Functional capacity evaluation in patients with chronic low back pain
Reneman, Michiel

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2004

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
10

Testing lifting capacity: validity of determining effort level by means of observation


Abstract

Study design: Video observation study.
Objectives: To establish the validity of determining effort level by visual observation of a lifting test.
Summary of background data: Determining effort level during a lifting test is critical for interpretation of test performance, yet the validity of these determinations has not been established in patients with chronic nonspecific low back pain (CLBP).
Methods: 15 Healthy subjects and 16 CLBP patients performed a standardized lifting test as outlined in the Isernhagen Work System Functional Capacity Evaluation (IWS FCE). The lifts were video taped and independently observed by 9 trained observers, who rated effort levels using a IWS categorical scale and a Borg CR-10 scale. External effort indices were established to control for effort at group level. Validity of the observer ratings was analyzed by means of sensitivity and specificity analysis and correlations between performances and observer ratings. Inter-rater reliability was analyzed by means of Intra-Class Coefficients (ICC) and Cohen’s Kappa.
Results: External indices differ significantly between patients and healthy subjects, indicating that at group level patients did not perform maximally. Submaximal performances were correctly rated in 85 to 90% (healthy subjects) and in 100% (CLBP) of the cases. ‘Maximal performances’ were correctly rated in 46 to 53% (healthy subjects) and in 5 to 7 % (CLBP) of the cases. Correlations between performances and observer ratings were $r=0.90$ to $r=0.92$ (healthy subjects) and $r=0.82$ (CLBP). Reliability: ICC $r=0.76$ (CLBP) to $r=0.87$ (healthy), Kappa K=0.50 (CLBP) to $r=0.58$ (healthy).
Conclusions: Effort level can be determined validly by means of visual observation.
Introduction

Chronic nonspecific low back pain (CLBP) often leads to a decrease in functional abilities. A decrease in functional abilities is related to work absenteeism, decreased productivity and wage compensation costs. Functional capacity evaluations (FCEs) aim to measure (decreased) functional abilities. FCEs are used in occupational rehabilitation programs for baseline, progress and discharge measurements, for return to work determinations, and to determine aspects of disability for insurance purposes.

To establish a person’s functional capacity, the evaluatee must perform to his or her maximum level of physical ability. The maximum performance that can be measured is the portion of capacity the evaluatee is willing to produce. The performance of the individual then depends on his or her ability to perform, and on his or her motivation to perform. It is, therefore, important to assess the extent to which a person is willing to perform to his or her physical maximum. This type of assessment requires a validated means to identify level of effort.

Methods developed to identify levels of effort (often referred to as sincerity of effort) are: Waddell’s non-organic signs, descriptions of pain behavior and symptom magnification, coefficients of variation, correlations between musculoskeletal evaluation and function, grip measurements and the relations between heart rate and pain intensity. Despite the widespread use of these methods, no evidence has been published that address their reliability and validity specific to the FCE setting. Another method of determining effort level is a standardized visual observation of indices of effort. Hazard et al. asserted that ‘a trained observer is better able to distinguish maximal from submaximal efforts than the most accurate physiological index’. The authors, however, did not substantiate their assertion with research findings.

Visual observations are an integral aspect of many FCEs, including the Isenhagen Work Systems (IWS) FCE. The inter- and intra-observer reliability of these observations has been studied on healthy subjects and on patients with CLBP in several occasions, all with good results. Only one paper is known in which the validity of visual observations is reported. The overall accuracy in identifying participants’ level of effort was 87%. Generalization of these results is limited because subjects with previous musculoskeletal pathology participated, which may not be representative for patients with current musculoskeletal injuries. Additionally, performances were either at maximal or at 50% of perceived maximal, suggesting a difference that is more outspoken than often presented in clinical practice. In the present study both healthy subjects and patients performed a lifting test in a wide range of effort levels (from light to maximal).

The primary aim of this study was to investigate the validity of determining effort levels during a lifting test by means of visual observations in a sample of...
healthy subjects and patients with CLBP. Secondary aims were to study the inter-rater reliability of observer ratings, and to study whether one rating method was superior to another.

