect sizes can be expected, dependent of the use of the chosen external criterion. Previous effect sizes calculated similarly are found of respectively 0.58, 1.01, and 2.02. The effect size of 0.58 found by Gommans et al can be explained by a low change score of 2.42 found in that study. This change score did not exceed the SRD of 6 points. Based upon this, on group level, a real change has not occurred according to the RMDQ-Dv. Pooled effect sizes were calculated to enable an adequate comparison of effect sizes with the thresholds of Cohen. All pooled effect sizes were well above 0.80, indicating good responsiveness, independent of the used external criterion.

In the present study, using SRD as an external criterion yields for both the effect sizes and the size of the AUC the highest statistic. It was expected that using this statistic as a criterion for change, almost all patients would be classified as improved correctly, because SRD as a cut off score for change is not a real external criterion, but it is based on the measurement properties of the instrument itself. The specificity to change is with the SRD as the cut-off for change per definition, for a normal distribution, equal to 95%. In our study however, the specificity to change was 100%, due to a skewed distribution of RMDQ change scores. This means that the estimation, if patients underwent a clinically important change, can be made with 100% accuracy in this group of patients. The next highest responsiveness statistic is found when using mean pain intensity difference as an external criterion, followed by the pain intensity scales ‘pain right now’, ‘pain when as least’ and ‘pain when as worst’. We chose to use all four pain intensity scales as separate external criteria, because we wanted to compare different external criteria in this responsiveness study. The mean pain intensity criterion was previously used in a study after responsiveness of the 100-mm Visual Analogue Scale, specifically aimed at patients with CLBP. It should be noted however, that the validity of the calculation of this mean pain intensity has not been investigated.
Table 2. Characteristics and responsiveness of the RMDQ-Dv, applying different external criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Global rating of change in complaints</th>
<th>Global rating of taking care of oneself</th>
<th>Δ pain intensity 'pain when as least'</th>
<th>Δ pain intensity 'pain when as worst'</th>
<th>Δ pain intensity 'pain right now'</th>
<th>Δ Mean pain intensity</th>
<th>Smallest Real Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Improved patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved (n [valid %])</td>
<td>39 [48.8]</td>
<td>37 [46.8]</td>
<td>38 [46.9]</td>
<td>30 [37.0]</td>
<td>36 [43.9]</td>
<td>31 [38.3]</td>
<td></td>
</tr>
<tr>
<td>SRM*</td>
<td>1.56</td>
<td>1.33</td>
<td>2.16</td>
<td>1.87</td>
<td>2.42</td>
<td>2.64</td>
<td></td>
</tr>
<tr>
<td>ESp**</td>
<td>1.68</td>
<td>1.50</td>
<td>2.09</td>
<td>2.07</td>
<td>2.29</td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>.82</td>
<td>.76</td>
<td>.92</td>
<td>.84</td>
<td>.94</td>
<td>.93</td>
<td></td>
</tr>
<tr>
<td>[95% c.i.][**]</td>
<td>[.73 to .91]</td>
<td>[.65 to .87]</td>
<td>[.86 to .98]</td>
<td>[.75 to .92]</td>
<td>[.89 to .98]</td>
<td>[.88 to .98]</td>
<td></td>
</tr>
</tbody>
</table>
**Improved patients with an initial score ≥6**

|               | (n [valid %]) | Initial RMDQ-score (mean [SD]) | Post RMDQ-score (mean [SD]) | Mean Δ RMDQ (mean [SD]) | SRM* | ESp** | AUC [95% c.i.]**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n [valid %])</td>
<td>33 [49.3]</td>
<td>12.8 [3.1]</td>
<td>5.1 [4.5]</td>
<td>7.8 [4.8]</td>
<td>1.62</td>
<td>1.98</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>35 [51.5]</td>
<td>13.5 [3.3]</td>
<td>4.8 [3.7]</td>
<td>8.7 [3.3]</td>
<td>2.63</td>
<td>2.45</td>
<td>.93</td>
</tr>
</tbody>
</table>

* SRM = standardised response mean = mean difference improved group / SD difference improved group
** ESp = pooled effect size = mean difference improved group / SDpooled improved group, in which SDpooled improved group = root mean square of the standard deviations before and after treatment of the improved group
*** AUC [95% c.i.] = Area Under Curve [95% confidence interval]
The high responsiveness statistic found for the RMDQ-Dv when using the pain intensity criteria means that a change in self-reported functional status is best comparable with a change in self-reported pain intensity scores, and consequently, pain intensity and functional status are closely related to each other.

