Rasch analysis of the 23-item version of the Roland Morris Disability Questionnaire

Kent, Peter; Grotle, Margreth; Dunn, Kate M; Albert, Hanne B; Lauridsen, Henrik H

Published in:
Journal of Rehabilitation Medicine

DOI:
10.2340/16501977-1935

Publication date:
2015

Document version
Final published version

Citation for published version (APA):

Terms of use
This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving.
If no other license is stated, these terms apply:
• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim.
Please direct all enquiries to puresupport@bib.sdu.dk

Download date: 07. Oct. 2019
INTRODUCTION

Back pain affects many different aspects of people’s lives, and an important index of this is how pain affects people’s ability to perform functional activities (activity limitation) (1). Measuring activity limitation not only provides an index of which functional activities are affected, but also identifies barriers to return to full participation in people’s social and occupational roles. The measurement of activity limitation has therefore become increasingly central in the clinical assessment of people with back pain.

Activity limitation is usually measured by questionnaires that are self-completed by patients. The Roland Morris Disability Questionnaire (RMDQ) is the most commonly used, self-reported questionnaire for measuring this construct in people with low back pain (LBP) (2), having been recommended as part of the core set of outcomes for assessing LBP (3, 4). Using traditional research methods based on classical test theory, the RMDQ has been shown to have measurement properties comparable to the available alternative questionnaires for measuring activity limitation in LBP (5–7).

Measurement science is progressing, however, and the performance criteria expected of questionnaires are becoming more stringent. For example, there is an increasing expectation that high-quality questionnaires produce scores for which there is evidence that they can be used mathematically to calculate scores such as means (interval level scores), and that they are stable across important clinical characteristics of patients, such as age and gender.

Rasch analysis is an example of a method of instrument development that is related to item-response theory. In the current context, Rasch analysis is a tool that can determine: whether a questionnaire can produce interval level measures (although this would require a score transformation); whether it measures only a single construct (unidimensionality), such as activity limitation; whether it contains questions on activities whose difficulty levels are evenly spaced; and whether it determines if these properties are constant across age, gender and other clinical variables (referred to as person factors) (8).

If a questionnaire does not fit the Rasch model, such analysis also provides diagnostic information as to which components require further development. There are 6 English-language versions of the RMDQ, and they vary in the number of questions they contain, their content and their response options (9–14). The two versions most
commonly used in research are the original 24-item version (RMDQ-24) (11) and a 23-item version (RMDQ-23) (10); all the others are shortened versions of these. The RMDQ-23 contains 23 items measuring activity limitation associated with back and leg pain. Each item is dichotomously scored yes/no according to the affirmation of an item statement associated with activity limitation.” The RMDQ-23 was created by Patrick et al. (10) by removing 5 RMDQ-24 items and adding 4 new items from the original source of the RMDQ items – the 136-item Sickness Impact Profile (15). These 4 new items were believed by Patrick et al. to increase the responsiveness of the RMDQ-23, although this was not reported in statistical terms. In addition, the RMDQ-23 has the practical advantage of asking about both back pain and back-related leg pain, making it also suitable for people with sciatica.

Four studies have used Rasch analysis to investigate the psychometric properties of the RMDQ-24 (16–19). Their results concur that the RMDQ-24 does not fit the Rasch model due to misfitting and redundant items, although in some samples they came close to fitting. They also concur that there is poor targeting of items to person ability, which in this case means that it lacks sufficient items effectively to assess people with relatively little activity limitation and those with relatively high activity limitation. However, there is also some inconsistency in their findings, with little concordance as to which items misfit and which items indicate a lack of unidimensionality. Each of those RMDQ-24 studies analysed a sample from a different country; Australia, Norway, Turkey and the UK.

It may be that this inconsistency of Rasch results for the RMDQ-24 is the outcome of technical differences between the studies, the different Rasch statistical packages used, and normal sampling variation. However, the psychometric properties of an instrument can also vary depending on the social and clinical characteristics of the study, such as the culture, language and care-setting of the participants, their pain severity, co-morbidity, and the distribution of specific and non-specific LBP in the sample (person factors). The interpretability of results would be enhanced by studies that use the same Rasch methods to analyse samples from different settings and thereby provide more stable insights into RMDQ psychometric properties, as unlike traditional methods, the Rasch Model is able to investigate whether items are biased by person factors by applying Differential Item Functioning (DIF) analyses.

