Physical performance tests and spinal pain –

Assessing impairments and activity limitations

Therese Ljungquist

From the Division of Physiotherapy,
Neurotec Department
and
Section for Personal Injury Prevention, Department of
Clinical Neuroscience,
Karolinska Institutet, Stockholm, Sweden

Stockholm 2002

Cover illustration: Markus Ljungqvist

Nog finns det mål och mening i vår färd -
men det är vägen, som är mödan värd.

ur Karin Boyes "I rörelse", 1927

Abstract

Background: Long-term spinal pain is a common health problem, often leading to disabilities. There is still no general agreement on what measures to use for evaluating disability in people with spinal pain. Performance-based tests are often used by physiotherapists for assessing impairments and activity limitations, but our knowledge of the clinimetric properties of such tests has been limited.
Aims: The overall aim of this thesis was to identify among eleven tests assembled in a test package, one or more that, based on clinimetric properties, could be used in clinical practice for a) assessing impairments and activity limitations in persons with long-term spinal pain, and b) contributing to a common basis for evaluation and treatment of persons with long-term spinal pain. The objective of Study I was to explore the underlying foundations for ratings made by physicians, physiotherapists and insurance officers involved in an individual's rehabilitation concerning rated need of rehabilitation and rated potential to benefit from rehabilitation, for persons with long-term spinal pain. The objectives of Studies II-V were to examine the reliability, validity and sensitivity to change of the physical performance tests, and to identify their clinical usefulness.
Methods: The physical performance tests examined were the Åstrand ergometry test; isometric endurance tests for neck and trunk flexion and extension; a dynamic endurance test for the lower extremity; two lifting tests (PILE tests) and three gait tests. The basis for treatment recommendations was examined with a questionnaire distributed to professionals involved in 214 persons' rehabilitation process. *Discriminative ability* of the physical performance tests was examined by comparing test performance for persons with long-term spinal pain and that of back-healthy persons. *Inter-rater agreement* was examined for persons with long-term spinal pain; *intra-rater repeatability* and *inter-rater repeatability* were examined for persons with long-term spinal pain as well as back-healthy persons. For *construct validity* we examined the possible effects of different related factors and background factors on test performance. *Sensitivity to change* was examined by relating the changes in performance to self-rated changes.
Results: The ratings of need for rehabilitation were based on duration of sick leave (physician) and on self-rated physical function (insurance officer). Most tests discriminated between persons with long-term spinal pain and back-healthy persons. Six of the tests were considered to have acceptable reliability when repeated over time and between raters. Persons with neck pain had generally better performance than those with low back pain, except in the cervical lifting test. When pain behaviour was high, the test performance went down. Rated pain and exertion levels during the tests also affected test performance. Background factors explained only at most 27 percent of the variation in performance scores. The sensitivity to change was moderate in most tests, but was greater for subjects with low performance at inclusion in the study.
Conclusions: There is an obvious need for a common basis and commonly accepted measures for evaluation and treatment in persons with long-term spinal pain. The cervical lifting test and the gait tests can without reservations be recommended for evaluating impairments and activity limitations in people with spinal pain. For use as outcome measures, the cervical lifting test, the gait test with burden and the stair-climbing test can be of interest. Physical performance tests and self-rated measures of disability complement each other, and might both be used as tools for describing disability and as outcome measures for persons with long-term spinal pain.

Keywords: *back pain, decision-making, disability evaluation, low-back pain, methods, motor skills, neck pain, needs assessment, outcome assessment, pain measurement, physical therapy, reproducibility of results.*

ISBN 91-7349-215-9

List of original papers

The thesis is based on the following publications, which will be referred to in the text by their Roman numerals (I-V):

I. Jensen I, Bodin L, Ljungquist T, Bergström G, Nygren Å. Assessing the needs of patients in pain: A matter of opinion? Spine 2000;25(21):2816-2823.

II. Ljungquist T, Fransson B, Harms-Ringdahl K, Björnham Å, Nygren Å. A physiotherapy test package for assessing back or neck dysfunction – discriminative ability for patients versus healthy control subjects. Physiotherapy Research International 1999;4(2):123-140.

III. Ljungquist T, Harms-Ringdahl K, Nygren Å, Jensen I. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. Physiotherapy Research International 1999;4(3):214-232.

IV. Ljungquist T, Nygren Å, Jensen I, Harms-Ringdahl K. Physical performance tests for people with long-term spinal pain – aspects of construct validity. Submitted.

V. Ljungquist T, Nygren Å, Jensen I, Harms-Ringdahl K. Physical performance tests for people with spinal pain – responsiveness to clinically important change. Submitted.

Some additional data are added in the Summary.

Contents

Definitions

Activity limitations – Difficulties an individual may have in executing activities (ICF 2001).

Back-healthy persons – Persons considering themselves as free from spinal pain conditions.

Basic responsiveness statistical package – Term used in this thesis for three statistical methods for examining sensitivity to change or responsiveness, consisting of a) Comparison between group differences with Wilcoxon signed ranks test, b) Correlation coefficients against other measures, and c) ROC curves (Deyo et al 1991).

Clinimetric properties - Measurement qualities of a measurement tool designed for direct clinical use. Includes reliability, validity and responsiveness.

Disability – A general term for impairments, activity limitations and participation restrictions, according to ICF 2001. Disability represents the problematic aspect of the classification.

Functioning - An umbrella term embracing all body functions, activities and participation, according to ICF 2001. Functioning represents the healthy aspect of the classification.

Impairments – Problems in body function or structure such as significant deviation or loss (ICF 2001).

Long-term spinal pain – Spinal pain persisting for at least twelve weeks (Abenheim et al 2000).

Musculoskeletal – Referring to the human muscle, joint and/or skeletal system.

Normative value – Value discriminating between persons with long-term spinal pain and back-healthy persons, derived from sensitivity and specificity cut-off values in Study II.

Outcome measure – A measure used as an indicator of change, related to a baseline measurement.

Pain behaviour - All behaviour communicating to others the fact that pain is being experienced (Fordyce 1976).

Perceived exertion – The overall perceived muscle strain, joint loading and effort on the cardio-respiratory system by a person while exercising.

Physical performance test – A test where the test person carries out a physical activity of any kind.

Repeatability coefficient – 2.77 x within-subject SD Two readings by the same method will be within 2.77 within-subject SD for 95 % of subjects (Bland and Altman 1999).

Spinal pain – Pain perceived as arising from the vertebral column or its adnexa (i.e. the structures attached to the vertebral column) (Merskey and Bogduk 1994).

Abbreviations

ASES = The Arthritis Self-efficacy Scale, a questionnaire designed for self-ratings of one's own capability of managing consequences of chronic arthritis (Lomi 1995)

CI = Confidence interval, statistical term for the interval within which the true value is likely to be found with a certain probability, usually 95 %

CR10 Scale = Borg´s 10-grade category scale with ratio properties (Borg 1982)

ICC = Intraclass correlation coefficient, a statistical method for examining differences in measurements between subjects

ICF = World Health Organisation International Classification of Functioning, Disability and Health (WHO 2001)

IQR = Inter-quartile range, the range between the first and the third quartile in a data set

kg = Kilograms

LBP = Low-back pain

m/s = Metres per second

PILE = Progressive Isoinertial Lifting Evaluation (Mayer et al 1988)

PPT = Physical performance test

PT = Physiotherapist

RCT = Randomized Controlled Trial, i.e. a study which has used an unbiased randomisation procedure and included a control group.

ROC curve = Receiver operating characteristic curve; a plot of sensitivity versus 1 minus specificity for each possible cut-off point, with the points joined by a line, forming a curve (Bland 2000)

RPE = Rated perceived exertion as rated by Borg´s 15-grade scale (Borg 1970)

TP = Test person, i.e. a person who is undergoing a test procedure.

s = Seconds

SD = Standard deviation of a measure, a measure of deviation from the mean

SF-36 = Short Form 36, a questionnaire designed for self-rated health-related quality of life (Mc Horney et al 1994)

SPSS – Statistical Package for the Social Sciences, a statistician programme used in this thesis

UAB = The University of Alabama in Birmingham Pain Behavior Scale (Richards et al 1982)

Introduction

In rehabilitation of long-term spinal pain, physiotherapists perform assessments to obtain a picture of the patient's physical capabilities and shortcomings. The consistency between assessment methods, techniques and recordings is currently low, as no "gold standard" yet exists in our field. It is therefore important to examine different kinds of measurement for possible use in physiotherapy practice, and for contributing to a future 'assessment instrument bank' for physiotherapists. In the mid-1990s, the need arose among our group of physiotherapists working in a rehabilitation company at seven locations in Sweden and one in Norway for a common test package. We wanted tests that could reveal information about the level of disability in persons with long-term spinal pain, since most of our patients had pain in the neck or in the low back. The ambition was an instrument to get a baseline assessment for optimal planning of rehabilitation. Moreover, we wanted an instrument which could be used as an outcome measure. It had to cover most of the possible disabilities we as physiotherapists wanted to evaluate in persons with spinal pain. It had to be easy to administrate, and non-expensive.

Pain

Pain is the primary reason for seeking medical help. Pain is a subjective multidimensional experience. It is influenced by many factors that interfere with the nociceptive signals to the central nervous system and the body-specific pain-modulating system. Pain has been defined as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage", by the International Association for the Study of Pain, IASP (Merskey and Bogduk 1994). Pain is part of life and plays an important protective role. Nevertheless, prolonged pain or pain that is perceived as uncontrollable affects quality of life (Turk and Melzack 1992, p.*xi*).
There are several aspects of pain experience. None can be measured without the co-operation of the person in pain. Pain site, pain intensity, and pain duration are all frequently measured aspects, both in the clinic and in research studies.

Spinal pain

The spine serves as the central core of our body. It houses and protects the spinal cord and related neural tissues. The spine also provides attachment sites for muscles and other structures, affording a stable foundation for all these elements. Spinal pain is, due to its universality, to be considered as a normal element in life. Almost everyone experiences spinal pain at some time in life (Nachemsson et al 2000, p.34).

Lumbar pain

Lumbar pain has been defined as

> "Pain perceived as arising from anywhere within a region bounded superiorly by an imaginary transverse line through the tip of the last thoracic spinous process, inferiorly by an imaginary transverse line through the tip of the first sacral spinous process, and laterally by vertical lines tangential to the lateral borders of the lumbar erectores spinae"
> (Merskey and Bogduk 1994, p.11)

Factors that influence the occurrence or severity of lumbar pain include earlier episodes of lumbar pain, work involving heavy lifting, work in awkward positions and whole-body vibrations. Job satisfaction as well as high demands and low control over one's job are psychosocial factors of importance (Nachemsson et al 2000, p.9-10). Psychological factors such as stress, distress, mood and pain behaviour play a role in first-time incidence and in recurrent lumbar pain (Linton 2000).

Thoracic pain arises from the thoracic spine, and is by far less common. Thoracic pain conditions have seldom been discussed in the literature. In this thesis, thoracic pain conditions are included in the general term 'lumbar pain'.

