Measurement properties of the Disability Rating Index in patients undergoing hip replacement

Helen Parsons1, Julie Bruce2, Juul Achten2, Matthew L. Costa2 and Nicholas R. Parsons1

Abstract

Objective. The aim of this study was to establish and validate the measurement properties of the Disability Rating Index (DRI) in a population of adults undergoing hip replacement.

Methods. One hundred and twenty-six adults participating in a randomized controlled trial completed the Oxford Hip Score, Harris Hip Score, DRI and EuroQol Group–Five Dimensions (EQ-5D) questionnaires at four time points. The structural validity of the DRI was assessed using principal component analysis. Cronbach’s α was used to determine the internal consistency and scale reliability was also assessed. Correlation between the DRI and the other functional and health-related quality of life scales was used to check criterion validity. DRI responsiveness was estimated and the interpretability of the scale was also assessed by checking for edge effects.

Results. Results of analyses showed that the DRI was internally consistent (Cronbach’s α = 0.92), had good association with both function-specific and general health-related quality of life scores and was sensitive to change (smallest detectable change = 2.7). No evidence of edge effects was found. Furthermore, structural assessment of the DRI revealed two novel subscales representing simple tasks and difficult tasks.

Conclusions. The DRI is structurally valid, responsive and concurs with functional assessment in adults undergoing hip replacement.

Key words: disability evaluation, patient reported outcomes, validity, DRI.

Introduction

Accurate assessment of health-related quality of life and function from the patient’s perspective is vital when determining the effectiveness of health interventions. Therefore it is important to use measurement scales that have been correctly tested and validated to avoid the introduction of bias [1, 2]. While many guidelines for validation are available [3–6], it is often difficult to establish whether a particular scale has been rigorously validated and, if so, for which patient populations [7].

This article focuses on the Disability Rating Index (DRI), a patient-reported outcome measure developed by Salén et al. in 1994 [8], with the aim of measuring physical disability of respondents within a clinical setting. The DRI evaluates a respondent’s disability by assessing activity and participation limitations. The DRI has been used in patients undergoing lumbar disc surgery [9], patients undergoing rehabilitation [10], women with pelvic girdle pain [11], patients with upper extremity disorders [12] and patients with chronic pain [13]. The DRI is also a core measure within the Swedish Quality Registry for Pain Rehabilitation [14]. The Salén et al. [8] study validated the DRI on healthy persons (n=1092) and a heterogeneous clinical population (n=366) with varying levels of disability, including elderly arthritis patients waiting for hip (n=23) or knee (n=24) replacement surgery and wheelchair-bound patients with multiple sclerosis (n=16). The validation consisted of assessing test–retest reliability, inter- and intrarater reproducibility, internal consistency, construct validity (including face validity) and responsiveness. The scale was found to be highly reliable (Cronbach’s α = 0.95 in a 1 day interval test), acceptable...
(high compliance in all populations) and well correlated ($\rho=0.69$) with the patient’s rating of their disability. However, the authors did not investigate the structure of the instrument using factor analysis of clinically relevant changes or consider whether any subscales may exist. In their systematic review, Grotle et al. [15] recommended that the DRI could be used without further validation for patients with lower back pain, but noted that assessment of the factor structure was inadequate.

Here we aim to validate the DRI in a population of younger adults undergoing hip replacement surgery. After a hip replacement, patients value improvements in their ability to participate in activities of daily life, alongside a reduction in pain and stiffness in the hip [16–18]. Therefore the DRI may be a useful measurement tool in this group of patients, and indeed for many other patient groups in whom participation in daily activities is a key outcome. The World Health Organization defines disability as an umbrella term for impairments, activity limitations and participation restrictions [19], also implying that disability rating could be useful in a wide range of different patient groups. We examined the DRI for internal consistency and structure using principal component analysis (PCA) and investigated the minimum clinically important change in the DRI occurring after surgical intervention for hip arthritis. The international initiative OMERACT [2] aims to improve outcome measurement in rheumatology and scrutinizes measures based on three criteria: truth, discrimination and feasibility [20]. Our analyses will help establish both the truth and discrimination criteria.