The RMDQ-23 has not been studied using Rasch analysis and it is not known whether its measurement properties are improved, compared with the RMDQ-24, due to its having 4 different questions. For example, 2 of the items identified as misfitting in the RMDQ-24 are not in the RMDQ-23. The measurement properties of the RMDQ-23 have been studied using traditional techniques and its internal consistency, concurrent validity, reproducibility and responsiveness found to be similar to other questionnaires measuring the same construct (7, 20). DIF in the RMDQ-23 has also been studied using a hybrid item response theory/logistic regression approach, which found that 18 of the 23 items displayed DIF on such person factors as age, education, marital status, employment status, surgical status and general health, though the authors conclude that this DIF had minimal impact on group-level scores (21). DIF occurs when the probability of endorsing an item varies across person factors, even when statistically controlling for the underlying construct being measured.

Therefore, the aims of this study were to use Rasch analysis to describe the psychometric properties of RMDQ-23 (DIF, unidimensionality, fit to the Rasch model, item difficulty, item misfit or overdiscrimination) and to determine how these were influenced by care setting and language/culture.

**METHODS**

**Samples**

The total sample comprised 1,000 patients with LBP: 500 from primary care in the UK, plus 250 from primary care and 250 from secondary care in Denmark. Patients with red flag diagnoses, such as cancer, fracture and spinal infection, were excluded. Ethics approval was obtained from the relevant local ethics committees, all participants received usual care from their clinicians and all participants provided written informed consent.

The UK cohort was a secondary analysis of data drawn from the Backpain Research in north Staffordshire (BaRnS) Study, in which 935 people, aged 30–59 years, were recruited from 5 general practices in North Staffordshire between October 2001 and October 2002. These general practices covered a heterogeneous population, both socio-economically and geographically. The sample included patients with non-specific LBP, of varying pain duration and different pain location (including pain radiating to the lower extremity) who were seeking care from general practitioners (GPs). Further details of patient recruitment and data collection are reported elsewhere (22). For the current study, data from a random sample of 500 people were selected.

The Danish secondary care sample consisted of consecutive, consenting patients attending the medical department of The Spine Centre of Southern Denmark, and originally matched the 935 sample size of the UK sample. This outpatient department is a non-surgical unit in a publicly-funded hospital that principally performs multidisciplinary assessment and treatment planning for primary care referrers (GPs, chiropractors and medical specialists) of patients with persistent spine pain. The Danish secondary care sample included patients seeking care with either non-specific LBP, stenosis or radiculopathy, of varying duration. Approximately 70% had pain radiating into the leg and approximately 47% had lower limb neurological signs. For the current study, data from a random sample of 250 people were selected and combined with those from Danish primary care to form the total Danish cohort of 500 people.

Patients self-completed questionnaires, included the RMDQ-23 (UK participants in the English language version (10) and Danish participants in the Danish version (24)), about the current back pain episode. *Pain intensity* was measured using numeric rating scales (scored 0–10) and re-coded as: no pain to mild pain (0–4), moderate...
pain (5–6) and severe pain (7–10), as recommended by Jensen et al. 
(25) and Zelman et al. (26). Pain duration was measured in the Danish 
cohort as months of duration and in the UK cohort as months since the 
last pain-free month, due to this difference in metric, was included in 
the Rasch analysis only in the Danish cohort. For the same reasons, 
previous episodes (coded as yes/no) was also included in the Rasch 
analysis only in the Danish cohort. Age was coded as <30, 30–39, 
40–49, 50–59, 60 years and above, as this classification allowed for 
sufficient numbers of participants in each group. 
The mean missing data on the RMDQ questions was 0.07% (range 
0–0.3%) and missing data in the descriptive variables ranged from 
0.5% to 2.8%. All the available data were used without imputation, 
as Rasch analysis is able to cope with missing data. 

Statistical analysis 
Data were analysed with SPSS version 16 (SPSS Statistics/IBM Corp, 
Chicago IL, USA) and RuMM2030 (RuMM Laboratory P/L, Duncraig 
WA, Australia). First, the frequency of responses, including missing data, 
for each item was analysed. Then the RMDQ-23 was assessed for 
its unidimensionality, item fit, redundancy and DIF using a dichoto-
mous Rasch model (8, 27, 28). This model assumes that as a patient’s 
activity limitation increases, the probability of a maximum score on 
the item increases. Patient 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 probability of answering “yes”. The adequacy 
of the fit of each item to the Rasch model was assessed as follows: 

Overall fit for the model for the scale. This was given by $\chi^2$ item-trait 
interaction statistic. A non-significant probability value indicates 
that there is no substantial deviation from the model and that the 
hierarchical ordering of the items is consistent across all levels of 
the underlying trait. 

Power of the scale to discriminate amongst respondents with different 
levels of activity limitation. This is indicated by the person separation 
index, which is equivalent to Cronbach’s alpha, as values of 0.8 and 
0.9 indicate that the scale can discriminate statistically between at 
least 3 and 4 groups, respectively (28). 