Cervical pain

Cervical pain has been defined as

> "Pain perceived as arising from anywhere within the region bounded superiorly by the superior nuchal line, inferiorly by an imaginary transverse line through the tip of the first thoracic spinous process, and laterally by sagittal planes tangential to the lateral borders of the neck. (Merskey and Bogduk 1994, p.11)

There has been much less research on cervical pain than on low-back pain. Factors which have been suggested to influence the occurrence and severity of cervical pain are work with repetitive and monotonous tasks, work with flexed or twisted trunk, and non-ergonomic design of the workplace (Nachemsson et al 2000, p.10, Hansson and Westerholm 2001). In the same way as for lumbar pain, job satisfaction and high demands from and low control of one's job seem important (Nachemsson et al 2000, p.10). And for cervical pain too, psychological factors play a role both in first-time incidence and in recurrent pain conditions (Linton 2000).
In this thesis, 'cervical pain' and 'neck pain' are used interchangeably.

Long-term spinal pain

Chronic spinal pain is defined as spinal pain persisting for at least twelve weeks (Abenheim et al 2000). The term 'long-term' is preferred before 'chronic' in this thesis due to the negative expectations associated with the latter word.

Approximately twenty percent of the persons sick-listed in Sweden for at least two months has different kinds of spinal pain (Hansson and Hansson 1999). In most cases, spinal pain is a state of short duration, but in about 25 % of the cases, the pain still persists after a year (Nachemsson et al 2000, p.35).

Long-term spinal pain consequences for disability

In a number of studies, evidence for present disability in persons with long-term spinal pain has emerged. Some examples are shown below.

Lumbar pain

Dehlin and co-workers reported on lower strength in the quadriceps muscle in nursing aids with LBP (Dehlin et al 1978). Troup and co-workers found that subjects with chronic LBP had lower lifting capacity than subjects without previous LBP or with mild LBP problems (Troup et al 1987). Reid and co-workers found that persons with LBP had lower strength in trunk flexion and extension compared to back-healthy controls (Reid et al 1991). Hultman and co-workers examined men with LBP and compared them with men with no history of LBP. They found that the men with chronic LBP had significantly lower strength and muscular endurance in trunk muscles (Hultman et al 1993).

Cervical pain

In 1966, Krout and Anderson published a study where 115 persons with neck pain of different duration and character were manually examined for weakness in the neck flexors. Ninety-five of the subjects were considered as having 'significant weakness' either bilaterally or unilaterally. Endurance training of the neck flexors resulted in complete reduction of, or markedly decreased, pain for the majority of the subjects (Krout and Anderson 1966).
Silverman and co-workers found that persons with long-term cervical pain had significantly lower strength in the neck flexors compared to neck-healthy controls (Silverman et al 1991).

Rehabilitation

Rehabilitation has been defined as "a goal-oriented and time-limited process aimed at enabling an impaired person to reach an optimum mental, physical and/or social functional level, thus providing her or him with the tools to change her or his own life. It can involve measures intended to compensate for a loss of function or a functional limitation (for example by technical aids) and other measures intended to facilitate social adjustment or readjustment." (United Nations 1982).
Strong evidence from RCTs show that back training, manual therapy and multidisciplinary treatment programmes are effective for pain relief and improvements in overall functioning for persons with long-term LBP (Nachemsson et al 2000, p.24 English Summary). Strong evidence concerning neck pain is lacking, but moderate evidence exists for the effectiveness of physical exercise (Nachemsson et al 2000, p.28 English Summary).

Consensus between rehabilitation actors

By analogy with the limited knowledge about the origin of spinal pain, consensus concerning how to treat spinal pain is only recently beginning to take shape (Nachemsson et 2000, Abenhaim et al 2000). The consensus among LBP researchers regarding principal care and treatment, evaluation and classification, however, has not yet spread to clinicians. Regarding neck pain, a consensus has not yet been established.

Evaluation. Lack of agreement regarding evaluation of patients has been described by Cherkin and co-workers. They studied how different physicians in a number of medical specialities examined and evaluated "typical LBP cases" as defined in a mailed folder. There were large differences between physicians (Cherkin et al 1994). Several studies (Sandström and Esbjörnsson 1986, Härkäpää 1992) have shown the importance of patients´ opinions and expectations for the outcome of rehabilitation efforts.
Opinions differ regarding the relative clinical importance of questionnaires versus "objective" tests for evaluation of disability. Deyo and co-workers suggested that questionnaires measuring health-related quality of life should be included in the outcome measures for persons with low-back pain, and that physiological and anatomical measures might even be unnecessary (Deyo et al 1994). However, self-reports of behaviour reveal how people *believe* they perform, which is not the same thing as how they *actually* perform (Fordyce 1984). Mooney stated in 1990 that an evaluation based on functional capacity testing "certainly is a great improvement over basing it on simple alteration in the patient's report of pain and function" (Mooney 1990, p.112), and Waddell and co-workers argue for the use of "some form of objective information" (Waddell et al 1993).

Classification. In studies concerning spinal pain, classification of subjects is a problem. There is no consensus on how to classify people with disability due to spinal pain. The earlier cited definitions of lumbar and cervical pain are basic anatomical definitions, but numerous clinical conditions exist. The Task Force on Taxonomy of the International Association for the Study of Pain have prepared a detailed classification guide for chronic spinal pain conditions, based on known anatomical or physiological morbidity (Merskey and Bogduk 1994). But since a specific cause of the pain condition can be shown clearly in only about 20% of people with spinal pain, (Nachemsson et al 2000, p.327), such classifications can seldom be used in the clinic. The Quebec Task Force on Spinal Disorders considered the most valuable classification system to be based on clinical signs and on symptoms (Spitzer et al 1987).
Topography, i.e. the site of pain, can be used for classification, e.g. in a classification scheme by Spangfort used in this thesis (Spangfort 1995).

Moffroid and co-workers proposed a classification scheme based on clinical physical measurements from the NIOSH Low Back Atlas (Moffroid et al 1994). Krause and Ragland proposed a scheme based on duration of work disability, and which took other biomedical, developmental and social characteristics into account (Krause and Ragland 1994). Identification of subgroups of persons at high risk for developing chronic pain conditions or going on long-term sick leave has recently been proposed. Skargren and Öberg identified five prognostic factors which could identify persons at high risk for a poor prognosis; a pain duration of at least one month, high Oswestry score (a questionnaire measuring the influence of pain on daily life, including social consequences), more than one pain location, low expectations of treatment and low well-being as measured on a six-point scale (Skargren and Öberg 1998). Linton developed a questionnaire for screening purposes, which identifies persons with back or neck pain with a poor prognosis for accumulated sick leave (Linton and Halldén 1998).
In the present thesis, the study participants regarded as patients were persons with long-term spinal pain conditions. They were or could be in question for a rehabilitation programme with a cognitive-behavioural approach, including physical exercise. They were either on sick leave or had recurrent sick leaves due to spinal pain.

Physiotherapy in rehabilitation for long-term spinal pain

"Physiotherapy is concerned with the ability of the individual to perceive, control and in a purposeful way use his or her body with regard to demands that come from the physical and social environment" (LSR 1998). "Physiotherapy aims to enhance health through movement and through different treatments and rehabilitation interventions which may improve, preserve, or compensate for disorders or health problems arising as a consequence of disease or injury, including physical and psychological overload" (translated from Broberg 1997 p.12).
As with the lack of knowledge of the origin of most spinal pain conditions, science-based knowledge of what treatment methods to use for the individual person is very limited. Physiotherapists' clinical experience and skills and the response from the patient guide the choices of adequate methods for the individual.
There is strong evidence that exercise prevents spinal problems (Linton and van Tulder 2001). There is, however, a great need for high-quality clinical studies for evaluating other treatment methods and for replication of others' study results.

Evidence-based physiotherapy

In recent years, the term 'evidence-based methods' has been introduced. To work with evidence-based methods is to use methods shown to

be useful in controlled clinical studies. The term has been introduced in several medical areas, including physiotherapy. Clinical studies examining the possible treatment effects of different physiotherapy modalities and treatment strategies are nowadays much more frequent than during the past decade. This is very promising and important for physiotherapy as a profession. It seems appropriate to stress the importance of clinicians being involved in the planning of these studies, since clinicians can contribute with their knowledge and experience, thus possibly avoiding research procedures and strategies less likely to show relevant results.

Remember that even though a certain treatment method has not yet been showed to have effects on a certain condition or population, this is no evidence for its inefficacy.

In research reports, the results concern almost exclusively *groups of subjects*. There will always be individual cases that react differently. The clinical experience of the physiotherapist, as well as the experience and the progress of the patient, are very important determinants of what treatment method to choose, but it is very important that all physiotherapists are aware of the possibilities and shortcomings of different treatment modalities. Not only this: they should also implement their knowledge in their clinical decision-making, so that our profession may become truly evidence-based.

ICF

When deciding what assessments we would like to use for a particular purpose, we need to define what we want to measure. A helpful instrument for defining what concepts to measure is the International Classification of Functioning, Disability and Health (ICF) (WHO 2001). In the ICF, the World Health Organisation has created a framework that can be used for classifying consequences of disease or dysfunction.

"In the context of health:
Body Functions are the physiological functions of body systems (including psychological functions).
Body Structures are anatomic parts of the body such as organs, limbs and their components.
Impairments are problems in body function or structure such as a significant deviation or loss.
Activity is the execution of a task or action by an individual.
Activity Limitations are difficulties an individual may have in executing activities.
Participation is involvement in a life situation.
Participation Restrictions are problems an individual may experience in involvement in life situations.
Environmental factors make up the physical, social and attitudinal environment in which people live and conduct their lives." (WHO 2001)

In this thesis, eleven physical performance tests were evaluated. The tests were designed to measure impairments and activity limitations for persons with long-term spinal pain.

Assessments in physiotherapy

"Physiotherapy, like medicine and law, will always remain partially an art, but without measurement it can be nothing more than art" (Rothstein 1985, p.1).

Assessment is the first step in the process of rehabilitation, and are important for identification and quantification of problems the individual may have, and of factors relevant for resolution of the problems (Wade 1998).
Assessments are necessary for a) establishing a baseline level of impairment or activity limitation for the individual, b) setting relevant goals for the treatment period, and c) evaluating the interventions. The assessments chosen should be adequate for the condition under examination. For clinical use, the assessment methods should be easy to perform in a clinical situation (for the physiotherapist as well as the patient) and to interpret. Physiotherapy assessments should also result in values that can be understood and accepted by other vocational groups in rehabilitation.
Physiotherapists are expected to evaluate disability in persons with spinal pain. The disability is often multidimensional, classified as impairment, activity limitation or participation restriction according to the WHO classification (WHO 2001). Hitherto, no consensus has been established on what measures are preferable.

Physiotherapists in general, in the author's experience, use outcome measures such as questionnaires or performance tests to a limited extent. In 1995 many Canadian physiotherapists were, according to a questionnaire, dissatisfied with their current methods for documenting their patients' progress (Basmaijan et al 1994, p.19). Jette advocated that physiotherapists should shift their interest from impairment measurements to disability measures (from 2001, 'activity limitations': author's comment) in their research efforts (Jette 1995).