**Methods**

**Patient sample**

All outcome data were collected within the Warwick Arthroplasty Trial (WAT) (trial registration ISRCTN 33354155, UKCRN 4093), described in full elsewhere [21, 22]. Briefly, 126 patients, aged >18 years (range 40–69; median age 57.5), 52 (41%) of whom were female, were recruited to compare the effectiveness of total hip replacement with resurfacing hip replacement in patients with severe hip arthritis. Patients were recruited from a large university teaching hospital in the UK. The WAT study was approved by the Coventry Research Ethics Committee (no. 07/Q2802/26) on 9 May 2007.

The primary outcome measurements were the Oxford Hip Score (OHS) [16] and the Harris Hip Score (HHS) [23]; secondary outcome measures included the DRI and the EuroQol Group–Five Dimensions (EQ-5D) [24]. Data were collected at baseline (preoperatively) and at 3, 6 and 12 months after surgery.

**The DRI**

The DRI questionnaire is self-administered and quick to complete (<5 min); it is a single scale comprising 12 items measured using a visual analogue scale (VAS (0–100)). Low scores denote little or no disability. The items are broadly grouped into three distinct sections: (i) questions 1–4, basic activities of daily life: dressing, outdoor walks, climbing stairs, sitting a long time; (ii) questions 5–8, daily physical activities: standing bent over a sink, carrying a bag, making a bed, running; and (iii) questions 9–12, work-related/more vigorous activities: light work, heavy work, lifting heavy objects and participating in exercise/sports.

**Comparison instruments**

**HHS**

The HSS [23] is a clinician-based outcome measure developed in the 1960s to assess function before and after hip surgery. The maximum score (best function) is 100.025, while the minimum score (worst function) is 0. A total score <70 is considered indicative of poor function, 70–79 is considered fair, 80–89 is good and 90–100 is excellent. The tool is known to suffer from ceiling effects [25]. For the WAT study, a change of 7 points was considered to be a minimal clinically important difference (MCID) [21].

**OHS**

The OHS [16] is a 12-item patient-reported outcome measure designed to assess hip function. At the point of administration, patients are asked to recall hip function (six items) and hip pain (six items, including overall pain level) over the last 4 weeks. Each item is scored from 0 (worst function) to 4 (best function), with the overall score ranging from 0 to 48, where 48 represents the best function. A total score of >41 represents excellent function, 34–41 is good, 27–33 is fair and <27 is poor [26]. A change of 5 points was defined to be the MCID in this population [21].

**EQ-5D**

The 3 level EQ-5D (EQ-5D-3L) was developed by the EuroQol Group [24] to give a standard measure of health-related quality of life. It is a patient-reported outcome measure and is split into two components: a health status classification (five-item score) and self-rated health (scored by a VAS). The five domains surveyed for health status are mobility, self-care, usual activities, pain and anxiety. Responses to the health status classification system are converted into an overall score using a utility algorithm for the UK population [27]. Health status is anchored at 1 (perfect health) to 0 (death); for the UK, health states range from 1.0 (no health problems) to −0.594 (severe health problems). A change of ~0.07 points was defined to be the MCID for the EQ-5D [28].

**Statistical methods**

All analyses were conducted using the statistical software R (R Project for Statistical Computing, Vienna, Austria) [29]. Data summaries and descriptive statistics were estimated at all data collection points. Because of the small amount of missing trial data (~4% of all possible participant responses were missing, with no incomplete responses), all analyses were conducted using complete cases only.