Overall person fit and item fit. These were assessed by inspecting the 
mean and standard deviation (SD) of the fit residuals. A mean value 
of approximately 0.0 and a SD of 1.0 were expected. Misfitting items 
were identified by fit residuals of greater than plus or minus 2.5 or 
a significant $\chi^2$ probability value (8, 29). High negative residuals 
are normally interpreted to indicate items are displaying an overdis-
crimination, which can occur for a variety of reasons, including item 
redundancy or item dependency. High positive residuals are interpreted 
as indicating misfitting items. 

Differential Item Functioning. DIF was investigated to identify items 
that operated differently for people of the same level of ability (i.e. 
having the same amount of activity limitation), but who differ on some 
other clinical feature. DIF by age (<30, 30–39, 40–49, 50–59, 60 years 
and above), gender, previous episodes (no, yes), pain intensity (no 
pain to mild pain, moderate pain and severe pain), country (Denmark, 
UK), and in the Danish cohort only, care-setting (primary, secondary 
care) were investigated and considered significant if the $\chi^2$ probability 
value was less than 0.05 (Bonferroni adjusted within RuMM2030 
<0.0022, (0.05/23)). Two types of DIF can be identified: uniform 
and non-uniform DIF. A uniform DIF indicates that 1 subgroup, such 
as males, is displaying a consistently greater ability to confirm an 
item than another subgroup (analysis of variance (ANOVA) main 
effect), such as females. A non-uniform DIF indicates that the ability 
differences are inconsistent amongst the subgroups (ANOVA interac-
tion effect). 

The presence of true DIF (30) was examined by running separate 
analyses on the 2 samples, sequentially splitting items displaying 
most DIF, beginning with the one showing greatest DIF. If an analysis 
resolving item DIF also removed DIF in other items, this confirmed 
artificial DIF in those other items, and we discarded this as an artefact 
of the DIF procedure. Only true DIF is reported. 

Unidimensionality. This was tested using the t-test method, as de-
scribed by Smith (31). 

Local item dependency. This was identified where any item displayed 
a correlation of item residuals that was 0.3 or more above the mean 
for all the items. As local response dependency can disrupt fit, inflate 
reliability estimates and result in spurious multimensionality, we also 
tested the effect of clustering locally dependant items into “testlets”, 
which created polyotomous items (super items) from those original 
dichotomous items (32). 

All analyses were conducted separately on each of the 2 cohorts 
and these samples were not pooled to carry out a cross-cultural DIF 
analysis due to the psychometric problems observed within each 
country sample. Initially, the Rasch models were run with all available 
participants and then repeated without the people whose scores were 
located either at the floor or the ceiling of the scale (extreme scores), 
with the differences being reported. Extreme scores indicate that those 
people’s activity limitation was beyond the range measured by the 
cale. We identified extreme score people by use of the default setting 
in RuMM2030, which is an “extreme person criterion” of 0.220. 

Sample-size calculations 
Linacre (29) proposed that the minimum sample size should be 20 
participants for every questionnaire item, and a preliminary analysis 
of some of our data using randomly selected and mutually exclusive 
samples of approximately 200 people showed there was considerable 
variability of results across samples of that size. We therefore opted 
to analyse a cohort of 500 people from each country, recognizing that 
this may have increased the possibility of over-identifying DIF (type 
et 2 error), despite Bonferroni adjustment. 

RESULTS 
Table I describes the demographic similarities and differences 
between the samples. Predictably and desirable, there was 
variability between the samples on almost all the demographic 
variables, due to differences in care-seeking patterns between 
countries and the different care settings. 

Endorsement frequency and item difficulty 
The endorsement frequencies of the RMDQ-23 in each cohort 
are shown in Table II. The difficulty order of the 23 items 
is reflected by the logit measure and suggests the relative 
degree of activity limitation for each item. In both cohorts, 
the item with the highest endorsement (73.8% UK, 92.6% 
Denmark) and lowest logit value (−2.35 UK, −2.80 Denmark) 
was item 2 (“I change positions frequently to try and get my 
back or leg comfortable”), as it was endorsed even by people 
with relatively low activity limitation. In contrast, the item with 
the lowest endorsement (3.2% UK, 8.8% Denmark) and 
highest logit value (4.43 UK, 3.44 Denmark) was item 19 (“I 
stay in bed most of the time because of my back or leg pain 
(scatica”) as it was endorsed only by people with relatively 
high activity limitation. 

Fig. 1 shows the item maps for both cohorts with the 23 
items plotted to the right and the number of people to the left
Rasch analysis of the RMDQ-23 of the vertical line (which reflects the difficulty level). The items endorsed by only the least physically able people with the most activity limitation are at the top (with increasing positive logit value) and the items endorsed even by the more physically able people with less activity limitation are at the bottom (increasingly negative logit value). Although the distribution of activity limitation was different in each cohort (item difficulty UK cohort mean –0.66 (SD 1.99), Danish cohort 0.39).