Physical performance tests

When searching Medline from 1966 up to March 2002, the term physical performance test (PPT) first appeared in an article by Baumgartner 1969, reporting of a study examining reliability for a 'broad jump test' and a 'side-step test' in healthy students. The term has in the 1990s been used for tests where the test persons (TP) perform some kind of physical activity. Test results are often expressed as time taken, number of repetitions managed or distance covered. No expensive equipment is typically needed for the testing. PPT are usually performed under

controlled conditions that may not reflect the reality of the patient's daily life. There are, though, several advantages connected with the use of PPT: a) they are little dependent on language or education level, b) they measure a different perspective of impairment and activity limitations in the observational aspect of the TP in motion; how he or she uses the body, the quality of movement, and body awareness aspects, c) PPT allows for collection of other information which emerges during the tests and which otherwise would not easily be revealed, such as aspects of attitude, beliefs and behaviour, d) PPT constitute a good opportunity for discussions about exercise habits and health, and e) they supplement the individual's self-rated opinions, which is important for completeness. PPT are safe for the participant, but screening by a physician is strongly recommended for avoiding infections, other diseases or risks which could affect results or which involve increased risks during the PPT (Gordon et al 1995).

Physical performance tests for long-term spinal pain

Some examples of PPTs known at the start of the present studies and which had been used for persons with spinal pain, are included in this review. Only PPTs which are easy to perform in a clinical situation, without need for expensive or cumbersome equipment, are included.
As can be seen below, there has been much less research on assessments of cervical pain conditions than on LBP conditions.

Impairments

Aerobic capacity. Cardiovascular fitness in persons with long-term LBP has been examined by Reilly and co-workers using time spent walking on a treadmill and cycling on an ergometer at constant resistance and speed (Reilly et al 1989). McQuade and co-workers used a submaximal test on a computerised bicycle ergometer (McQuade et al 1988). Lindström et al used the Åstrand submaximal ergometry test for persons with subacute LBP (Lindström et al 1992).

Muscular strength or endurance. Isometric muscular endurance of the abdominal muscles (held in a curl-up position) in sub-acute LBP subjects was used by McQuade and co-workers (McQuade et al 1988) and by Lindström and co-workers (Lindström et al 1992). Dynamic sit-ups to tolerance was used for persons with chronic pain by Harding and co-workers (Harding et al 1994), and by Alaranta and co-workers for persons with chronic LBP (Alaranta et al 1994).
The Sörensen test, or a slightly modified version, has been used for persons with LBP by several authors (McQuade et al 1988, Lindström et al 1992, Alaranta et al 1994). It is an isometric endurance test for the back extensor muscles. In the original test, the TP lies prone with the legs strapped to a bench. The TP is asked to hold the unsupported trunk horizontally until exhaustion, maximum 240 s (Biering-Sörensen

1984). A dynamic version of the test was used by Alaranta and co-workers (Alaranta et al 1994).
In an arm strength test, the TP was asked to pull down the arms from a 90- degree elevated position against a dynamometer. This test was used for persons with sub-acute LBP (Lindström et al 1992).

Activity limitations

Walking tests. Several different gait tests have been reported by Harding and co-workers: six-minute and ten-minute walking tests, where the distance managed is measured, and maximal walking speed at different distances (Harding et al 1994).
Stair-climbing. In several studies, different types of climbing test have been used (Lindström et al 1992, Harding et al 1994).
In repeated-sit-to-stand-on-a-chair, the number of repetitions managed in a certain time is recorded (Harding et al 1994).
Repetitive squatting. The TP was asked to stand with the feet 15 cm apart and then to squat until the thighs were horizontal, hereafter returning to the standing position. The procedure was repeated as many times as possible with a maximum of 50 (Alaranta 1994, two sources).
Pushing and pulling. A heavy wheeled object was to be moved five metres (Lindström et al 1992). The ability was rated on a three-graded scale (without difficulty - some difficulty – cannot manage).
Lifting tests. Lindström and co-workers (1992) asked the TP to lift a load of maximum tolerable weight to different heights (floor to 90-cm-high table, table to 25 and 50 cm high boxes placed on table) and in different techniques (from one table to another at right-angles to the first, and from the table to a shelf placed under another table). Di Fabio and co-workers used a maximum isometric lift, MIL, for persons with LBP. The TP was required to pull a dynamometer handle attached to the floor upwards as much as possible without aggravating pain (Di Fabio et al 1995). Mayer and co-workers developed the PILE tests (Progressive Isoinertial Lifting Evaluation), where the TP is asked to lift a plastic box containing bottles of augmenting weight in a self-selected manner; in the lumbar test from floor to a shelf at 76 cm (floor-to-waist), and in the cervical test from the shelf at 76 cm to another shelf at 137 cm (waist-to-shoulder). The maximum weight managed for each test, together with heart rate and reason for ending the test, were recorded (Mayer et al 1988).

In only a few of the cited studies are clinimetric properties examined or described.
In the study by Harding and co-workers, the sit-ups, the walking test, the stair-climbing test, and the repeated sit-to-stand tests were all shown to have high reliability between raters, and the scores were improved by treatment (Harding et al 1994).
The MIL test was responsive to change after a physiotherapy intervention using multiple treatment methods (Di Fabio et al 1995).

The PILE tests were responsive to change after a 'functional restoration program' (Mayer et al 1988 b).

Since 1995, increased interest in PPT has resulted in a number of published studies in the field, some of which will be discussed later.

Associations between physical performance tests and other measures

The association between physical measurements of impairments such as muscle strength and flexibility and the severity of the pain condition is controversial (Turk and Melzack 1992, p.6).

Lumbar pain

The first study that revealed the relationship between decreased back extensor endurance and LBP was the classical paper by Biering-Sörensen from 1984. His results were replicated by Luoto and co-workers in 1995. In a review article, Rodriquez and co-workers found a high correlation between chronic LBP and decreased muscular strength and endurance in trunk muscles (Rodriquez et al 1992). In an often-cited study, high overall fitness level, as measured by flexibility, an isometric lifting strength test and an ergometry test, was shown to be related to a lowered risk of recurrent LBP (Cady et al 1979). Hirsch and co-workers found that persons with LBP who expressed excessive illness behaviour according to the Waddell Score performed significantly worse on tests of lumbar isometric strength (Hirsch et al 1991).

Other authors have not produced such straightforward results; Mellin and co-workers found low correlations between improvements in strength and mobility and self-rated improvement (Mellin et al 1989, Part II). Hazard and co-workers (Hazard et al 1994) found low-to-moderate correlations between a combined measure of lifting capacity (PILE tests) and trunk range of motion and pain intensity as rated on VAS,and disability as rated on Oswestry pain questionnaires, in persons with chronic LBP. Lindström and co-workers found no clear correlations between individual physical capacity and other measures such as psychological capacity, physical work demands and LBP (Lindström et al 1994).

Cervical pain

Rodriquez and co-workers suggest in their review article that there is a correlation to decreased muscular strength and endurance for chronic cervical pain, as for LBP, but found only two studies to support this suggestion: (Kraut and Anderson 1966, Silverman et al 1991).

Salén and co-workers reported moderate correlations between performance of an obstacle course and self-rated disability for persons with neck-shoulder or low-back pain (Salén et al 1994)

Clinimetric properties

The measurement qualities of a questionnaire, are termed its *psychometric* properties. The term '*clinimetric*' was first used by Feinstein in 1967, and was suggested as a way of categorising clinical data into measurable units (Feinstein 1967, Feinstein 1983). The term has later been used when discussing qualities for measurement developed merely for reality use in the clinic (Dijkers 1999). Psychometric/clinimetric properties can roughly be divided into reliability, validity and sensitivity to change/responsiveness.

Reliability

Reliability is defined as the consistency of a measurement when all conditions are held constant (Rothstein 1985, p.5). For measurements to be considered reliable, they must be comparable when performed with the same subject by numerous raters (inter-rater reliability) or when performed on several occasions with the same subject by the same rater (intra-rater reliability).
To know whether the individual's body function and activity have improved after our intervention, our measurement instruments have to be reliable, that is, we have to know how great the variation of the test result would normally be to be able to interpret the test result.
The term ***inter-rater agreement*** is used in this thesis for describing the degree of agreement between two physiotherapists testing 21 TPs simultaneously (Study III).

Repeatability is the degree of consistency over test occasions. The term is used in this thesis for describing degree of stability in PPT results over test occasions (Study III).

Validity

Validity is a complex concept, which is not easily established. The traditional definition runs: "the extent to which an instrument measures what it is intended to measure" (McDowell and Newell 1987). A more clinically applicable definition runs "The degree to which a useful interpretation can be inferred from a measurement"(Task Force on Standards for Measurements in Physical Therapy 1991). Valid applications of a test may thus go beyond the purpose for which the method was originally designed (McDowell and Newell 1987). If in e.g. a lifting test the TP turns out to be able to lift 10 kg:s as maximum – how do we interpret the clinical meaning for that individual?
There are several aspects of validity to consider, and no clear boundaries can be drawn between the different aspects (Basmajian 1995). The aspects of validity discussed in this thesis are:
Construct validity – how far the measurement instrument fulfils the theoretical assumptions underlying its construction. When a measurement instrument has construct validity, it fulfils the hypotheses concerning clinical applicability set up by the constructor. Construct validity has also been descibed as a continuing process of reciprocal verification of the measuring instrument and the theory

of the construct it is meant to measure (Angoff 1988).
Content validity – how well the contents of the instrument represent aspects of the topic being studied (Rothstein 1985).
Face validity is an aspect of content validity, and means that the instrument seems adequate according to experienced professionals. Face validity as the patient sees it is also important to remember – if our measurements are not experienced as meaningful to the individual, he or she may not perform optimally (Rothstein 1985).

Sensitivity to change / responsiveness

The ability to detect clinically important changes over time is called 'sensitivity to change' or 'responsiveness' (Deyo et al 1991). Recently, there have been proposals to use these terms differently: sensitivity to change should designate the ability of a measure to detect any change in health status, whereas responsiveness should mean the ability of a measure to detect clinically important change (Stratford et al 2002). When we want to evaluate our interventions, we must use measurement instruments that are reliable and valid. If these instruments, however, are not sensitive to change/responsive, we will not be able to detect any improvements. There are difficulties in knowing whether an instrument has the ability to assess sensitivity to change or responsiveness, because change is not always apparent. In the field of spinal pain, there are no 'gold standard' measures to tell us that a change really has taken place. And who is to decide what constitutes a 'clinically important change'? Stratford and co-workers define it as "a change that is important to the patient, clinician, or both" (Stratford et al 1999, p. 110). Another question to consider is the following: must negative results always indicate that the instrument is not sensitive enough, or can it be that there are no changes?

Aims

General aims

The overall aim of this thesis was to identify among eleven physical performance tests assembled in a test package, one or more that, based on clinimetric properties, could be used in clinical practice for

a) assessing impairments and activity limitations in persons with long-term spinal pain, and
b) contributing to a common basis for evaluation and treatment of persons with long-term spinal pain.