**Structural validity**

Structural validity is concerned with the construction of a measurement instrument, identifying underlying elements
such as the existence of subscales and examining the behaviour of individual items that comprise the instrument. Conceptually it is a subtype of construct validity, in that it is a check by which all aspects of the latent traits are measured. Internal consistency assesses the similarity of items within a test and the relation with the overall score. Terwee et al. [3] suggest that analyses should be performed on a sample size of seven times the number of items (this equates to 84 for the DRI) or 100 participants, whichever is greater.

As no information was available on the structure of the DRI, PCA was utilized to investigate the factor composition of the DRI. To ensure that the instrument structure was not biased by repeated responses, it was decided that each participant should be entered into the structural analysis once. However, limiting the analysis to one particular time point was not deemed to be appropriate, as not all levels of hip function were represented at a single time point (i.e. respondents entered the study with low levels of function and improved postoperatively). Hence, for each participant in the WAT study (n = 126), an instance of the DRI was randomly sampled from the four time points (baseline, 3 months, 6 months and 12 months). This yielded a subsample consisting of data from 34 study participants at baseline (27%), 38 at 3 months (30%), 29 at 6 months (23%) and 25 at 12 months (20%). The mean DRI score in the subsample was 37.6 (range 0–91.6). The subsample was then entered into a model with varmax rotation. A scree plot was used to determine the appropriate number of PCs. Absolute item loadings > 0.45 were considered to represent sufficient correlation with a PC to warrant inclusion within a subscale.

Internal consistency

Using the subsample, the internal consistency of the whole DRI, as well as any identified subscales (extracted PCs), were calculated using Cronbach’s $\alpha$. Values between 0.70 and 0.95 were considered acceptable, indicating that items have good to excellent agreement with the score total [6].

Criterion validity

Criterion validity assesses whether the instrument has good correlation with other externally validated and well-accepted measures [3] (where possible, the gold standard measurement), with defined hypotheses stated a priori [6]. To investigate the external validity of the DRI, Pearson’s correlation coefficient was used to assess the association between the DRI and the three other instruments, with an expectation of a moderate to high positive correlation ($\rho \geq 0.7$), as poor hip function was likely to be the main cause of disability for this patient group and poor health-related quality of life is associated with disability.

Reliability, responsiveness and interpretability

The intraclass correlation coefficient (ICC) was used to measure reliability through agreement for continuous measures [3, 4, 6], where reliability is defined as the amount that a scale is influenced by random error [3]. Reliability is ideally calculated using a test-retest format in a short time frame [6], but here we compare stable respondents between 6 and 12 months; a respondent was considered stable if the change in clinician-reported HHS between the two time points was less than the MCID (7 points). This subsample of 66 patients (56% of the valid total cohort) was used to assess reliability (ICC) and the absolute measurement error or agreement (standard error of the model; SEM) using a two-way random effects model [6, 30]. An ICC of $\geq 0.7$ is required for good reliability [30].

The interpretability of an instrument relates to how easily qualitative interpretations can be assigned to an instrument’s results, while its responsiveness is the ability to detect changes in the latent traits under study, usually over time. The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) checklist [4] recommends that any edge effects (ceiling or floor effects) and an MCID be reported as part of interpretability.

Here we report the responsiveness as the smallest detectable change (SDC), which is the smallest change in scores that can be shown to be greater than measurement error. The SDC can be calculated both for individuals and for groups as $SDC_{\text{ind}} = \text{SEM} \times 1.96 \times \sqrt{2}$ or $SDC_{\text{group}} = 1.96 \times \sqrt{2} \times \text{SEM}/n$, where $n$ is the group size [6]. An MCID was calculated by comparing changes in DRI between 6 and 12 months for three groups of patients defined by the HHS: stable (change in HHS $< 7$ points), improved (HHS increase $\geq 7$ points) and worsened (HHS decrease $\geq 7$ points). To be considered appropriate, the MCID must be greater than the SDC.

Results

Descriptive statistics for each of the four instruments completed at each time point in the WAT study are shown in Table 1. Item level responses of the DRI can be found in supplementary Table S1, available at Rheumatology Online.