Table I. Characteristics of the UK and Danish samples

<table>
<thead>
<tr>
<th>Item Statement</th>
<th>UK primary care sample (n = 500)</th>
<th>Denmark primary care sample (n = 250)</th>
<th>Denmark secondary care sample (n = 250)</th>
<th>Differences between samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years, median (IQR) [range]</td>
<td>46 (39–53) [30–59]</td>
<td>47 (39–55) [18–71]</td>
<td>40 (33–50) [2–81]</td>
<td>Danish secondary care &lt; the others&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Pain intensity (NRS 0–10), median (IQR) [range]</td>
<td>4 (2–6) [0–10]</td>
<td>7 (6–8) [1–10]</td>
<td>6 (5–8) [0–10]</td>
<td>UK primary care &lt; the others, Danish secondary care &lt; Danish primary care&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Activity limitation (RMDQ 0–23), median (IQR) [range]</td>
<td>8 (4–15) [0–23]</td>
<td>14 (10–18) [0–22]</td>
<td>12 (8–17) [0–23]</td>
<td>UK primary care &lt; the others, Danish secondary care &lt; Danish primary care&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Gender, female</td>
<td>56.0</td>
<td>56.0</td>
<td>51.2</td>
<td>No statistically significant differences&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Duration of this pain episode, less than 3 months since onset of episode, %</td>
<td>21.9</td>
<td>66.7</td>
<td>17.3</td>
<td>Danish secondary care &lt; Danish primary care&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Reported previous low back pain episodes, %</td>
<td>82.4</td>
<td>87.0</td>
<td>70.5</td>
<td>Danish secondary care &lt; the others&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>Kruskal-Wallis with Mann-Whitney post-hoc comparisons, indicating which samples have statistically significant lower scores than other samples.

<sup>b</sup>Bonferroni-adjusted χ<sup>2</sup>.

NRS: numeric rating scale; RMDQ: Roland Morris Disability Questionnaire.

Table II. endorsement frequency and fit statistics, from Rasch analyses of the 23-item version of the Roland Morris Disability Questionnaire cohorts

<table>
<thead>
<tr>
<th>Item Statement</th>
<th>UK cohort (n = 500)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Danish cohort (n = 500)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>UK cohort (n = 500)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Danish cohort (n = 500)&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Statement</td>
<td>Endorsement n (%)</td>
<td>Logit measure&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Fit residual&lt;sup&gt;d&lt;/sup&gt;</td>
<td>χ&lt;sup&gt;2&lt;/sup&gt; p-value&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>1 Stay home</td>
<td>112 (22.4)</td>
<td>1.40</td>
<td>–2.14</td>
<td>0.0288</td>
</tr>
<tr>
<td>2 Change position frequently</td>
<td>369 (73.8)</td>
<td>–2.35</td>
<td>0.39</td>
<td>0.1634</td>
</tr>
<tr>
<td>3 Walk more slowly</td>
<td>258 (51.6)</td>
<td>–0.94</td>
<td>–3.29&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.0009&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>4 Not doing jobs around the house</td>
<td>146 (29.2)</td>
<td>0.75</td>
<td>–1.30</td>
<td>0.4463</td>
</tr>
<tr>
<td>5 Use a handrail up stairs</td>
<td>148 (29.6)</td>
<td>0.70</td>
<td>–1.55</td>
<td>0.4170</td>
</tr>
<tr>
<td>6 Hold onto something to get out of easy chair</td>
<td>175 (35.0)</td>
<td>0.31</td>
<td>–2.16</td>
<td>0.0844</td>
</tr>
<tr>
<td>7 Dressed more slowly</td>
<td>213 (42.6)</td>
<td>–0.25</td>
<td>–0.40</td>
<td>0.6646</td>
</tr>
<tr>
<td>8 Stand for short periods</td>
<td>193 (38.6)</td>
<td>0.00</td>
<td>1.35</td>
<td>0.6009</td>
</tr>
<tr>
<td>9 Try not to bend or kneel down</td>
<td>247 (49.4)</td>
<td>–0.71</td>
<td>–0.62</td>
<td>0.9515</td>
</tr>
<tr>
<td>10 Difficult to get out of a chair</td>
<td>159 (31.8)</td>
<td>0.54</td>
<td>–0.81</td>
<td>0.4749</td>
</tr>
<tr>
<td>11 Back or leg painful almost all the time</td>
<td>211 (42.4)</td>
<td>–0.62</td>
<td>1.49</td>
<td>0.3080</td>
</tr>
<tr>
<td>12 Difficult to turn over in bed</td>
<td>245 (49.0)</td>
<td>–0.06</td>
<td>1.26</td>
<td>0.6307</td>
</tr>
<tr>
<td>13 Trouble putting on shoes or socks</td>
<td>239 (47.8)</td>
<td>–0.58</td>
<td>–0.56</td>
<td>0.9322</td>
</tr>
<tr>
<td>14 Only walk short distances</td>
<td>166 (33.2)</td>
<td>0.43</td>
<td>–1.64</td>
<td>0.1035</td>
</tr>
<tr>
<td>15 Sleep less well</td>
<td>283 (56.6)</td>
<td>–1.06</td>
<td>2.29</td>
<td>0.0050</td>
</tr>
<tr>
<td>16 Avoid heavy jobs around the house</td>
<td>301 (60.2)</td>
<td>–1.51</td>
<td>1.48</td>
<td>0.1079</td>
</tr>
<tr>
<td>17 More irritable and bad tempered</td>
<td>190 (38.0)</td>
<td>0.06</td>
<td>1.45</td>
<td>0.0461</td>
</tr>
<tr>
<td>18 Go upstairs more slowly</td>
<td>166 (33.2)</td>
<td>–0.04</td>
<td>1.26</td>
<td>0.6307</td>
</tr>
<tr>
<td>19 Stay in bed most of the time</td>
<td>16 (3.2)</td>
<td>4.43</td>
<td>–0.56</td>
<td>0.6403</td>
</tr>
<tr>
<td>20 Decreased sexual activity</td>
<td>187 (37.4)</td>
<td>0.12</td>
<td>–0.62</td>
<td>0.6307</td>
</tr>
<tr>
<td>21 Keep rubbing or holding areas of my body that hurt</td>
<td>276 (55.2)</td>
<td>–0.59</td>
<td>2.34</td>
<td>0.0748</td>
</tr>
<tr>
<td>22 Doing less of daily work around the house</td>
<td>219 (43.8)</td>
<td>–0.34</td>
<td>–4.41&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.0001&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>23 Express concern about my health</td>
<td>127 (25.4)</td>
<td>0.88</td>
<td>1.82</td>
<td>0.0000&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>51 with extreme scores. 449 valid scores. 8 class intervals.