**Methods**

**Participants**
15 Healthy volunteers and 16 patients with CLBP were included in this study. The healthy subjects consisted of a convenience sample of 7 males and 8 females. Their mean age was 29.5 years (range 18 to 53 years, sd 10.8 years), mean length was 180 cm (sd 9 cm). The patient group consisted of 12 males and 4 females with CLBP. They were selected for rehabilitation treatment by a physiatrist and referred to an FCE as part of a regular assessment protocol. Their mean age was 39.6 years (range 27 to 50 years, sd 7.1), mean length was 179 cm (sd 6 cm). Excluded were patients with co-morbidity (additional diagnoses unrelated to low back pain) and specific diagnoses related to low back pain (for example disk herniations). The duration of low back pain was less than 6 months in 2 patients, between 6 and 12 months in 3 patients, and more than 12 months in 9 patients. The patients’ mean self-rated disability on a Roland Morris Disability Scale\(^\text{12}\) was 11.1 points (sd 4.7) on a scale ranging from 0 to 24 points. Informed written consent was obtained from all participants. None of the participants used medication that influences heart rates. The ethics commission of the University Hospital Groningen approved video taping of the patients’ performance.

**Materials**
Standardized materials of the IWS FCE were used: a plastic receptacle (dimensions 30 × 40 × 26 cm) with handles on each side and metal weights of 2, 4 and 10 kg. Heart rate was measured by means of a heart rate monitor. Commercially available video equipment was used. The Borg Category-Ratio scale\(^\text{13}\) (CR-10 scale, Appendix I) was visualized on paper. The CR-10 scale was available for the patients, the healthy subjects and the observers. Available for the observers only was the IWS categorical effort level scale\(^\text{14}\), in which categories light, moderate, heavy and maximal were operationally defined (Appendix II).

**Procedures**
After a general introduction of the procedures and signing of informed consent, the healthy subjects were instructed on how to perform the lift. All subjects were explained that they were expected to perform maximally. They were also informed they could terminate testing whenever they felt unsafe or unwilling to proceed. No heart rate ceiling was set on the monitor. The evaluator performed the lift once for further explanation. The subjects were allowed to use lifting techniques of their own preference; ‘safe’ lifting techniques were not emphasized. The meaning and utility of the CR-10 scale was explained. The healthy subjects were instructed to
rate their effort level in either of 4 categories: light, moderate, heavy or maximal. The healthy subjects were asked to rate their effort after each set of lifts, using both ratings (categorical and CR-10).

Patients were introduced to the general procedures of the IWS FCE\textsuperscript{14}. All instructions and procedures were equal to the healthy subjects. Differences included: due to general safety procedures, test termination criterion of 85\% of a patient’s age related maximum heart rate was used ((220-age) \times 85\%) and the patients were asked to self-rate their effort level by means of the CR-10 scale only.

The lifting protocol was as follows (Figure 1): lifting the receptacle (weights not visible) from a 76 cm shelf, turning 90°, lowering the receptacle until it touched the floor, lifting towards an upright position, turning back 90°, and returning the receptacle to its original position. This was repeated 5 times with the same weight (one set of 5 repetitions). Before every set the heart rate was recorded. After each set heart rate, duration of the test and self-rating of effort was recorded. The weight in the receptacle was increased in 4 to 7 steps until a maximum was reached.

The video camera was mounted on a tripod. A lateral view was obtained when the subject initially grasped the receptacle and a frontal view was obtained when the subject lowered and lifted. All sets of lifts were recorded. The lifts of the patients and healthy subjects were edited in random order. The sequence of all the lifts of one person remained intact on the tape. Observers were instructed to the general procedures, and to the use of both rating methods (Categorical and CR-10). The observers were instructed to rate effort levels only, and to disregard whether the lifts were performed using ‘proper’ body mechanics. The observers did not know whether the person stopped or lifted a higher amount of weight subsequently after a certain weight. They were not informed on whether the person was a healthy subject or a patient. No information was available about heart rates or self-ratings of effort. There was no sound on the tape. Observers were blinded for each other’s ratings. Rewinding of the tape was not allowed. In total, 134 sets of lifts were edited on the video, consisting of 71 sets of healthy subjects and 63 sets of CLBP patients. The length of the tape was 90 minutes. Technical problems prevented videotaping of maximal performance of 1 patient. Additionally, 5 observers were unable to rate the last 4 lifts (numbers 131 to 134) due to technical problems.