Structural validity and internal consistency

DRI structure

After examination of the scree plot (Fig. 1), it was clear that two factors were sufficient to describe the data, as these explained 66% of the total variation within the data. The first component (PC1) accounted for the majority (41%) of the total variation, with the second component (PC2) explaining the remainder. The DRI items were well explained by the model, as all item communalities (item correlation with other items of that PC) were $> 0.5$ (Table 2). Hence the two PC model was a good approximation to the DRI data.

Although the DRI was broadly split into three sections by the original authors, the two PCs imply the possible existence of two distinct subscales. Items and their corresponding PCs are shown in Table 2, where it can be seen that sections 1 and 2 correspond to PC1 and section 3 corresponds to PC2. Item 8 (running) and item 9 (light
work) are found in a different PC. This grouping is also
reflected in the overall item means (Table 2), with the
items in PC1 scoring 30 points and items in PC2 scoring
much higher.
Cronbach’s α for the full scale was 0.92 (95% CI 0.89,
0.94), indicating high reliability. The reliability remained
high if any item was removed (see Table 2 for details),
indicating high interitem correlation. This is also reflected
in the high correlations of individual items and the overall
DRI score, the lowest of which, item 8, was 0.59 (95% CI
0.46, 0.69). Item S.D. values were all similar, ranging from
23 to 38 points (Table 2).
Subscales
As each item loaded onto a single PC, PC1 and PC2
were both considered subscales of the DRI. As shown in
Table 2, PC1 (subscale S1) consisted of all items grouped
as basic activities of daily life and three of the four items
grouped as daily physical activities. These items are all
relatively simple tasks. PC2 (subscale S2), however, com-
prises more complex and difficult tasks: all of the work
related/vigorous activities and running from the daily
physical activities group. This lends itself to headings of
the subscales of simple tasks for S1 and difficult tasks
for S2.
Fig. 2 shows boxplots of the two subscales at each time
point, showing that mean scores were higher in S2 than in
S1. Both subscales reached the lowest estimation of dis-
ability (0) at the postoperative time points, with only S1
reaching 0 preoperatively. S2 (difficult tasks) reached the
maximum disability (100) at all data collection points, but
the maximum for S1 was 97.7, preoperatively. Subscale
S1 had a Cronbach’s α of 0.92 (95% CI 0.89, 0.94), with
S2 exhibiting a lower internal consistency of 0.84 (95% CI
0.77, 0.89). These values denote good consistency.
Criterion validity
Correlations between DRI and hip function (HHS and
OHS) and quality of life (EQ-5D) are shown in Table 3.
The correlations are all negative, as higher values of the
HHS, OHS and EQ-5D indicate better function or quality of
life, but higher disability for the DRI. The smallest correl-
ation with the HHS was −0.67, however, all CIs include
good correlation between the two scales at all times. For