<sup>b</sup>12 with extreme scores. 488 valid scores. 7 class intervals.

<sup>c</sup>Higher score logits represent harder items to endorse, that is, activities that were difficult (answered “yes”) only for people with the most activity limitation due to low back pain.

<sup>d</sup>Fit residuals with greater than plus or minus 2.5 are marked; negative values indicate overdiscrimination for reasons including redundant or dependant items. Positive values misfitting items.

<sup>e</sup>χ<sup>2</sup> probability value (p-value) < 0.002 (0.05/23).
there was the same pattern of there being some poor distribution of items, with the the RMDQ-23 lacking items towards either end of the difficulty scale.

**Item misfit and overdiscrimination**

Table II also shows that, in the UK cohort, one of the 23 items (item 23) displayed evidence of item misfit ($\chi^2$ $p$-values $<0.002$) and 3 items (items 3, 18 and 22) displayed evidence of overdiscrimination (fit statistics $<-2.5$). In the Danish cohort, 4 items (items 15, 17, 21 and 23) displayed evidence of item misfit (fit statistics $>2.5$ or $\chi^2$ $p$-values $<0.002$) and 3 items (items 3, 7 and 14) displayed evidence of item overdiscrimination (fit statistics $<-2.5$).

**Fit to the Rasch model**

In both cohorts, there was significant deviation ($p<0.0001$) from the model on the $\chi^2$ item-trait interaction statistics (see Table III). However, the person separation reliability index (PSI) values approximated 0.9, indicating an ability to discriminate between participants with 3 or 4 different levels of disability.

In the UK cohort, 51 of 500 had extreme scores. The fit statistics of the data with the misfitting people, were PSI of 0.873, mean person location --0.67 (SD 1.93), and without the misfitting people were PSI 0.867, mean person location --0.40 (SD 1.63). In the Danish cohort, 12 of the 500 had extreme scores. The fit statistics with the misfitting people were PSI 0.846, mean person location 0.38 (SD 1.50), and without the misfitting people were PSI 0.831, mean person location 0.39 (SD 1.36).

The $t$-tests showed evidence of RMDQ-23 multidimensionality in both cohorts (Table III).