Specific aims

The more specific aims were

- to explore the underlying foundations for ratings made by physicians, physiotherapists and insurance officers involved in an individual's rehabilitation concerning
  a) rated need of rehabilitation and b) rated potential to benefit from rehabilitation,
  for persons with long-term spinal pain (Study I).

- to examine
  a) the intra- and inter-rater reliability (Study III),
  b) construct validity (Study II and IV), and
  c) sensitivity to change / responsiveness (Study V)
  of the physical performance tests.

- to identify the clinical usefulness of the eleven physical performance tests from the studies performed.

Methods and subjects

Overview of subjects in the studies

In the present thesis, the study participants regarded as patients were persons with long-term spinal pain conditions. They were or could be in question for a rehabilitation programme with a cognitive-behavioural approach, including physical exercise. They were either on sick leave or had recurrent sick leaves due to spinal pain. In Studies I, IV and V, persons with long-term spinal pain participating in a national randomised controlled multicentre study (RCT) constituted the study group. They were identified from the AGS insurance register, which covers 2.5 million employees in Sweden. The RCT was conducted to evaluate the outcome of a behavioural medicine rehabilitation programme and that of its two main components (behaviour-oriented physiotherapy and cognitive behavioural therapy), as compared to that of a "treatment-as-usual" control group (Jensen et al 2001). The inclusion criteria in the RCT were spinal pain, sick leave for between one month and six months for spinal pain, fluency in Swedish, and age between 18 and 60 years. Basic demographic and background data are shown in Table I.

In Figure 1, the relations between Study I, Study IV and Study V are illustrated. In Study I, we had data on only 217 of the 235 persons included in the RCT, because the questionnaire concerning rehabilitation needs and potential was not distributed to the first eighteen subjects included. In Study I,

Table 1. Demographic and background data for the RCT participants (n = 235) expressed in median values with inter-quartile range (IQR), and in numbers (%).

	Median(IQR)
Age (years)	46.0 (17)
Number of days of sick leave 1 year before inclusion	144.0 (79.0)
3 months before inclusion	90.0 (26.8)
Pain duration current condition (months)	8.0(26.0)
	Number (%)
Men	106 (45)
Women	129 (55)

273 subjects participating in another study were also used as a reference group for validating the predictive power of the subject's ratings of beliefs concerning existing effective treatments and future coping ability. In Study IV, all 235 subjects of the RCT were included, while Study V covered only the 186 subjects who stayed in the RCT for the whole 6-month follow-up period (Figure 1).

In Study II-III the subjects were a) 15 patients with long-term spinal pain referred to a rehabilitation clinic in southern Sweden, HI Blekinge, b) 15 persons, who had no complaints of spinal pain and were matched by age, gender and occupation, identified by an occupational nurse at an occupational health service centre, c) 21 persons participating in the RCT, d) 12 patients with long-term spinal pain referred consecutively to HI Blekinge, 4 patients referred to HI Stockholm, and 8 persons included in a pilot study preceding the RCT, and e) a further 23 persons (11 + 12) with no complaints of spinal pain (Figure 2).

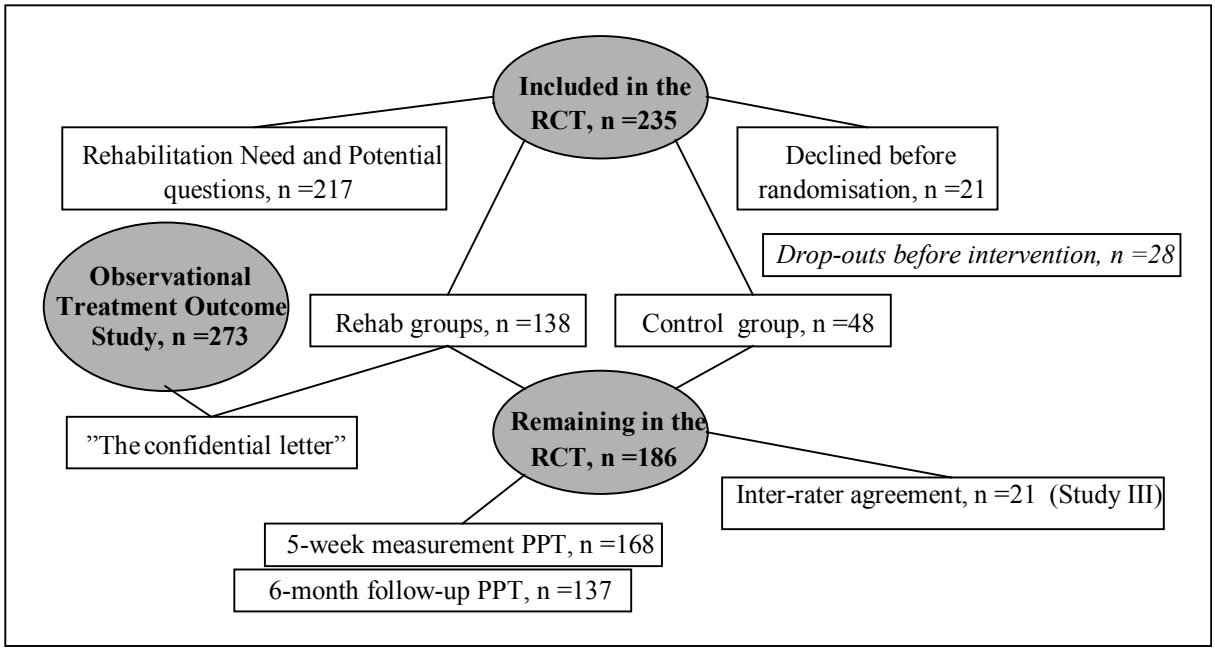

Figure 1. Relations between Studies I, IV and V with regard to study participants.

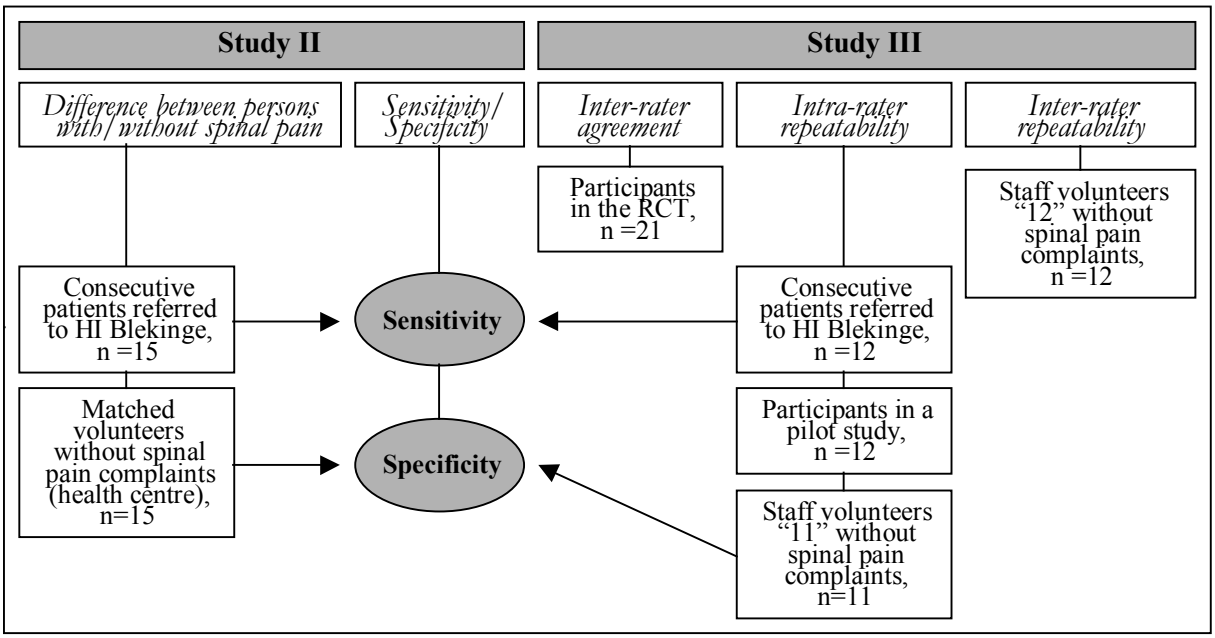

Figure 2. Origin of subjects participating in Studies II – III. HI Blekinge is a rehabilitation clinic in southern Sweden.

Study designs

Study I

Two hundred and thirty-five persons with long-term spinal pain participating in a national randomised controlled study (RCT) were included. In a prospective study, a questionnaire concerning the participants' rated need for rehabilitation and their rated potential for improving from rehabilitation was distributed to professionals involved in each participant's rehabilitation, to the attending physician, to the attending physiotherapist, and to the health insurance officer in charge. In addition, the screening physician attached to the RCT was asked to rate the same constructs. The participants randomised to any of the rehabilitation programmes were asked to rate their belief in existing treatment modalities effectiveness for their pain condition, and also to rate their belief in their ability to learn to cope with the pain ('The confidential letter'). The ratings of the professionals were compared for agreement, and all ratings were analysed for predictive value for self-rated health and sick leave at 6-month follow-up.

Study II

The *discriminative ability* of the physical performance tests for persons with long-term spinal pain versus persons with no spinal pain complaints was examined in two ways: 1) Fifteen persons with long-term spinal pain and fifteen age-, gender- and occupation-matched back-healthy persons went through the test package once. The persons with long-term spinal pain were consecutive patients

referred to HI Blekinge. The PPTs were part of the usual examination procedure at the particular rehabilitation clinic. The matched back-healthy persons were identified from a local occupational health service centre by an occupational nurse. They were contacted and informed about the study by one of the authors. All contacted except one agreed to take part. 2) The sensitivity and specificity of the tests were analysed after including the test results from the first test occasion for a further 12 persons with long-term spinal pain from HI Blekinge and 11 unmatched, back-healthy participants from Stockholm, participating in Study III (Figure 2).

Study III

The *reliability* of the test package was examined for four samples: 21 persons with long-term spinal pain participating in the RCT went through the test package once under physiotherapeutic guidance, with another physiotherapist acting as a passive co-assessor *(inter-rater agreement)*. Altogether, three physiotherapists were involved, but only two at a time. The degree of agreement in PPT results between the two raters was examined.
Twenty-four persons with long-term spinal pain and eleven 'back-healthy' persons underwent the test package three times within a week, administered by the same physiotherapist *(intra-rater repeatability)*. The persons with long-term spinal pain were recruited from HI Blekinge, from HI rehabilitation clinics in Stockholm, and from a pilot study preceding the RCT. The 'back-healthy' persons were recruited from the RCT staff and from the staff in the rehabilitation clinics in Stockholm. They considered themselves as 'back-healthy', and had no complaints of pain in the back or in the neck.
Another 12 back-healthy persons went through the test package three times within a week, administered each time by a different physiotherapist *(inter-rater repeatability).* These persons were recruited from staff connected to the RCT in Stockholm. The degree of individual variation in PPT results between test occasions was examined for the three repeatability samples.

Study IV

In Study IV, the 235 persons participating in the RCT were included. All but four went through the physical performance tests at inclusion in the RCT. Several hypotheses concerning the effect of related factors and background factors on test results were developed. C*onstruct validity* was examined for the six tests considered to have acceptable reliability (Study III) by testing the hypotheses developed.