### Table 1: Descriptive statistics of instrument use at each time point

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Time point</th>
<th>n</th>
<th>Mean</th>
<th>s.d.</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean difference</th>
<th>Significance</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRI</td>
<td>Baseline</td>
<td>126</td>
<td>57.5</td>
<td>17.3</td>
<td>17.35</td>
<td>98.5</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>3 months</td>
<td>118</td>
<td>34.7</td>
<td>19.31</td>
<td>0.0</td>
<td>89.3</td>
<td>22.6</td>
<td>P &lt; 0.01</td>
<td>−1.30</td>
</tr>
<tr>
<td></td>
<td>6 months</td>
<td>120</td>
<td>31.6</td>
<td>22.05</td>
<td>0.0</td>
<td>85.0</td>
<td>26.1</td>
<td>P &lt; 0.01</td>
<td>−1.50</td>
</tr>
<tr>
<td></td>
<td>12 months</td>
<td>119</td>
<td>31.4</td>
<td>24.19</td>
<td>0.0</td>
<td>97.1</td>
<td>26.6</td>
<td>P &lt; 0.01</td>
<td>−1.50</td>
</tr>
<tr>
<td>HHS</td>
<td>Baseline</td>
<td>126</td>
<td>49.4</td>
<td>13.8</td>
<td>22.3</td>
<td>78.4</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>3 months</td>
<td>119</td>
<td>79.8</td>
<td>17.79</td>
<td>28.0</td>
<td>100.0</td>
<td>30.4</td>
<td>P &lt; 0.01</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td>6 months</td>
<td>120</td>
<td>85.8</td>
<td>15.09</td>
<td>30.5</td>
<td>100.0</td>
<td>36.3</td>
<td>P &lt; 0.01</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td>12 months</td>
<td>120</td>
<td>85.2</td>
<td>18.27</td>
<td>27.0</td>
<td>100.0</td>
<td>35.8</td>
<td>P &lt; 0.01</td>
<td>2.59</td>
</tr>
<tr>
<td>OHS</td>
<td>Baseline</td>
<td>126</td>
<td>19.3</td>
<td>7.82</td>
<td>5</td>
<td>41</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>3 months</td>
<td>119</td>
<td>35.9</td>
<td>9.59</td>
<td>7</td>
<td>48</td>
<td>16.6</td>
<td>P &lt; 0.01</td>
<td>2.13</td>
</tr>
<tr>
<td></td>
<td>6 months</td>
<td>122</td>
<td>39.4</td>
<td>8.74</td>
<td>12</td>
<td>48</td>
<td>20.2</td>
<td>P &lt; 0.01</td>
<td>2.58</td>
</tr>
<tr>
<td></td>
<td>12 months</td>
<td>120</td>
<td>39.2</td>
<td>10.39</td>
<td>6</td>
<td>48</td>
<td>20.0</td>
<td>P &lt; 0.01</td>
<td>2.56</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Baseline</td>
<td>126</td>
<td>0.3396</td>
<td>0.3357</td>
<td>−0.35</td>
<td>0.76</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>3 months</td>
<td>119</td>
<td>0.7092</td>
<td>0.2580</td>
<td>−0.67</td>
<td>1.00</td>
<td>0.368</td>
<td>P &lt; 0.01</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>6 months</td>
<td>122</td>
<td>0.7715</td>
<td>0.2668</td>
<td>−0.24</td>
<td>1.00</td>
<td>0.438</td>
<td>P &lt; 0.01</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>12 months</td>
<td>120</td>
<td>0.7553</td>
<td>0.3089</td>
<td>−0.18</td>
<td>1.00</td>
<td>0.420</td>
<td>P &lt; 0.01</td>
<td>1.25</td>
</tr>
</tbody>
</table>

*DRI and HHS values have been rounded to 1 decimal point; no rounding has been applied to EQ-5D and OHS values.
*Significance is of mean change from baseline calculated by a paired t-test. The effect size is calculated as the mean change divided by the baseline standard deviation. DRI: Disability Rating Index; HHS: Harris Hip Score; OHS: Oxford Hip Score.
the OHS, the smallest correlation was \( r = 0.77 \), indicating good consistency between the two scales. For the DRI and quality of life, the association is fair to good, as the correlation has larger changes between time points and wider CIs. The smallest correlation (\( r = 0.59 \)) was at baseline, before hip surgery.