**Differential Item Functioning**

Systematic DIF by care setting (tested Danish cohort only) was found for 2 items: items 7 and 13 (uniform DIF). DIF by age was found for item 17 (uniform DIF) in both the UK and Danish cohorts, DIF by gender for item 12 (uniform DIF) in both cohorts, and DIF by pain intensity was found for items 1 and 11 (uniform DIF) in the UK cohort and item 23 (uniform DIF) in the Danish cohort. Systematic DIF by pain duration and previous back pain episodes was also tested in the Danish cohort only, with items 7 and 13 showing DIF by pain duration (uniform DIF), but no items showed DIF by previous episodes.

Additional detail of the characteristics of DIF found within each of the 2 cohorts is reported in Table IV and an overall summary of RMDQ-23 item performance across both cohorts displayed in Table V. The only consistency across the cohorts was that females were more likely to answer yes to having difficulty turning over in bed, and that people of various ages differed in their being likely to answer yes to being more irritable and bad-tempered, indicating that all other DIF findings were specific to one country.

**Local item dependency**

Local item dependency was identified between 3 pairs of items in the UK cohort and 4 pairs of items in the Danish cohort (Table III). Clustering locally dependent items into testlets and re-running the analysis resulted in evidence of unidimensionality and fewer misfitting/overdiscriminating items, but the RMDQ23 still did not fit the Rasch model, as the fit residuals and $\chi^2$ results were still unsatisfactory (Table VI).

**DISCUSSION**

The aims of this study were to use Rasch analysis to describe the psychometric properties of RMDQ-23 (DIF, unidimensionality, fit to the Rasch model, item difficulty, item misfit and overdiscrimination/redundancy) and to determine how these were influenced by care setting and language/culture. The first finding was that the RMDQ-23 is a unidimensional measure of activity limitation, if local item dependency issues are accommodated. Furthermore, items 3 and 15 in the UK cohort and items 14, 15, 17, 21 and 23 in the Danish cohort showed misfit.
Table III. Detailed results from Rasch analyses of UK and Danish cohorts of the 23-item version of the Roland Morris Disability Questionnaire

<table>
<thead>
<tr>
<th>Fit to model</th>
<th>UK primary care cohort n = 500a</th>
<th>DK primary and secondary care cohort n = 500b</th>
</tr>
</thead>
<tbody>
<tr>
<td>χ², p-value</td>
<td>Highly significant deviation from Rasch model χ² = 272.46 (df = 161), p &lt; 0.0001</td>
<td>Highly significant deviation from Rasch model χ² = 310.68 (df = 138), p &lt; 0.0001</td>
</tr>
<tr>
<td>Multidimensionality, t-tests, % (CI)c</td>
<td>6.7 (4.7–8.7) Misfitting items Item 23 (χ² p-value 0.0000)</td>
<td>8.0 (6.1–9.9) Misfitting items Item 15 (χ² p-value 0.0001) Item 17 (fit residual 2.92) Item 21 (fit residual 4.44 and χ² p-value 0.0000) Item 23 (fit residual 4.05 and χ² p-value 0.0000)</td>
</tr>
<tr>
<td>Item misfit</td>
<td>Misfitting items</td>
<td>Misfitting items</td>
</tr>
<tr>
<td>Local dependencies</td>
<td>Items: 4 and 22 5 and 18 7 and 13</td>
<td>Items: 3 and 18 6 and 10 7 and 13 16 and 22</td>
</tr>
<tr>
<td>Systematic DIFd by care setting</td>
<td>Not tested</td>
<td>Uniform DIF for items 7, 13</td>
</tr>
<tr>
<td>Systematic DIF by age</td>
<td>Uniform DIF for item 17</td>
<td>Uniform DIF for items 17</td>
</tr>
<tr>
<td>Systematic DIF by gender</td>
<td>Uniform DIF for item 12</td>
<td>Uniform DIF for item 12</td>
</tr>
<tr>
<td>Systematic DIF by pain intensity</td>
<td>Uniform DIF for items 1 and 11</td>
<td>Uniform DIF for item 23</td>
</tr>
<tr>
<td>Systematic DIF by pain duratione</td>
<td>Not tested</td>
<td>Uniform DIF for items 7 and 13</td>
</tr>
<tr>
<td>Systematic DIF by previous episodesf</td>
<td>Not tested</td>
<td>No DIF</td>
</tr>
</tbody>
</table>

All results based on a Bonferroni adjusted p-value < 0.0022.

*500 included, 51 with extreme scores, 449 valid scores, 8 class intervals.

*500 included, 12 with extreme scores, 488 valid scores, 7 class intervals.

t-test method as described by Smith (31).

High negative fit residuals indicate overdiscriminating items, that may be due to item redundancy or to other reasons, including item dependancy.

Only real DIF is reported (tested by the method described by Andrich et al. 2011 (30)). No non-uniform DIF was found.

Only comparably collected in the Danish samples.