Study V

Study V included 186 persons participating in the RCT for the whole treatment period.The physical performance tests were administered on inclusion in the RCT, after 5 weeks, and after 6 months. *Sensitivity to change /responsiveness* was examined for the six PPTs considered to have acceptable reliability according to

Study III by relating the changes in PPT results to the self-rated changes in general health, pain affecting work, and self-efficacy; and by considering differences in change scores in persons with low performance and high performance, respectively, at baseline. The degree of change for persons rating improvement/deterioration was also considered.

Measurements

Table 2. Measurements used in Study I-V.

	Study				
	I	**II**	**III**	**IV**	**V**
Åstrand test		X	X		
Isometric endurance tests		X	X		
Step-on-stool test		X	X	X	X
PILE lifting tests		X	X	X	X
Self-selected walking speed		X	X	X	X
CR10		X	X	X	X
RPE		X	X	X	X
UAB		X	X	X	X
Pain drawing	X			X	X
SF-36	X			X	X
ASES, pain dimension	X				X
Rehabilitation Needs and Potential	X				
Background data	X			X	X
Medical examination	X				
Sick-leave and disability pension	X			X	

The physical performance tests

The test package was based on existing tests used in earlier studies as well as specially developed tests. Some tests were designed mainly for measuring dysfunction due to lumbar pain and others for measuring dysfunction due to cervical pain.

Face validity and content validity. The choice of tests was discussed by a selected group of skilled physiotherapists. The group's suggestion was discussed and revised in a group of fifteen physiotherapists throughout Sweden. All were experienced in rehabilitation for persons with musculoskeletal pain. It was decided that all subjects should perform all tests regardless of pain site, since in our experience pain in one site tends to influence other areas too. To estimate approximate normal values, where such values were not established in earlier studies, we had staff members at a rehabilitation clinic perform the tests. Since the values collected from these persons were considered very high, we chose to set the end points of the actual tests a little lower, based on clinical judgement of what limit could be clinically relevant and not too time-consuming. The face value and

content validity properties were considered fulfilled by this process of assembling the tests.

The definitive test package, consisting of eleven separate tests, took about one hour to complete.

Assessment of impairments

– ***Bicycle ergometry, the Åstrand test*** (Åstrand and Rhyming 1954). The test person (TP) cycled on an ergometer bicycle with a fitness computer, Monark Ergomedic 829E , for 6 minutes or until steady state was achieved. The resistance was chosen from the TP´s heart rate during the first 2 minutes, to achieve a steady-state heart rate of at least 120 beats per minute, a value which represents the limit for possible calculation of the VO_2 ml·kg^{-1}·min^{-1} (Åstrand and Rodahl 1986). Oxygen consumption was estimated from the known linear relationship between heart rate and oxygen consumption at submaximal workloads (Åstrand and Rodahl 1986). Test result was expressed as VO_2 ml·kg^{-1}·min^{-1}. The TP was instructed not to eat or smoke, or to perform excessive physical activities for at least two hours before the tests. The ergometer weight was calibrated regularly.
The main aim with the inclusion of the test was to measure cardiovascular capacity. We considered high fitness levels to bring certain resistance to spinal pain, and that therefore cardiovascular fitness training should be included in the rehabilitation programme (Cady et al 1979). Moreover, the test had a "warm-up" function for safer performance in the other tests.

– ***Isometric endurance in neck flexors and in neck extensors.*** In ***the neck flexor test***, the TP lay supine on a bench with a goniometer ad modum Myrin fixed just above the ear and a weight of 0.5 kg on the forehead. The TP was asked to retract the chin and to lift the head from the bench to 10° of cervical spine flexion, and to maintain that position. The time managed was measured with a stopwatch, in seconds. The test was discontinued after 60 s. A revised version of the test has later been descibed by Alricsson et al 2001.
In ***the neck extensor test*** modified from Harms-Ringdahl et al 1991, the TP lay prone on a bench with head and cervical spine unsupported. A weight of 1.5 kg for women and 2 kg for men was placed on the back of the head, and the TP was asked to hold
For ***the trunk extensor test***, modified from Biering-Sörensen (Biering-Sörensen 1984), the TP lay on an angle table, an adjustable medical exercise therapy bench (Holten 1976). In the start position, the hip angle was 55°. This position was chosen to achieve a more favourable working moment for the hip extensor muscles (Németh et al 1983). The feet were fastened under a cylinder, and the upper body leaned against the front of the angle table. The TP lifted the upper body up to horizontal level, without accenting the lumbar lordosis, with the hands placed on the sacrum and retracted chin. The time managed in that position was recorded in seconds. The test was discontinued after 360 s in Study II, and after 180 s in Studies III-V.

the head steady with the chin retracted and the cervical spine in a zero position for as long as possible. The time managed was recorded in seconds. The test was discontinued after 360 s in Study II and after 180 s in Studies III-V.

– ***Isometric endurance in trunk flexors and extensors.*** For ***the trunk flexor test***, the TP lay supine on a bench with knees flexed, heels about 0.30 m from buttocks. The TP rounded the cervical and thoracic spine and lifted the arms until the palms were level with the knees, so that the angulus inferior of the scapula was barely lifted from the bench. The time the TP managed to hold steady was recorded in seconds. The test was discontinued after 90 s.

The main aim with the inclusion of the isometric endurance tests was to measure muscular endurance. Based on experiences from earlier studies, muscular endurance was assumed to be related to spinal problems (Krout and Anderson 1966, Biering-Sörensen 1984, Reid et al 1991, Silverman et al 1991).

– ***Dynamic endurance for lower extremities, the step-on-stool test,*** (with approval from Selles, personal communication 1997). The TP was asked to step up on and down from a specially designed, solid wooden stool. The step height was 0.40 m for women, 0.44 m for men. The legs were tested separately, so that one leg at a time was the ”working leg”, and the other the ”supporting leg”. The

number of steps managed was recorded. The test was discontinued after 100 steps in Study II and after 50 steps in Studies III-V.
This test was developed in the PT group involved in the test package assembling. The main aim with the inclusion of the test was to measure muscular endurance. Muscular fitness in lower extremities has been considered important for persons with back problems, based on the empirical knowledge that flexion in the hips and knees helps to decrease the pressure on the back in situations were the trunk is flexed. In a study by Lee and co-workers (Lee et al 1995), strength in lower extremity seemed to be as affected as trunk strength in persons with back problems.

Assessment of activity limitations

– ***Lumbar and cervical lifting tests, PILE tests*** (Mayer et al 1988). The lifting tests were performed standing in front of bookshelves with shelves at 0.76 m and 1.37 m from the floor. The TP was asked to lift weights in a plastic box from floor to waist (0 - 0.76 m) for the ***PILE lumbar test***, or from waist to shoulder height (0.76 - 1.37 m) for the ***PILE cervical test***. The initial weight was 3.6 kg for women and 5.9 kg for men. A ´lifting movement´ involved a single transfer from one level to the next and back again. After every four such lifting movements (=20 s), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last four lifting movements was recorded and used as a test result (Mayer et al 1988). The weights consisted of plastic bottles filled with sand, and their weight was checked and adjusted regularly with a digitalized letter balance. The tests were discontinued if the heart rate, as measured by an electronic pulse counter attached to the thorax on the TP, was at 85 % of the estimated maximal level, adjusted for age, or if the weight lifted was level with 55 % of body weight.
The main aim with the inclusion of the lifting tests was to measure activity limitations in a physically demanding, potentially pain-provoking task. Lifting tests were considered important to include because many persons with back or neck pain report difficulties in lifting things. The PILE tests were chosen because they were safe, had been thoroughly described, and because they had shown sensitivity to change (Mayer et al 1988).

– ***Three gait tests.*** Self-selected walking speed was measured with a stop-watch:

1) In ***the gait test***, the TP was asked to walk at a comfortable speed 20 m along a corridor and to turn around where 20 m was marked.
2) In ***the gait test with burden***, the TP repeated the procedure, now carrying one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. These two tests were discontinued after 50 s.
3) In ***the stair-climbing test***, the subjects were asked to walk up and down a flight of stairs at a comfortable speed, preferably without support. The stairs had 18 to 20 steps, the number differing

between clinics. To standardise the measurements, the heights of the stairs were recalculated into a standardised height for Study V. The test was discontinued after 35 s.

The gait tests were developed by the PT group involved in the test package assembling. The main aim with the inclusion of the tests was to measure activity limitations regarding walking in a situation as close to real life as possible. The weight-carrying was assumed to reveal special difficulties for persons with problems in the upper limb and the cervical spine. We assumed that the comfortable speed would be lower when the patient was in pain. Persons with LBP, waiting-listed for surgery, walk more slowly than pain-free controls (Khodadadeh and Eisenstein 1993). When walking at the self-selected speed, the energy consumption is likely to be smallest (Ralston 1958). We also considered self-selected speed to be more applicable to real life for persons with long-term spinal pain, compared to maximal speed.

Procedure for the physical performance tests

A detailed manual was developed for standardising methods, instructions to test takers and interpretations. The tests were arranged in a fixed order which allowed them all to be performed consecutively without pausing, taking account of the possible fatigability in muscle groups active in the previous test. The order is presented in Table 3. The only exception from the fixed order was that if the TP managed more than 20 steps on the step-on-stool test with the dominant leg, the neck extensor test was performed before the step-on-stool test with the non-dominant leg.

Table 3. The physical performance tests in the fixed order of testing.

1	Åstrand test
2	Neck flexor test
3	PILE lumbar test
4	Back extensor test
5	Step-on-stool test
6	Neck extensor test
7	PILE cervical test
8	Trunk flexor test
9	Gait test
10	Gait test with burden
11	Stair-climbing test

Before the Åstrand test, body weight and body height were recorded. The TPs wore exercise clothing. The test leader, i.e. the physiotherapist, asked the TPs to try their hardest, but to take their pain and fatigue into account. It was emphasised that they could discontinue each test any time or decline a test completely. The TP was told that he or she was responsible for the limit chosen in each test. The test leader demonstrated and explained each test. The TP was allowed to try the technique for the step-on-stool test and for the PILE tests before starting the tests. During the testing, two corrections of technique or speed were allowed before the PT was to decide to discontinue a test. The administrating PT did not encourage the TPs in any way during testing.

Perceived pain intensity

After each test, the TP was asked whether he or she had experienced any pain during the test, and if so, the TP was asked to rate the perceived pain intensity on the CR10 Scale (Borg 1982). The CR10 scale is a **c**ategory scale with certain **r**atio properties with **10** scale steps, and with an additional possibility to rate 'maximal pain' (= 11), see Table 4. It was assumed that it would be important to obtain knowledge about possible pain intensity level during testing, while we hypothesised a certain correlation between pain intensity and test performance.