Observers
Nine functional capacity evaluators observed the videotape. Four observers were occupational therapists and 5 were physical therapists. Four observers were employees of 3 different facilities in the Netherlands, and 5 observers were employees of IWS in the USA. All observers were trained in the IWS observational criteria. None of the observers was experienced in the use of the CR-10 scale to express effort level.
Data analysis

For each set of lifts completed, the following variables were entered in a database: effort levels rated by the healthy subjects, patients and observers (categorical and CR-10), the lifted weight (kg), heart rates (beats per minute), target maximal heart rate (220-age), percentage of target maximal heart rate, the duration of each set (seconds), and actual performance. Actual performance (in percentage) was calculated as percentage of maximal performance (=100%). Heart rate increase adjusted for age was entered as a separate variable. For example: a 40 year old male had a heart rate prior to the lifts of 100 bpm and after the lifts of 125 bpm. His maximum heart rate was 180 bpm (220-40=180). His percentage of maximum heart rate was 69.4% ((125/180) × 100%). The heart rate increase of 25 bpm was 25% ((125-100)/100) × 100%). His heart rate increase adjusted for age was 13.8% ((25/180) × 100%).

The performances of healthy subjects and CLBP patients were compared to analyze level of effort. Valid external indices of effort at group level were used for this comparison (duration, heart rate and self-rating of performance intensity). It was expected that the performance indices of the patients would not increase as much as the healthy subjects. The differences in performance indices were analyzed using a Mann-Whitney U-test (α<0.05).

Whether the observers were able to correctly identify individual effort levels by means of visual observation was determined. To determine whether the observer ratings were reflective of differences in performance, maximum performances of patients and healthy subjects were set at 100% (regardless whether or not this was a ‘true maximum’). Sets of lifts performed prior to maximum were coded as a percentage of this performance. For example if maximum performance was 50 kg, then a performance of 10 kg was coded as 20%, a performance of 45 kg as 90%, etc. Mean observer ratings of maximal performance was calculated. The specificity of the ratings (correct rating of submaximal performances) was calculated as a percentage of all lifts preceding maximal performance. The sensitivity of the ratings (correct rating of maximal performances) was calculated as a percentage of all maximal performances (the last set of weight lifted). For these analyses, CR-10 ratings exceeding 10 were coded as ‘maximal’.

Relationships between observer rated effort levels and actual performances were expressed in Pearson and Spearman correlation coefficients.

Differences between the rating methods were computed with a Wilcoxon test (significant when p≤0.05). Inter-rater reliability of the ratings of the 9 observers was analyzed using Intra Class Correlation method one-way random (CR-10 scale) and a mean weighted Kappa coefficient (categorical data) as indices for reliability. Interpretation of ICCs: r<0.49: weak relationship, r≥0.50 moderate relationship, and r≥0.75 strong relationship. Interpretation of Kappa coefficients: c.41 to c.60 moderate agreement, c.61 to c.80 substantial agreement, c.81 and higher excellent agreement.
Kappa analyses were performed using AGREE statistical software. All other analyses were performed using SPSS 10.0.

Results

Indices of performance
As shown in Table 1, the means of indices for effort were significantly different between the healthy subjects and patients. The healthy subjects reached a higher percentage of target heart rate, took longer to perform the lifts and scored higher on the CR-10 scale. There was no difference between healthy subjects and patients in heart rate before the lift. There were no differences between males and females in all indices in both groups, except for maximal amount of weight lifted. Maximum performance was significantly different between healthy subjects and patients for both genders separately. The objective indices of effort indicate that the patients as a group did not reach maximal physical performance. The patients’ self-rating indicates this performance to be ‘very heavy’, as expressed in a mean CR-10 rating of 7.7. The objective indices of effort indicate that the healthy subjects as a group did reach or approach maximal performance. The healthy subjects’ self-rating indicates this performance to be ‘maximal’, as expressed in a mean CR-10 rating of 13.0. Two healthy subjects self-rated very high with scores of 20 and 25 on the CR-10 scale. Removal of these 2 outliers resulted in a mean self-rating of 11.5 (SD 1.0).