**TABLE 2** Descriptive statistics for the principal component analysis sample (\( n = 126 \))

<table>
<thead>
<tr>
<th>DRI item</th>
<th>PC1 loadings</th>
<th>PC2 loadings</th>
<th>Item communality</th>
<th>Correlation with total score (95% CI)</th>
<th>Mean (S.D.)</th>
<th>Cronbach's ( \alpha ) if item is removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common basic activities of daily life</td>
<td>Q1: Dressing</td>
<td>0.73</td>
<td>0.24</td>
<td>0.59</td>
<td>0.71 (0.61, 0.78)</td>
<td>16 (23)</td>
</tr>
<tr>
<td></td>
<td>Q2: Outdoor walks</td>
<td>0.75</td>
<td>0.38</td>
<td>0.71</td>
<td>0.82 (0.75, 0.87)</td>
<td>30 (31)</td>
</tr>
<tr>
<td></td>
<td>Q3: Climbing stairs</td>
<td>0.82</td>
<td>0.24</td>
<td>0.73</td>
<td>0.78 (0.70, 0.84)</td>
<td>26 (31)</td>
</tr>
<tr>
<td></td>
<td>Q4: Sitting longer time</td>
<td>0.58</td>
<td>0.43</td>
<td>0.52</td>
<td>0.72 (0.62, 0.79)</td>
<td>29 (30)</td>
</tr>
<tr>
<td>More demanding daily physical activities</td>
<td>Q5: Standing bent over a sink</td>
<td>0.77</td>
<td>0.23</td>
<td>0.65</td>
<td>0.75 (0.66, 0.81)</td>
<td>28 (30)</td>
</tr>
<tr>
<td></td>
<td>Q6: Carrying a bag</td>
<td>0.80</td>
<td>0.24</td>
<td>0.69</td>
<td>0.76 (0.68, 0.83)</td>
<td>26 (28)</td>
</tr>
<tr>
<td></td>
<td>Q7: Making a bed</td>
<td>0.76</td>
<td>0.19</td>
<td>0.62</td>
<td>0.72 (0.63, 0.80)</td>
<td>25 (32)</td>
</tr>
<tr>
<td></td>
<td>Q8: Running</td>
<td>0.09</td>
<td>0.82</td>
<td>0.68</td>
<td>0.59 (0.46, 0.69)</td>
<td>76 (33)</td>
</tr>
<tr>
<td>Work-related or more vigorous activities</td>
<td>Q9: Light work</td>
<td>0.79</td>
<td>0.21</td>
<td>0.67</td>
<td>0.74 (0.65, 0.81)</td>
<td>20 (25)</td>
</tr>
<tr>
<td></td>
<td>Q10: Heavy work</td>
<td>0.43</td>
<td>0.78</td>
<td>0.80</td>
<td>0.83 (0.76, 0.87)</td>
<td>56 (37)</td>
</tr>
<tr>
<td></td>
<td>Q11: Lifting heavy objects</td>
<td>0.37</td>
<td>0.76</td>
<td>0.71</td>
<td>0.76 (0.68, 0.83)</td>
<td>58 (38)</td>
</tr>
<tr>
<td></td>
<td>Q12: Participating in exercise/sports</td>
<td>0.21</td>
<td>0.73</td>
<td>0.57</td>
<td>0.63 (0.51, 0.72)</td>
<td>61 (37)</td>
</tr>
</tbody>
</table>

Loadings with an absolute value > 0.45 are bold. Cronbach’s \( \alpha \) for the full scale is 0.92. PC: principal component.

**Fig. 2** Boxplots of each DRI subscale at each data collection point

Light grey boxes denote subscale 1 (S1), simple tasks; dark grey boxes denote subscale 2 (S2), difficult tasks. DRI: Disability Rating Index.

the OHS, the smallest correlation was \( -0.77 \), indicating good consistency between the two scales. For the DRI and quality of life, the association is fair to good, as the correlation has larger changes between time points and wider CIs. The smallest correlation (\( -0.59 \)) was at baseline, before hip surgery.

**Reliability, responsiveness and interpretability**

As shown in Table 1, disability measured using the DRI decreases significantly after the cohort undergoes surgery. Compared with baseline preoperative measurements, at the 12-month follow-up point, the majority of patients (89.9%) had functional gains, as they show a reduction in DRI scores. There were no instances of ceiling effects in the dataset, as the highest DRI score was 98.5 at baseline (Table 3). There were nine instances (<2% of all responses) of the minimum DRI recorded, which is far fewer than the 15% of respondents at the lowest score, which McHorney and Tarlov [31] suggest as indicative of a floor effect.