DIF: Differential Item Functioning; CI: confidence interval.

Table IV. Summary of real differential item functioning (DIF)

<table>
<thead>
<tr>
<th>DIF by:</th>
<th>UK cohort</th>
<th>Danish cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care setting</td>
<td>Not tested</td>
<td>Patients in primary care had a higher probability of affirming item 7 [dressed more slowly] and item 13 [trouble putting on shoes or socks]</td>
</tr>
<tr>
<td>Age group</td>
<td>Patients aged 40–49 years had a higher probability of affirming item 17 [More irritable and bad tempered]</td>
<td>Patients aged 30–39 had higher probability of affirming item 17 [more irritable and bad tempered]</td>
</tr>
<tr>
<td>Gender</td>
<td>Female patients had a higher probability of affirming item 12 [difficult to turn over in bed]</td>
<td>Female patients had a higher probability of affirming item 12 [difficult to turn over in bed]</td>
</tr>
<tr>
<td>Pain intensity</td>
<td>Patients with high pain intensity had a higher probability of affirming item 1 [stay home], whereas patients with mild to moderate pain intensity had a higher probability of affirming item 11 [difficult to get out of a chair]</td>
<td>Patients with no to mild pain have higher probability of affirming item 23 [express concern about my health]</td>
</tr>
<tr>
<td>Pain duration</td>
<td>Not tested</td>
<td>Patients with pain duration of 0–2 weeks had higher probability of affirming item 7 [dressed more slowly] and item 13 [trouble putting on shoes or socks]</td>
</tr>
</tbody>
</table>

Abbreviated item wording in square brackets.
Table V. Summary of 23-item version of the Roland Morris Disability Questionnaire item performance in the UK and Danish cohorts

<table>
<thead>
<tr>
<th>Item</th>
<th>Statement</th>
<th>Local dependence</th>
<th>Real DIF by Care setting</th>
<th>Age</th>
<th>Gender</th>
<th>Pain intensity</th>
<th>Pain duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stay home</td>
<td>Redund (UK, DK)</td>
<td>To item 10 (DK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Change position</td>
<td></td>
<td>To item 18 (DK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Walk more slowly</td>
<td>Redund (UK)</td>
<td>To item 22 (DK)</td>
<td></td>
<td></td>
<td>Uniform (DK)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Not doing jobs around the house</td>
<td></td>
<td>To item 18 (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Use a handrail up stairs</td>
<td></td>
<td>To item 18 (UK)</td>
<td></td>
<td></td>
<td>Uniform (DK)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Hold onto something to get out of easy chair</td>
<td></td>
<td>To item 22 (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Dressed more slowly</td>
<td></td>
<td>To item 13 (UK)</td>
<td></td>
<td></td>
<td>Uniform (DK)</td>
<td>Uniform (DK)</td>
</tr>
<tr>
<td>8</td>
<td>Stand for short periods</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>9</td>
<td>Try not to bend or kneel down</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>10</td>
<td>Difficult to get out of a chair</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>11</td>
<td>Back or leg painful almost all the time</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>12</td>
<td>Difficult to turn over in bed</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>13</td>
<td>Trouble putting on shoes or socks</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>14</td>
<td>Only walk short distances</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>15</td>
<td>Sleep less well</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>16</td>
<td>Avoid heavy jobs around the house</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>17</td>
<td>More irritable and bad tempered</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>18</td>
<td>Go upstairs more slowly</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>19</td>
<td>Stay in bed most of the time</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>20</td>
<td>Decreased sexual activity</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>21</td>
<td>Keep rubbing or holding areas of my body that hurt</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>22</td>
<td>Doing less of daily work around the house</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
<tr>
<td>23</td>
<td>Express concern about my health</td>
<td></td>
<td>Uniform (UK)</td>
<td></td>
<td></td>
<td>Uniform (UK)</td>
<td>Uniform (UK)</td>
</tr>
</tbody>
</table>

Misfit: misfitting item; Redund: redundancy or dependency; Uniform: systematic real uniform; DIF: Differential Item Functioning.