Table 1 Indices of effort at maximum performance

<table>
<thead>
<tr>
<th>Indices</th>
<th>Healthy subjects</th>
<th></th>
<th>CLBP</th>
<th></th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>mean</td>
<td>sd</td>
<td>n</td>
<td>mean</td>
</tr>
<tr>
<td>Maximal performance (kg) – males</td>
<td>7</td>
<td>73.4</td>
<td>15.7</td>
<td>12</td>
<td>36.3</td>
</tr>
<tr>
<td>Maximal performance (kg) – females</td>
<td>8</td>
<td>40.1</td>
<td>14.0</td>
<td>3</td>
<td>12.0</td>
</tr>
<tr>
<td>Duration of the lifts (sec)</td>
<td>15</td>
<td>38.1</td>
<td>11.7</td>
<td>16</td>
<td>25.9</td>
</tr>
<tr>
<td>Self-rated performance (CR-10 scale)</td>
<td>15</td>
<td>13.0</td>
<td>4.1</td>
<td>11</td>
<td>7.7</td>
</tr>
<tr>
<td>Heart rate before (bpm)</td>
<td>15</td>
<td>109.9</td>
<td>13.7</td>
<td>15</td>
<td>110.7</td>
</tr>
<tr>
<td>Heart rate after (bpm)</td>
<td>15</td>
<td>156.7</td>
<td>13.7</td>
<td>15</td>
<td>131.5</td>
</tr>
<tr>
<td>Heart rate increase (%)</td>
<td>15</td>
<td>43.9</td>
<td>14.5</td>
<td>15</td>
<td>19.2</td>
</tr>
<tr>
<td>Heart rate increase adjusted for age (%)</td>
<td>15</td>
<td>24.6</td>
<td>14.5</td>
<td>15</td>
<td>11.4</td>
</tr>
<tr>
<td>Percentage maximal heart rate (%)</td>
<td>15</td>
<td>82.5</td>
<td>7.9</td>
<td>15</td>
<td>73.0</td>
</tr>
</tbody>
</table>

CR-10: Category-Ratio scale
bpm: beats per minute
*: Indicates p<0.001
Observer ratings of performance
The mean observer rating of the patients’ maximal performance was 6.6 (sd 1.6, 95% CI 5.6 to 7.5), and of the healthy subjects’ maximal performance was 9.5 (sd 1.5, 95% CI 8.6 to 10.5). The results of analyses of the specificity and sensitivity of differentiating between maximal and submaximal effort are presented in Table 2 (healthy subjects) and Table 3 (CLBP patients). Specificity is presented in the Tables for 2 different scenarios: the first in which only the lifts just prior to maximal performance were included (referred to as ‘selected performances’ in Tables 2 and 3), the second in which all sets of lifts preceding maximal performance were included (‘all performances’). For the healthy subjects, the lifts ‘unjustly’ rated as not maximal were rated as heavy in 98.8% of the performances and as moderate in 1.2% of the performances. The mean CR-10 rating for the same performances was 8.8 (sd 1.3, 95% CI 7.6 to 9.8). For the patients, the lifts ‘unjustly’ rated as not maximal were rated as heavy in 74.1% of the performances, as moderate in 25.1% of the performances, and as light in the remaining 0.8% of the performances. The mean CR-10 rating for the same performances was 6.6 (sd 1.6, 95% CI 5.6 to 7.5). The correlations between observer rated effort levels and actual effort levels were high (Table 4). The correlation coefficients concerning the ratings of the healthy subjects were significantly higher than the correlations of the patients (p<0.05).