The ICC was calculated as 0.86, suggesting good reliability, since the majority of the variation in stable respondents occurs between, rather than within, subjects. The SEM of the overall DRI score was calculated to be 7.9; SDC\(_{\text{ind}}\) was 21.9 and SDC\(_{\text{group}}\) was 2.7. Hence, when using the DRI for clinical assessment of an individual, changes of <22 points cannot be distinguished from error, but when using the DRI in groups, much smaller differences are detectable.

The change in DRI between the 6- and 12-month follow-up points by stability group is shown in Fig. 3: 66 of the 118 valid respondents (56%) were considered stable, with a mean change in DRI of 0.6 points (S.D. 11.2); 26 (22%) were considered to have improved clinically, with a mean DRI decrease of 11.0 (S.D. 16.9) and the remaining 26 (22%) had worsened according to their HHS, with a mean change in DRI of 6.9 (S.D. 15.1). This suggests that the smallest difference between clinically stable and unstable patients (i.e. the MCID) in this population is ~6.9 points.

**Discussion**

The DRI is now becoming increasingly important for outcome assessment across a wide spectrum of clinical studies, particularly in musculoskeletal medicine, orthopaedics, surgery and rehabilitation research. The main reason for this is that it provides global, rather than a localized or joint-/function-specific, outcome assessment. For instance, two large multicentre studies recently funded by the UK National Institute for Health Research (NIHR), Wound Management of Open Lower Limb Fractures (WOLLF) [32] and Fixation of Distal Tibia Fractures (FixDT) [33], are both using DRI as the primary outcome in the absence of a more appropriate function-specific measure. It is important that we understand the measurement properties of DRI and, in particular, how it is...
TABLE 3 Edge effects and criterion validity

<table>
<thead>
<tr>
<th>Time point</th>
<th>n</th>
<th>HHS correlation (95% CI)</th>
<th>OHS correlation (95% CI)</th>
<th>EQ-5D correlation (95% CI)</th>
<th>DRI edge effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>126</td>
<td>-0.71, [-0.79, -0.61]</td>
<td>-0.81, [-0.73, -0.86]</td>
<td>-0.64, [-0.73, -0.52]</td>
<td>11.1–98.5, 0 (0)</td>
</tr>
<tr>
<td>3 months</td>
<td>118</td>
<td>-0.69, [-0.77, -0.57]</td>
<td>-0.79, [-0.72, -0.85]</td>
<td>-0.59, [-0.70, -0.46]</td>
<td>0–89.3, 3 (2.5)</td>
</tr>
<tr>
<td>6 months</td>
<td>120</td>
<td>-0.67, [-0.76, -0.67]</td>
<td>-0.77, [-0.69, -0.84]</td>
<td>-0.73, [-0.80, -0.63]</td>
<td>0–85.0, 5 (4.2)</td>
</tr>
<tr>
<td>12 months</td>
<td>119</td>
<td>-0.71, [-0.79, -0.61]</td>
<td>-0.80, [-0.73, -0.86]</td>
<td>-0.76, [-0.83, -0.67]</td>
<td>0–97.1, 1 (0.8)</td>
</tr>
</tbody>
</table>

Pearson’s correlations are shown between quality of life and hip function outcomes and the DRI at each data collection point. DRI: Disability Rating Index; HHS: Harris Hip Score; OHS: Oxford Hip Score.

FIG. 3 Boxplots of DRI change between the 6 and 12 month follow-up point

Sixty-six of the 118 valid respondents (56%) were clinically stable, 26 (22%) had improved and 26 (22%) had worsened. DRI: Disability Rating Index.

structured and relates to other more specific measures. The WAT data provide a particularly rich resource to investigate these issues and, to our knowledge, the work presented here is the first structural validation of the DRI. We provide evidence to suggest that the DRI has two separate and distinct subscales, labelled according to the items they contain as simple tasks and difficult tasks.