Perceived exertion

After each test, subjects rated the perceived exertion during the test on the Rated Perceived Exertion (RPE) Scale (Borg 1970). The scale is based on the principle that a person can perceive and interpret feelings of muscle strain, joint loading and efforts on the cardiorespiratory system while exercising. The scale has fifteen points ranging from 6 to 20 (Table 4), each corresponding to an approximate range of heart rates if multiplied by the constant ten (± 10 heart beats/minute). A rating of 12 to 13 corresponds to 60-80 % of VO_2 max in most healthy individuals, while a rating of 16 to 17 corresponds to 90 % of VO_2 max approximately (Williams & Eston 1989). The scale has been used as a measure of overall strain. The participant is asked to rate his or her perception of exertion, i.e. how heavy and strenuous the exercise feels.

Table 4. The Borg Scales used for ratings of pain intensity and perceived exertion during the physical performance tests.

Borg´s CR-10 Scale (Borg 1982)		Borg´s RPE scale (Borg 1970)	
0	Nothing at all	6	
0.5	Just noticeable	7	Extremely light
1	Very weak	8	
2	Weak	9	Very light
3	Moderate	10	
4	Somewhat strong	11	Fairly light
5	Strong	12	
6		13	Somewhat hard
7	Very strong	14	

8	
9	
10	Extremely strong
•	Maximal

15	Hard (heavy)
16	
17	Very hard
18	
19	Extremely hard
20	

In this thesis, the RPE scale was used mainly to rate subjective local exertion and fatigue from the working muscles involved, but the cardiorespiratory strain was naturally included in the overall perception.
One assumption when adding the Borg Scales was that the perceived exertion could change during a rehabilitation period, so that the individual could make the same performance with less exertion. Another assumption was that it would be possible to understand the individual's reason for ending a test – if not because of pain, perhaps it would be because of high perceived exertion. If, on the other hand, neither pain nor high perceived exertion was present as determinators of performance, perhaps the TP's motivation for performing the test could be low.
In Study III, the RPE scale was used for encouraging participants to regulate their performance on the second and third test occasions according to the RPE rated, and considered relevant, on the first test occasion. This type of application of the RPE scale is reportedly valid (Williams & Eston 1989).

Pain behaviour

After completing the test package, the participant's pain behaviour during the testing procedure was rated on the UAB Pain Behavioral Scale (Richards et al 1982) by the PT in charge. Only behaviour which could be seen or heard during the testing procedure was recorded, no questions were asked concerning e.g. medication consumption (Table 5). Pain behaviour was hypothesised as correlating with test performance. The UAB Scale possesses high inter-rater reliability and repeatability over time (Richards et al 1982). Pain behaviour as rated by the UAB scale is moderately inversely correlated to observed physical activity, suggesting that the two concepts are related, but not "mirror images" (Richards et al 1982). Two factors have been identified from ratings on the UAB scale, one for facial/audible pain behaviour and one for motor pain behaviour (Öhlund et al 1994). UAB score correlates positively to time till return to work, that is, the higher pain behaviour, the longer the time off work (Öhlund et al 1994).

Pain drawing

Pain site was identified from pain drawings filled in by the participants on inclusion in the RCT (Studies I, IV and V), and interpreted by the project physician according to the 'topographic classification of spinal pain' (Spangfort 1995). Pain drawings are frequently used in the clinic as well as in research studies, and show properties of construct validity for persons with LBP (Ohnmeiss 2000) and those with cervical pain (Toomingas 1999).

SF-36

The Short Form 36 (SF-36) is a questionnaire commonly used for measuring self-rated health-related quality of life. The scale has 36 items designed for measuring eight physical and mental health constructs

(McHorney et al 1994). It has been recommended for use as part of a 'standard questionnaire package' in rehabilitation following spinal pain (Deyo et al 1998). SF-36 was used in the logistic regressions in Study I examining the possible explanatory effect of different background factors on ratings of rehabilitation needs and potential. Single questions included in the SF-36 were used in Studies IV and V. In Study IV, we used the question "How much bodily pain have you had during the past 4 weeks?", rated on a scale ranging from one to six, higher rating meaning more pain. The hypothesis was that people rating pain intensity as "severe" or "very severe" would have lower test performance. In Study V, two questions were used; a) "Compared to one year ago, how would you rate your health in general now?", and b) "During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?", both rated on scales from one to five, higher rating meaning worse general health/more pain disturbance. These questions were used as outcome measures, hypothesised to have a high potential for improvement.

Arthritis Self-efficacy Scale (ASES), pain dimension

Measures of self-efficacy have been used frequently in recent years. Self-efficacy concerns the feeling of how successfully one could cope with different tasks, situations or symptoms. The concept is responsive to change after exercise intervention (Stenström 1994) and is important for outcome after rehabilitation (Söderlund 2001). In a recent study, perceived self-efficacy was shown to be inversely correlated with pain intensity and pain interference with daily life (Lin 1998). The Arthritis Self-efficacy Scale (ASES) was used in this thesis. The Swedish version shows construct validity in that a) the scale could discriminate between persons with chronic pain and persons with rheumatoid arthritis, and b) the scale correlated with other health-status measurements in the expected direction. The ASES has also showed sensitivity to change for women with fibromyalgia participating in a rehabilitation programme of self-management education and physical training (Lomi 1995).
The scale has three subscales: self-efficacy concerning pain, disability, and other symptoms. The pain subscale, including five questions, was used in Studies I and V. The questions range from 0 to 100, with 10-point intervals. Higher score indicates better self-efficacy, in this case higher belief in one's own capacity to cope with pain. In Study I, the mean value of the five questions was used in the logistic regressions examining the possible explanatory effect of different background factors on ratings of rehabilitation needs and potential. In Study V, the subject's median value for the five questions was used as an outcome measure, hypothesised to have a high potential for improvement.

Table 5. The UAB Pain Behavior Scale. Reprinted from Pain;14, Richards et al. Assessing pain behavior: The Pain Behavior Scale, p. 395, Copyright (1982), with permission from Elsevier Science.

1. Vocal Complaints, Verbal	None Occasional Frequent	0 0.5 1
2. Vocal Complaints, Non-verbal	None Occasional Frequent	0 0.5 1
3. Down-time	None 0-60 min >60 min	0 0.5 1
4. Facial Grimaces	None Mild and/or infrequent Severe and/or frequent	0 0.5 1
5. Standing Posture	Normal Mildly impaired Distorted	0 0.5 1
6. Mobility	No visible impairment Mild limp and/or mildly impaired walking Marked limp and/or labored walking	0 0.5 1
7. Body Language	None Occasional Frequent	0 0.5 1
8. Use of visible supportive equipment	None Occasional Frequent	0 0.5 1
9. Stationary movement	Sits or stands still Occasional shifts of position Constant movement, position shifts	0 0.5 1
10. Medication	None Non-narcotic analgesic and/or psychogenic medications as described Demands for increased dosage or frequency, and/or narcotics, and/or medication abuse	0 0.5 1
Total		

Rehabilitation Needs and Potential

A two-item questionnaire concerning the need of the participants in the RCT for rehabilitative intervention and their potential for benefiting from such an intervention was used in Study I. The questionnaire was sent to the attending physician and physiotherapist, and to the social insurance officer in charge of the case. The information enclosed was that the person had agreed to take part in an examination included in a study designed to investigate the natural course of spinal pain. The respondents were asked to rate the person's overall need for any type of rehabilitation and his/her potential for benefiting from rehabilitation. The ratings (on a 0-10 scale) were to be done regardless of

type or availability of intervention, and regardless of possible factors to consider.

Patient's beliefs concerning effective interventions and pain coping ability

Two questions were used in Study I to assess the person's own belief about whether there was any possible intervention that could relieve his or her pain condition (modified after Borkovec and Sidney 1972), and about belief in his or her ability to learn to cope with the pain (inspired by Bandura 1977). The questions were given to 120 persons participating in the RCT, and to 273 persons participating in an observational multicentre treatment outcome study at the end of their first day at the rehabilitation clinic. They were informed that these questions were not known to the staff at the clinic (the letter was called 'the confidential letter'), and that the questionnaire was to be posted in the enclosed stamped addressed envelope.

These questions were:

1. How certain are you that you can learn to cope with your pain? (0 = Not sure at all – 10 = Positive)
2. How sure are you that there is some existing treatment that could help you effectively? (0 = Not sure at all – 10 = Positive)

Background data

The participants were asked to provide personal and background data. In Studies II-III, questions concerning name, date of birth, profession, duration of any sick-leave and reason for sick leave were answered. In Studies I, IV and V, subjects answered a complete background questionnaire, including personal data, medical state, former treatments, pain sites and any related issues about present and previous pain.

Medical examination

The participants in the RCT were examined by a physician according to a standardised status formula. Measurements of spinal range of motion were included in the examination, and were used in Study I as independent variables in the logistic regression examining possible factors explaining the ratings of the professionals.

Sick-leave and disability pension

Sick-leave data for Studies I, IV, and V and data concerning disability pensions for Study I, were obtained from the National Health Insurance Authority (NHIA), which covers all employees in Sweden. Only sick leave periods exceeding fourteen days were included due to missing data for shorter periods.
Numbers of days on part-time sick leave were converted into a normative value of full days of sick leave.
In Studies II and III, all patients except three were on full-time sick leave when tested.

Statistical methods

An overview of the statistical methods used in this thesis is shown in Table 6. All statistical analyses used the SPSS for Windows (SPSS 1999).

In ***Study I,*** the agreement between professionals was analysed with intraclass correlation coefficients and Kappa statistics. Only Kappa statistics were reported, as the ICC showed values of <0.17. For examination of the possible effect of different background factors on the ratings of rehabilitation needs and potential, logistic regressions were used.

Multiple, logistic, and Cox regressions were used in the prediction analyses to reveal possible predictors for self-rated health-related quality of life (SF-36) and work status after 6 months.

In ***Study II***, the differences between the matched groups were analysed with the Wilcoxon signed ranks test for related samples, and sensitivity and specificity were calculated according to standard procedures (Bland 2000). A hierarchical cluster analysis was performed for revealing homogenous groups of PPT.

Limits of agreement (Bland and Altman 1988) were used for interpreting inter-rater agreement. Twice the within-subject standard deviation for the measurements between test occasions and/or raters were used for interpretation of intra- and inter-rater repeatability (***Study III***). In this summary, 2.77 x within-subject SD, which might be a more appropriate figure representing the value below which the difference between two measurements would lie with 95 % probability (Bland and Altman 1999), is used. In the summary, $ICC_{2,1}$, is also included for a more complete interpretation of the 'Patients' sample. The ICC represents the proportion of variation in test performance which emerges between subjects. The higher the value, the lower variation within each individual.

In ***Study IV***, Mann Whitney tests were used to test the hypotheses concerning factors affecting the PPT results. The significance levels were adjusted by Bonferroni corrections (Bland 2000). Regressions (multiple and ordinal) were used to examine the explanatory degree of several background factors on PPT results. Sensitivity and specificity was calculated for the stair-climbing test.