Table 2  Sensitivity and specificity of the ratings for healthy subjects (n=15)

<table>
<thead>
<tr>
<th>Observer rating</th>
<th>Selected submaximal performances</th>
<th>All submaximal performances</th>
<th>Maximal performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cat</td>
<td>CR-10</td>
<td>Cat</td>
</tr>
<tr>
<td>Submaximal (%)</td>
<td>*85.1</td>
<td>*89.6</td>
<td>*95.8</td>
</tr>
<tr>
<td>Maximal (%)</td>
<td>14.9</td>
<td>10.4</td>
<td>4.2</td>
</tr>
<tr>
<td>Total ratings (N)</td>
<td>134</td>
<td>135</td>
<td>502</td>
</tr>
</tbody>
</table>

Selected performances: includes only the lift just prior to maximal performance
All performances: includes all lifts prior to maximal performance
Cat: category
CR-10: CR-10 scale
*: Specificity of the rating
**: Sensitivity of the rating
Table 3  Sensitivity and specificity of the ratings for CLBP patients (n=16)

<table>
<thead>
<tr>
<th>Observer rating</th>
<th>Selected submaximal performances</th>
<th>All submaximal performances</th>
<th>Maximal performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cat CR-10</td>
<td>Cat CR-10</td>
<td>Cat CR-10</td>
</tr>
<tr>
<td>Submaximal (%)</td>
<td>*100 *100</td>
<td>*100 *100</td>
<td>93.1 94.7</td>
</tr>
<tr>
<td>Maximal (%)</td>
<td>0 0</td>
<td>0 0</td>
<td>**6.9 **5.3</td>
</tr>
<tr>
<td>Total ratings (N)</td>
<td>122 122</td>
<td>410 411</td>
<td>130 131</td>
</tr>
</tbody>
</table>

Selected performances: includes only the lift just prior to maximal performance
All performances: includes all lifts prior to maximal performance
Cat: category.
CR-10: CR-10 scale
*: Specificity of the rating
**: Sensitivity of the rating

Table 4  Mean correlation between levels of effort and observer rated effort levels (SD)

<table>
<thead>
<tr>
<th></th>
<th>CR-10(^1)</th>
<th>Category(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy subjects</td>
<td>0.90 (0.02)</td>
<td>0.92 (0.03)</td>
</tr>
<tr>
<td>Patients</td>
<td>0.82 (0.05)</td>
<td>0.82 (0.05)</td>
</tr>
</tbody>
</table>

\(^1\) Pearson correlation
\(^2\) Spearman rank correlation

Differences between the 2 rating scales were minor or non-existing. Correlation coefficients, specificity and sensitivity were not significantly different from each other (p<0.05). The inter-rater reliability between the 9 observers are for healthy subjects: CR-10 ratings ICC=0.87 (95%CI=0.69 to 0.91) and categorical ratings Kappa=0.58. The inter-rater reliability between the 9 observers are for CLBP patients: CR-10 ratings ICC=0.76 (95%CI=0.69 to 0.83) and categorical ratings Kappa=0.50.

Discussion

Overseeing the results of this study, it appears that effort levels can be determined validly and reliably by means of observation. This conclusion, however, is derived from the observer ratings and the assumption that the indices of performance are indeed valid indices for effort during a lifting test.
Indices of performance

All objective indices of effort were significantly different between the patients and healthy subjects, indicating submaximal performance of the patients. The patients’ self-rating of effort (‘very heavy’) is in concordance with the objective indices. The healthy subjects clearly performed higher on all indices, including self-rating of effort (‘maximal’). In the absence of a gold standard, however, it cannot be concluded that the healthy subjects’ performance was maximal.

The use of heart rate increase as an indicator for effort in dynamic work is used widely and is based firmly in exercise physiology. In a general sense, the observation that an increase in workload is accompanied by an increase in heart rate is undisputed. The variance in individual responses, however, is too wide to use heart rate as a valid indicator in individual lifting tests. In this study heart rate increase is used as an indicator of effort at group level only. The patients’ heart rate increase of 19.2% and the healthy subjects’ increase of 43.9% supports the assumption that the patients did not perform maximally. Another well-known phenomenon in exercise physiology is that the velocity of a movement decreases when an increase of power is demanded. No criterion for distinction between submaximal or maximal performance has been set for duration of lifts, however, a longer duration indicates a higher level of effort. The mean duration of the maximal performance was longer for the healthy subjects compared to the patients (38.1 versus 25.9 seconds respectively), supporting the assumption that the patients did not perform maximally. Self-ratings of maximal performances of the patients (mean of 7.7) were lower than the healthy subjects (mean of 13.0). This indicates that the patients themselves rated their maximal performance as ‘very heavy’, which is short of maximal performance. Taking all indices in perspective, it is concluded the patients did not perform to physical maximum on the lifting test.