In the original study by Salén et al. [8], the authors subdivided the instrument into three sections. However, in our population of patients undergoing hip surgery for arthritis, we did not find that these categories aligned with the observed subscales. The original categories of daily physical activities and work related or more vigorous activities each had one item split into the other subscale. The problematic items were light work and running. Further work is required to confirm that this structure is indeed a more general characteristic of the DRI and not an artefact of this particular dataset or unique to patients suffering from dysfunction of the hip joint.

We have shown that the DRI can detect changes in function postoperatively, with no evidence of edge effects. By comparing the average changes in groups of study participants who experienced clinical changes in function with those who remained clinically stable suggests that the DRI MCID for this population could be as small as 6.9 points. This potential MCID is larger than the SDC_{group}, but not the SDC_{ind}, indicating that in our study population, the DRI was able to assess group changes but was less sensitive to change on an individual basis. However, this problem of detecting clinically relevant changes for an individual by a patient-reported outcome measure is not unique and has been reported elsewhere [34, 35]. Further work is needed to establish a definitive DRI MCID, potentially using multiple methods [7] or anchor points identified as relevant or important to patients [36].

Our analyses showed that there was good association between the DRI and the functional assessment scores, with the OHS correlation greater than the HHS correlation at all times. This may be due to the administration method: the OHS, like the DRI, is self-completed, while the HHS is administered by a clinician. The correlation between the DRI and the health-related quality of life (EQ-5D) score was not large enough to be considered good agreement in the first two data collection points, but increased in later follow-up periods. This implies that while quality of life and disability may be linked, they are not as closely associated as disability and function in the population under investigation. Furthermore, the EQ-5D is a utility-based measurement that aims to capture the societal views of health-related quality of life.

The WAT study was not designed to validate the DRI, making some aspects of the validation difficult to determine. In particular, MCIDs for each of the subscales were not calculated and the overall MCID given may be an overestimate, as only coarse groupings of improved and worsened were used, when ideally only patients showing minimal change would be included. Further validation is required using more than three anchor points for other patient groups, as well as confirming the MCID. Although the DRI shows potential for use in many different groups of patients, this study was clearly limited to a specific population and, as such, caution should be exercised in generalizing conclusions and recommendations more widely.
We anticipate that the DRI will correlate well with other functional scores where dysfunction is hypothesized to be a cause of disability (e.g. knee OA or ankle fracture). Also, the DRI will provide a realistic option for assessing lower limb function for other conditions where no function-specific patient-reported outcome measure exists. The DRI may also be of use in studies where the outcome of interest is not restricted to the function of a single joint. For instance, if an intervention aimed to improve mobility in OA patients, the DRI could capture change in patients with hip problems, knee problems or a combination of both; an instrument designed to capture single-joint function, although specific, may overlook wider mobility and functional improvement.

In summary, the metric properties of the DRI explored here lead us to conclude that this outcome measure is robust, reliable and sensitive and, as such, we endorse its use across a wide range of clinical and research settings, including randomized controlled trials, routine patient follow-up, audits and service evaluations.

Rheumatology key messages

- The Disability Rating Index is a valid, reliable patient-reported outcome for patients undergoing hip replacement.
- The Disability Rating Index may have wider applications for assessing lower limb function in research studies.

Acknowledgements

This article used data collected during the WAT, which was funded by the Research for Patient Benefit scheme of the National Institute of Health Research (grant reference no. PB-PG-0706-10080). The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute of Health Research or the Department of Health. M.L.C., J.A., J.B. and N.R.P. conceived of the study. H.P. planned and conducted the analysis. All authors contributed to the interpretation of results and read and approved the final manuscript.

Disclosure statement: The authors have declared no conflicts of interest.

Supplementary data

Supplementary data are available at Rheumatology Online.

References


28 Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14: 1523–32.