In ***Study V***, Chi-square analyses were used to examine differences between the different rehabilitation groups, between subjects with differing compliance, and between men and women. Wilcoxon signed ranks tests were used for detecting differences in PPT results between test occasions for a) persons rating improvement on the outcome measures studied, and b) persons not rating improvement. Wilcoxon signed ranks tests were also used for examining differences in PPT results between test occasions for a) persons performing worse than the group median value at inclusion ('least fit'), and b) persons performing at least as well as the group median value on inclusion ('more fit'). The significance levels were adjusted by Bonferroni corrections (Bland 2000) when the analyses performed for a PPT exceeded one single analysis. Spearman correlations were used for

correlating changes in the PPT results with rated changes in the outcome measures. ROC curves were used for detecting the possible discriminative value of the PPT for persons rating improvement/not rating improvement on the outcome measures.

Effect sizes were calculated for persons rating improvement on the outcome measures, and for persons rating deterioration on the outcome measures at the 5-week measurement or at the 6-month follow-up.

Table 6. Statistical methods used in the different studies in this thesis.

	Study				
	I	**II**	**III**	**IV**	**V**
Kappa	X				
ICC	X				
Wilcoxon signed ranks test		X			X
Cluster analysis		X			
Diagnostic tests		X		X	
ROC curves		X			X
Limits of agreement			X		
2 x within-subject SD			X		
Mann Whitney U tests				X	X
Regressions (logistic, Cox, multiple, ordinal)	X			X	
Chi-square					X
Spearman correlations					X
Effect sizes					X

Results

Inter-professional judgements (Study I)

There was no consensus between different professionals involved in the individual's rehabilitation process regarding a person's need of rehabilitation or his or her potential to benefit from such rehabilitation (Kappa values < 0.20). The ratings of need for rehabilitation were based on the duration of sick leave (attending physician) and physical functioning according to SF-36 (the insurance officer in charge), i.e., the more sick leave/rated disability, the more the rated need for rehabilitation. The judgements about benefiting from rehabilitation were in most cases based on age, i.e. the higher age, the less rated potential. The variable best predicting health status and return to work after six months was the patient's ratings: The stronger the patient's belief concerning the existence of effective treatments, and the stronger the belief that he or she could learn to cope with the pain, the better the health, and the less the sick-listing at the six-month follow-up.

Reliability (Study III)

The inter-rater agreement was acceptable, though somewhat low for the neck extensor test, the step-on-stool test and the stair-climbing test. The repeatability coefficients and $ICC_{2,1}$ for three test occasions for 24 persons with long-term spinal pain (intra-rater repeatability) are shown in Table 7. For a more conclusive interpretation, median PPT results for the firsts of the three test occasions are also shown. Our strategy when deciding what PPTs were to be considered reliable in Study III was that the repeatability figure should be below or at a value representing 15% of the possible range in at least two of the three samples examined. When re-considering the repeatability figures as represented by 2.77 x within-subject SD, the step-on-stool test was not considered reliable, and neither was the PILE lumbar test.

For the neck flexor test, a systematic difference between test occasions was revealed in the 'Patients' sample, i.e. the third test value was best. For the PILE lumbar test, there was likewise a systematic difference in the 'Patients' sample, i.e. the test performance was best at the first test occasion.

Table 7. Repeatability coefficients and intraclass correlation coefficients for 24 persons with long-term spinal pain performing the physical performance tests at three test occasions in one week. Median (inter-quartile range) is given for the first test occasion.

	Test performance – median (IQR)	Value representing 15 % of possible range	Repeatability coefficient (2.77 x within-subject SD)	Intraclass correlation coefficients (95 % C.I.), method 2,1
Åstrand test VO_2	32.2 (16.5) ml kg^{-1} min^{-1}	6.3 ml kg^{-1} min^{-1}	21.9 ml kg^{-1} min^{-1}	.89 (.78-.96)
Isometric endurance:				
neck flexors	28.5 (42) s	9 s	18.5 s	.90 (.80-.95)
neck extensors	97 (145) s	27 s	80.0 s	.83 (.70-.92)
trunk flexors	33.5 (44.5) s	13.5 s	36.5 s	.84 (.72-.92)
trunk extensors	67 (120) s	27 s	77.5 s	.82 (.68-.91)
Step-on-stool test:				
dominant leg	20 (21) steps	7.5 steps	9.0 steps	.95 (.90-.98)
non-dominant leg	16.5 (16.5) steps		19.5 steps	.76 (.59-.88)
Lumbar lifting test				
men	14.9 (10.12) kg	5.61 kg	5.5 kg	.91 (.83-.96)
women	8.1 (5.05) kg	2.91 kg	5.0 kg	.88 (.77-.94)
Cervical lifting test				
men	10.4 (10.13) kg	5.61 kg	5.5 kg	.94 (.90-.98)
women	5.9 (2.85) kg	2.91 kg	1.66 kg	.96 (.92-.98)
Gait test	32.5 (7) s	4.5 s	5.0 s	.91 (.83-.96)
Gait test with burden	35 (8) s	4.5 s	4.5 s	.95 (.90-.98)
Stair-climbing	23 (8) s	3.45 s	2.5 s	.97 (.93-.99)

Construct validity (Study II, Study IV)

Significant differences in test performance between 15 persons with spinal pain and 15 matched back-healthy persons were found for all tests but the trunk flexor test. In Figure 3, boxplots for the two groups are showed for two of the PPTs considered reliable.

High sensitivity and moderate-to-high specificity were found for all tests except the trunk flexor test and the Åstrand test.

Sensitivity and specificity for the stair-climbing test were high; for a cut-off value of .29 m/s, the sensitivity was .86 and the specificity .83.

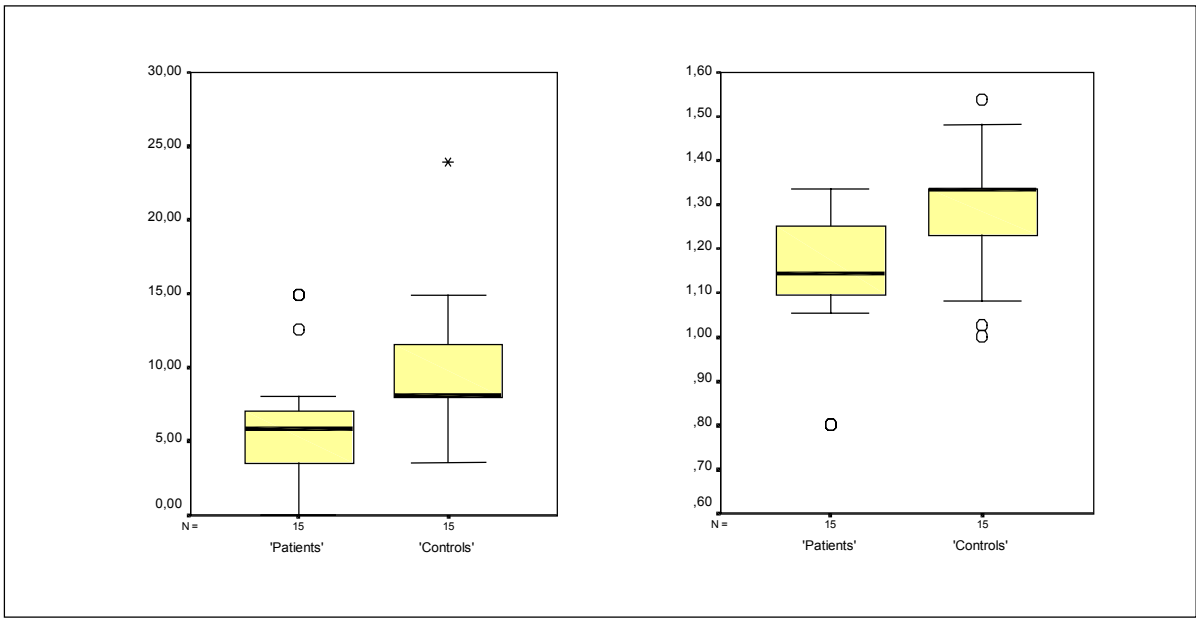

Figure 3. Results of the cervical lifting test in kg (left) and the gait test with burden in m/s (right) for the 15 person with spinal pain ('Patients') and the 15 back-healthy persons ('Controls').

Lumbar pain hampered the performance on PPTs more than neck pain (Table 8). Persons with lumbar pain as their main complaint (n = 107) performed significantly worse than persons with neck pain as their main complaint (n = 92) on all tests but the cervical lifting test.
The median value for the men with neck pain as their main complaint reached the normative value (derived from the cut-off values in Study II) for the step-on-stool test and all the gait tests, while the median value for the women with neck pain as their main complaint only reached the normative value for the stair-climbing test. At most 30% of those with neck pain as their main complaint reached the normative value in the lifting tests and in addition, for women, the gait test with burden.

Additional Mann Whitney U analyses. Persons with *neck pain only* (16 men and 18 women) had significantly better results on the stair-climbing test compared to persons with *lumbar pain only* (33 men and 23 women, both genders considered together. In the cervical lifting test, persons with *neck pain only* performed significantly worse than those with *lumbar pain only*.
Median values for persons with *neck pain only* were slightly higher for the step-on-stool test and the gait tests than those for persons with neck pain as their main complaint. On the lumbar lifting test, the persons with *neck pain only* showed a slightly lower capacity than persons with neck pain as their main complaint, whereas in the cervical lifting test, the women with *lumbar pain only* showed a slightly higher capacity than women with mainly lumbar pain.

High ratings of pain intensity on the CR10 Scale during PPTs hampered the results for most tests. High rated pain behaviour was likewise connected

with low PPT results. These lowered performances were particularly obvious for women. Subjects rating high perceived exertion on the RPE Scale during the gait tests performed less well, while men rating high on RPE during the lifting tests had a higher test performance (Table 8). According to additional analyses, CR10 ratings in general correlated slightly more highly to PPT results than RPE ratings, except for the gait test (Spearman correlations). Only at most 27 % of the variations in test performances were explained by the background factors analysed in regression analyses. Age, gender and neck pain as main complaint had most impact on test performances (n = 231).

Table 8. Variables significantly related to physical performance test results, as revealed in Mann Whitney U tests (Study IV). A = all, M = men, W = women. A minus sign is indicating a negative relationship, and a plus sign a positive relationship.

PPT	Neck main pain site			LBP main pain site			CR10 > 4			RPE ≥ 15			UAB ≥ 3			Pain > 4 (SF-36)			Exercise ≥ twice a week		
	A	M	W	A	M	W	A	M	W	A	M	W	A	M	W	A	M	W	A	M	W
Step-on-stool test				-			-	-	-				-	-	-			-	+		
Lumbar lifting test				-			-	-			+		-		-			-			
Cervical lifting test							-		-		+		-		-			-			
Gait test				-						-		-	-	-	-						
Gait test with burden				-	-		-		-	-		-	-	-	-						
Stair-climbing test				-		-	-		-	-		-	-	-	-						

Sensitivity to change / Responsiveness (Study V)

Sensitivity to change in absolute values according to the 'basic responsiveness statistical package' was revealed for the gait test with burden, the stair-climbing test and, for women, the cervical lifting test. For subjects with low initial performance, sensitivity to change was high for most tests in terms of absolute change. Moderate-to-high effect sizes were found for all the six examined PPT. For men, the lumbar lifting test, the gait test with burden and the step-on-stool test was responsive for improvements as well as deterioration, while for women, the gait test with burden and the stair-climbing test were responsive in both directions. According to effect sizes, the cervical lifting test was the most responsive test for improvements in men, while in women, the cervical lifting test, the gait test with burden and the stair-climbing test were likewise responsive for improvements.