Observer rating of performance

When the patients’ performances were submaximal, the observer ratings were expected to indicate just that. Indeed, the observers rated the maximal performances of the patients as 6.6 (CR-10 scale) and of the healthy subjects as 9.5. Specificity was high for both groups, indicating that submaximal performances were correctly identified as submaximal. Sensitivity was very low in the patient group, indicating that the observers almost never rated ‘maximal performances’ as such. Retrospectively, the patients did not perform maximally, thus high sensitivity scores may not be expected. It is not known whether the healthy subjects have performed to their maximum, because of the absence of a valid criterion to serve as a gold standard. The sensitivity of the ratings of healthy subjects is much higher than the patients’, however, they do not exceed 53.4%. In the 46.6% that the observer did not rate maximal performance ‘correctly’, the performances were almost always rated as heavy (categorical scale) or rated ‘extremely heavy’ (mean of 9.5 on the CR-10 scale). With regards to clinical interpretation of the specificity/sensitivity analysis, the following can be concluded: when a performance by healthy subjects is rated by the observer as light or moderate, it is a submaximal performance; when it is rated
as heavy, it can either be a heavy or a maximal performance, and when it is rated as maximal, it is almost always a maximal performance. For patients it can be concluded that a performance rated by an observer as submaximal is almost always indeed a submaximal performance. The ability to correctly identify maximal performances of CLBP patients is unknown, because maximal performances did not occur.

Rating expressions
With regards to the validity of the ratings, it makes no difference whether the ratings are expressed in CR-10 scale or category. Differences between the ratings were minor and statistically non-significant. With regard to the inter-rater reliability of the ratings, however, the rating method does matter. Inter-rater reliability was very good when ratings were expressed in CR-10 scale. Reliability of ratings expressed in category was moderate with Kappa values below the 0.60 for both groups. Inter-rater reliability between patients and healthy subjects differed: observers rated the healthy subjects more consistently than the patients. It is hypothesized that pain behaviors, such as the presence of antalgic movement patterns, were mistakenly interpreted as effort by one observer more than the other. However, no validated instrument is currently available to assess pain behaviors specific to the FCE setting. This should be subject for further research.

Relevance
As stated in the introduction, the maximum performance that can be measured is the portion of capacity the evaluatee is willing to produce. The performance of the individual depends on his or her ability to perform, and on his or her motivation to perform. Detection of effort level is not a goal by itself, but a means to interpret the performance of the individual testee. This is relevant for clinical use and for future research. Rational decisions regarding the design of a rehabilitation program, return to work recommendations and disability determinations cannot be made without a validated interpretation of the performances to determine the ‘meaning’ of the scores. It is incorrect to assume that the evaluatee’s performance would be equal to his or her capacity. Clinical decisions should be different when performances are reflective of maximal or submaximal physical capacities. For example, occupational rehabilitation programs are often designed to resolve the difference between the patient’s performance and the anticipated workload. However, the content of the program should depend on whether the reason for this difference is a lack of capacity or the ‘unwillingness’ to perform maximally. Consequently, it is not only important to assess the extent to which a person performs to his or her physical maximum, but also to identify the reason(s) why he or she performs as such. Theoretically, the magnitude of the performance is resultant of several factors, such as physical abilities, understanding of effort required during testing, and psychological factors such as fear avoidance beliefs, pain, pain experiences, self-efficacy and outcome expectancies, as well as unconscious or conscious symptom magnification. These and other psychological factors have been postulated to prevent people from performing to their maximum ability and
Terminating an activity prior to reaching a physical maximum\textsuperscript{16,18,19}. However, evidence of the strength of these relationships is often unavailable. Knowing the results of this study, further research can now be executed to study the determinants of the patients’ performances.