Table VI. Comparison 23-item version of the Roland Morris Disability Questionnaire item performance, when locally dependent items were clustered into “testlets”

<table>
<thead>
<tr>
<th></th>
<th>UK cohort (n = 500)</th>
<th>Danish cohort (n = 5,500)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original analysis</td>
<td>Analysis with testlets</td>
</tr>
<tr>
<td>Items fit residual, mean (SD)</td>
<td>-0.64 (1.82)</td>
<td>4 and 22, 5 and 18, 7 and 13</td>
</tr>
<tr>
<td>Person fit residual, mean (SD)</td>
<td>-0.14 (0.57)</td>
<td>-0.55 (1.64)</td>
</tr>
<tr>
<td>$\chi^2$ interaction</td>
<td>272.25 (df=161)</td>
<td>230.61 (df=140)</td>
</tr>
<tr>
<td>Unidimensionality, % tests &lt;5% (95% CI)</td>
<td>6.7 (4.7–8.7)</td>
<td>4.9* (2.9–6.9)</td>
</tr>
<tr>
<td>Items displaying misfit</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Items displaying overdiscrimination</td>
<td>3, 18, 22</td>
<td>3, 6*</td>
</tr>
</tbody>
</table>

*aThe assumption of unidimensionality was fulfilled when locally dependent items were clustered.
*bThe number of misfitting or overdiscriminating items was reduced when locally dependent items were clustered.
to the Rasch model. For items 17, 21 and 23 this is a plausible finding as these items (“I am more irritable and bad tempered with people than usual”, “I keep rubbing or holding areas of my body that hurt or are uncomfortable” and “I often express concern to other people over what might be happening to my health”) do not appear to have face validity as measures of activity limitation. While this finding differs from previous Rasch analysis of the RMDQ-24, in that items 21 and 23 are unique to the RMDQ-23 version, it was not consistent, as only item 15 displayed misfit in both cohorts. That item misfit was not eliminated by accommodating local item dependency indicates that misfit was caused by other factors.

The second finding was that one item consistently showed overdiscrimination (item 3) and 4 items inconsistently showed overdiscrimination (items 18 and 22 in the UK cohort only, and items 7 and 14 in the Danish cohort only). This result is also reflected in Fig. 1, where most of these items bunch together with other items that have the same item difficulty, suggesting some item redundancy. There was also some poor targeting of items to person ability, specifically that the RMDQ-23 has few items for people with relatively little activity limitation and also for people with relatively high activity limitation. This mirrors the non-ideal targeting of items that other authors have found during Rasch analysis of the RMDQ-24 (16–18).

An additional finding was that 4 items displayed real DIF in the UK cohort and 5 items in the Danish cohort, with only 2 items showing consistent DIF across countries (items 12 by gender and 17 by age). In the Danish cohort, real systematic DIF by care setting was found for 2 items, indicating that people seeking care in Denmark responded culturally or linguistically in different ways to these items, depending on the care setting.

A strength of this study is that, to our knowledge, it is the first to use Rasch analysis to investigate the measurement properties of the RMDQ-23. DIF has only previously been studied in the RMDQ-23 using a non-Rasch approach on data from the USA, finding that 18 of the 23-items displayed DIF on such factors as age, education, marital status, employment status, surgical status and general health scores (21). All of the items that displayed DIF in our cohorts also displayed DIF in the US study.

Other strengths of the current study are the size of the cohorts and the ability to quantify the measurement properties of the RMDQ-23 questionnaire across 2 cultures/languages, 2 types of care-settings, and in people with diverse episode duration. However, it would also be ideal to replicate the current findings in other samples, such as other languages, cultures and other types of patients, such as surgical patients. As there have been no previously published Rasch analyses of the RMDQ-23, direct comparisons of our results with previous findings are constrained to those involving the RMDQ-24. A limitation of the study is that the available data did not allow analysis of the separate influences of culture and language.

In conclusion, this Rasch analysis showed that the measurement properties of the RMDQ-23 are not ideal. It is unidimensional if local item dependency issues are accommodated, but it contains evidence of misfitting or overdiscriminating items, has sparse items to adequately assess people with relatively little or relatively high activity limitation, and the scoring of most items is affected by 1 or more common clinical characteristics (age, gender, pain intensity, care setting and country). These results are similar to those from Rasch analysis of the RMDQ-24 (16–19), which have also shown that version of the questionnaire to contain some misfitting items and to have poor targeting of items with high and low activity limitation. Therefore, in our view it is timely to reconsider whether: (i) the RMDQ should be reconstructed using an instrument development approach based on item-response theory, including the consideration of different items and response options; or (ii) the use of alternative questionnaires should be recommended, such as the Oswestry Disability Index, that have shown evidence of fitting the Rasch model (33); or (iii) a completely new questionnaire should be developed, perhaps utilizing a computer adaptive testing delivery platform (34).

ACKNOWLEDGEMENTS

The authors would like to thank Lise Hestbæk for access to the Danish primary care data used in this study.

Kate M. Dunn is funded through a Research Career Development Fellowship from the Wellcome Trust [083572]. Peter Kent is partially funded by a grant from the Danish Foundation for Chiropractic Research and Postgraduate Education. No benefits in any form have been, or will be, received from a commercial party related directly or indirectly to the subject of this manuscript.

REFERENCES


J Rehabil Med 47