The relationship between pain and self-reported functional status has been investigated in a post hoc analysis. Pearson correlation coefficients were calculated between the different initial pain intensity scores and initial RMDQ-Dv scores. Correlation coefficients ranged between 0.43 and 0.56 (all were significant at p<0.001, table 3). Additionally, Pearson's correlation coefficients were calculated between the different change pain intensity scores and RMDQ-Dv change scores. Correlation coefficients were ranging from 0.66 to 0.85 (table 1). Although initial pain scores and self-reported functional status measures are not strongly related (correlation coefficients ranged from 0.27 to 0.69) it seems that a change in self-reported pain intensity accompanies a considerable change in self-reported functional status, as Beurskens already suggested.

Table 3. Relationships between self-reported functional status and initial pain scores*

<table>
<thead>
<tr>
<th>Initial RMDQ-score</th>
<th>Initial mean pain intensity</th>
<th>Initial pain intensity 'pain when as least'</th>
<th>Initial pain intensity 'pain when as worst'</th>
<th>Initial pain intensity 'pain right now'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.56</td>
<td>0.43</td>
<td>0.44</td>
<td>0.55</td>
</tr>
</tbody>
</table>

* All correlation coefficients were significant at p < 0.001.

A discrepancy exists between clinical assessment of patients with CLBP, and scientific purposes of determining the responsiveness of the RMDQ-Dv in this group of patients. Looking at the SRD of the RMDQ-Dv, in our study, 10-11% of the patients is not 'certain' of having a limitation in self-reported functional status due to CLBP (initial RMDQ-score < SRD), and can not show any improvement according to the SRD. However, these patients were treated for their low back complaints, and they did improve according to the external criteria global rating of complaints, global rating of taking care of oneself and self-reported change in pain intensity. This discrepancy can be explained as follows: firstly, patients' limitations concern other domains than functional status in daily life, which is assessed by the RMDQ. Secondly, an improvement on one of the items of the RMDQ is that important for the patient, that he or she feels really improved, despite the absence of change on other items, the so called patient priority. For
these reasons, patients did not improve in self-reported functional status as is measured by the RMDQ, but they may have reached their personal treatment goals. Thirdly, the global perceived effect measures are only used after treatment. Because patients usually underestimate their pre-treatment state, their assessment of being improved after treatment may therefore be biased, resulting in an overestimation of improvement on the global ratings. The pain intensity criteria scales are used before and after treatment, but the scales ‘pain when as worst’ and ‘pain when as least’ also refer to a previous state. Because this discrepancy between clinical assessment and research purposes, we decided to dichotomise our presentation of results into two groups: all improved patients and the group of improved patients with an initial RMDQ score $\geq 6$. Lower responsiveness statistics are found for all improved patients compared to the improved group with an initial RMDQ score $\geq 6$, when using global rating of change in complaints, global rating of taking care of oneself, and change in pain intensity scores ‘pain when as least’, pain when as worst’ and ‘pain right now’ as external criteria. These differences in responsiveness statistics between both groups can be explained by baseline score variability. If baseline-score variability decreases, which occurs when excluding patients with an initial score $\geq 6$, the responsiveness of a measurement increases. When using the criteria of mean pain intensity or SRD, exactly the same patients were classified as ‘improved’ for all patients as for the group with an initial score $\geq 6$, thus no change in baseline score variability occurs. Additionally, the ability of the RMDQ to detect changes diminishes when small limitations exist in self-reported functional status, as assessed by the RMDQ.

It can be concluded that considerable differences were found in responsiveness statistics, when using different external criteria. However, good responsiveness is found for the RMDQ, independent of the used external criterion and therefore, the RMDQ is able to distinguish between improved and non-improved patients. This means that the RMDQ-Dv can be used to determine whether a patient has changed in self-reported functional status after treatment.
REFERENCES

Chapter 3