**Limitations of the study**

Observing a tape without knowledge of the subject’s diagnosis, symptoms, and heart rate, and without sound and communications with the subject is obviously different from a clinical setting. It is unknown how these differences translate to everyday practice, and, thus, to judge whether the observer ratings are better or worse in a clinical setting. Real life observations are preferable to the rigid angle and focus, as well as the technical imperfections of videotape. From this, it is expected that observer ratings would improve in a clinical setting. Prior knowledge of the diagnosis and symptoms, and the availability of sounds (including pain behaviors such as moaning, sighing and verbal complaints) may lead to effort determinations that are less valid than presented in this paper. To optimize generalization of the results of this study, clinicians are advised to focus their observations on signs of effort, and to disregard non-specific pain behaviors as much as possible.

The healthy subjects were selected not only on their healthy status, but also on their willingness to perform maximally. This sample of healthy subjects may not be a random sample from the open population. Additionally, it is uncertain whether the subjects did perform to their maximum abilities. FCEs consist of many tests. Only a lifting test was studied. The observational criteria (Appendix II) are designed to apply to other activities and patient groups as well. However, validity for those tests and patient groups is not established.

**Strength of the study**

In contrast to other studies\textsuperscript{5,11}, the subjects were not asked to perform either maximal (100\%) or submaximal (50\%), but to perform through a wide range of performances (from light to maximum). Consequently, patients performing between 1\% and 99\% of their maximum would all be rated submaximal. The wide variety of effort levels that would all be labeled as submaximal in other studies was acknowledged in this study. The results of the reliability analysis confirm the need to express variety of effort with more detail. The use of a CR-10 scale is recommended over a crude 4-point categorical scale. Additionally, the performances and observer ratings of both CLBP patients and healthy subjects were studied. Generalization of the results is also enhanced, because of the number of and variety in observers (4 facilities, 2 countries, PT and OT). Detailed analyses of variances in ratings between the observers were outside the scope of this study.
Conclusions

Observer ratings for the healthy subjects correspond highly with actual performances. Submaximal performances of patients and healthy subjects are well identified by means of observation (sensitivity). Maximal performances of healthy subjects can be identified with acceptable specificity. The specificity of the ratings of patients is unknown, because the patients did not perform maximally. Correlations between performances and observer ratings were high. It is concluded that effort level can be validly determined by means of visual observation. Ratings should preferably be expressed in a CR-10 scale, instead of a categorical rating method.

Acknowledgement

The authors acknowledge the participants and the observers for donating their time and effort to this study.

References

16. Watson PJ. Non-physiological determinants of physical performance in musculoskeletal pain. Syllabus IASP refresher courses on pain management held in conjunction with the 9th world congress on pain, August 22-27, 1999, Vienna, Austria.
Appendix I
Borg CR-10 scale

<table>
<thead>
<tr>
<th>Rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No effort at all</td>
</tr>
<tr>
<td>0.5</td>
<td>Extremely light</td>
</tr>
<tr>
<td>1</td>
<td>Very light</td>
</tr>
<tr>
<td>2</td>
<td>Light</td>
</tr>
<tr>
<td>3</td>
<td>Moderate</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Heavy</td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Very heavy</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Extremely heavy</td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>..</td>
<td>Maximal</td>
</tr>
</tbody>
</table>

Appendix II
IWS FCE observational criteria for level of effort

<table>
<thead>
<tr>
<th>Light</th>
<th>moderate</th>
<th>Heavy</th>
<th>Maximal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Muscle recruitment</td>
<td>Prime movers only, no accessory muscles, no trunk and neck stabilizers</td>
<td>Recruitment of accessory muscles and trunk and neck stabilizers</td>
<td>Pronounced recruitment of accessory muscles and trunk and neck stabilizers Bulging of accessory muscles and trunk and neck stabilizers</td>
</tr>
<tr>
<td>Base of support</td>
<td>Natural stance</td>
<td>Stable base</td>
<td>Wider base</td>
</tr>
<tr>
<td>Posture</td>
<td>Upright posture</td>
<td>Beginning of counter balance</td>
<td>Increasing counter balance</td>
</tr>
<tr>
<td>Control and movement pattern</td>
<td>Easy movement patterns</td>
<td>Smooth movements</td>
<td>Begins to use momentum</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Difficult but not maximal</td>
</tr>
</tbody>
</table>