When change was defined as at least 15 % of possible range, the value suggested to represent clinically important change, responsiveness was demonstrated for the cervical lifting test for men with low initial performance as a group.
The proportion of subjects with low initial performance was high among those who improved at least 15 % of possible range.

An overview over the clinimetric properties revealed for the PPTs is presented in Table 9.

Practicality

The practicality of assessments is defined as the usefulness of a test based on issues relating to personnel, time, equipment, cost of administration, and impact on the person taking the test (Task Force on Standards for Measurement in Physical Therapy 1991). These properties, shown for the present physical performance tests, are summarised in Table 10. The full range of possible test values was used in all tests.

Table 9. Overview of the clinimetric properties revealed for the physical performance tests. All PPTs were examined only for the first three columns.

PPT	ICC (2,1) 'Patients'	Intra-rater repeatability coefficients 'Patients'	Diff. patients - controls	Diff. neck-back pain	Related to ratings on CR10, RPE, UAB	Partly explained by background factors as listed (regressions)	Sensitivity to change[1]	Responsive-ness[2]
Åstrand test VO_2	89 (.78-.96)	22.2 ml $kg^{-1}min^{-1}$	–					
Isometric endurance tests; Neck flexors Neck extensors Trunk flexors Trunk extensors	 .90 (.80-.95) .83 (.70-.92) .84 (.72-.92) .82 (.68-.91)	 18.5 s 80.0 s 36.5 s 77.5 s	 X X – X					
Step-on-stool test; Dominant leg Non-dominant leg	 .95 (.90-.98) .76 (.59-.88)	 9.0 steps 19.5 steps	 X X	X	High CR10 – low performance High UAB- low test value	Age – Woman – Severe pain - Neck pain + Sick leave + Exercise +	Mod ES x 1 men	Large ES x 1 men
PILE lumbar test Max. weight	.91 (.83-.96)	5.5 / 5.0 kg	X	X	High CR10 – low test value men High RPE- high test value men High UAB- low test value	Age – Woman – Severe pain – Neck pain + >1 pain site +	Mod ES x 5	Large ES x 2
PILE cervical test Max. weight	.94 (.90-.98)	5.5 / 1.66 kg	X	X	High CR10 – low test value women High RPE- high test value men High UAB- low test value women	Woman – Neck pain – Age +/-	Base.stat. women Mod ES x 5	≥ 15% for 'least fit' men
Gait test	.91 (.83-.96)	5.0 s	X	X	High RPE – low performance High UAB – low performance	Age – Woman – Neck pain + Sick leave +	Mod ES x 2	Large ES women x 1
Gait test with burden	.95 (.90-.98)	4.5 s	X	X men	High CR10- low performance High RPE – low performance High UAB – low performance	Age – Severe pain - Neck pain +	Base.stat Mod ES x 8	
Stair-climbing	.97 (.93-.99)	2.5 s	X	X women	High CR10- low performance women High RPE – low performance High UAB – low performance	Age – >1 pain site – Neck Pain +	Base.stat Mod ES x 3 women	Large ES women x 1

[1] Sensitivity to change as shown by the 'basic statistical test package' (Base.stat.) or moderate effect sizes (mod ES)

[2] Responsiveness to clinically important change as shown by a change ≥ 15 % of possible range or large effect sizes (large ES)

Table 10. Overview for practicality of the physical performance tests. The proportion of persons who declined a test completely,
or presented floor or ceiling performance, is shown in the first three columns.

PPT	Decliners n=231 %	Floor effect n=231 %	Ceiling effect n= 231 %	Max. time taken min:s	Equipment needed	Space needed	Test leader skills
Åstrand test	2	13	2	15	Ergometer bicycle, heart rate monitoring, stop-watch	A space with a window, 2 x 2 m minimum	Medical education
Isometric endurance;					Stop-watch	1 x 2 m	Medical education
Neck flexor test	0	2	23	2	Myrin inclinometer Weight 0.5 kg		
Neck extensor test	0.5	0.5	37	4	Weight 1.5 / 2 kg		
Trunk flexor test	0.5	8	13	2.5			
Trunk extensor test	2	3	17	4	Angle table		
Step-on-stool test (dom/non-dominant leg)	2/4	4/2	25/20	5/5	Robust stool, height 44/40 cm	1 x 2 m	Common sense
PILE lumbar test	3	3	1	5	Robust shelves, box, weights, letter scale, pulse counter, stop-watch	1 x 2 m	Medical education
PILE cervical test	3	5	-	5	Same as for PILE lumbar test	1 x 2 m	Medical education
Gait test	0	2	-	1	Stop-watch, tape measure	A corridor, minimum 20 m	Common sense
Gait test with burden	6	7	-	1.5	+ carrier bags, weights 4+4+8 kg		
Stair-climbing	2	8	-	1	Stop-watch	Staircase, minimum 15 stairs	Common sense

General discussion

There is a need for a common basis for treatment recommendations and evaluations to establish clinical consensus among rehabilitation professionals regarding the evaluation and judgement of persons with long-term spinal pain. Thus, commonly accepted measurements are needed. Patient's expectations and beliefs are obviously important to include when performing treatment studies. There are indications that these expectations and beliefs could be more important than the treatment itself (Kalaoukalani et al 2001), thus they should be controlled for.
The often weak correlations between PPT and other measures of disability speak in favour of using PPT for supplementing self-ratings of artificial, and that the TP could be affected by the test situation in different ways. Many PPTs seem despite these possible drawbacks to possess important clinimetric properties, i.e. reliability, validity and sensitivity to change (Harding et al 1994, Simmonds et al 1998).

disability. Together with self-ratings of perceived pain and exertion, as well as the administering PT's ratings of pain behaviour, PPT could contribute to the overall understanding of a person with long-term spinal pain. Short-comings of questionnaires developed for measuring self-reported activity limitations are that the subjects possibly give the answers corresponding to a typical day, thereby referring to their use of different strategies for managing tasks of daily living. The answers also are affected by the subject's belief in his or her current physical function. Questionnaires and PPT may both be affected by motivation, and by familiarity with the formulations/test situation. Shortcomings of PPT are that they may not measure a 'real' capacity, since the test situation is

Questionnaires and PPTs alike depend on patient co-operation. When speaking of objectivity, neither questionnaires nor PPTs are objective, even if we often want to view PPTs as objective. Bohannon in 1989 defined 'objective measures' as those which depend not primarily on the judgement of the examiner. While we

can never be entirely objective in our measurements as human beings, we should seek to measure as objectively as possible.

Methodological discussion

Inter-professional judgements (Study I)

The questionnaire covering the need and potential for rehabilitation was not evaluated psychometrically. We wanted to investigate the use of professional judgement as in routine practice, and clinicians are confronted with these types of non-standardised question routinely when they are asked to judge the patient's work capacity and functional limitation for compensation purposes. Medical examination results including measurement of range of motion and the different measures of self-rated health were not obtained on the same day as the experts' ratings of need and potential, but within three weeks.

Reliability (Study III)

We did not prepare for good agreement or repeatability in advance by a certain training procedure, since we wanted the test situation to be as normal as could be. All the physiotherapists had a detailed manual, which we had discussed and demonstrated before each PT had begun to use the PPTs, but this was at least some months before the study period. In the inter-rater agreement study, the three PTs (including the author) were all involved in the testing procedure for the RCT study, while in the inter-rater repeatability study, two of the PTs had been using the PPTs in their clinical reality for some months. This indicates that the figures provided in our study were partly generalisable to a clinical situation.
When performing the inter-rater repeatability study, we decided to use the Borg RPE scale for regulating performance. TPs were asked on the first test occasion to discontinue the neck flexor test when they reached a perceived exertion level, which they felt as appropriate. The TP then rated the exertion level on the RPE scale. The other PPTs were then discontinued at the same RPE level. On subsequent test occasions, the TP was asked to discontinue the PPT at the same RPE level as the one chosen on the first occasion. This approach has been described as useful for regulating exercise levels (Dunbar 1993), and was used as an attempt as standardisation of effort between test occasions, after our experience that TPs performance motivation could differ somewhat between test occasions (in the intra-rater repeatability study).
The high proportions reaching 'best possible' test values, i.e. the point when the test was terminated by the PT, revealed for some tests could have biased the results (Table 10). It was considered natural to include all participants in the analyses, since the test performances in a clinical test situation would be discontinued at a certain time. Be this as it may, the tests shown to have highest proportions of 'best possible' test values were the isometric endurance tests, which all were considered to have too high variability between test

occasions, independently of the high proportion of 'best possible' values. In the repeatability study, we chose to report the size of the differences within the same subject between test occasions, for an easier interpretation of the figures. These figures were in this summary complemented by the $ICC_{2,1}$ for the 'Patients' sample. ICC show the proportion of between-subject variation for the measurements, i.e. the larger the ICC, the less the variation within the same subject, which is the information we are interested in. However, the ICC alone does not give enough information for interpreting the clinical use of the respective PPT in individual follow-ups – then the size of the differences, here shown as 2.77 x within-subject SD, should be judged too (Bland and Altman 1996). This repeatability coefficient can serve as a base for evaluating a clinically significant change in performance.

Construct validity (Study II and IV)

When discussing validity, logic is a key concept. Statistical analyses help in revealing indications of evidence, but they can never replace our reasoning in deciding what constitutes validity in the particular situation, for the particular population. The construct validity of a test can scarcely ever be said to be thoroughly investigated. There will always be more angles of approach. It is said to be important that hypotheses are made up in advance, before conducting the study. On the other hand, we probably have something to learn from the results of the studies. In Study IV, for example, we found only minor evidence of poor test results on the PPTs for persons with more than one pain site. This finding suggests a reconsideration of our hypothesis that persons with multiple pain sites would have low PPT results.

Sensitivity to change / responsiveness (Study V)

The major problem when examining sensitivity to change or responsiveness is the lack of 'gold standard' for assessing whether a change has really taken place (Lurie 2001). Williams and Myers asked eight persons with LBP about what "recovery" meant to them. These persons said "Getting back to the way I was before the injury" or "Getting back to normal" (Williams and Myers 1998). A question where the satisfaction with the state of 'back to normal' was to be rated seems like a good suggestion for an outcome measure.

The self-rated concepts used in this thesis as the 'standards' against which changes in the PPTs were compared were considered likely to improve from the rehabilitation. The concepts turned out to be responsive to differing degrees; 82% of the subjects participating in the six-month follow-up rated their general health improved compared to the previous year, whereas only 42% rated their self-efficacy in coping with pain as better compared to the rate on inclusion. The proportion that rated the 'disturbance from pain at work' as decreased was in-between.

In our study, figures representing 15 % of the range of test values possibly obtained for a particular test

represented 'clinically important change'. These figures, derived from Study III, represented the value below which the difference between two measurements would lie with 95 % probability (Bland 2000). Vandermeulen and co-workers suggest that an increase of 14% (of the initial performance, author's comment)"is likely approaching the level of minimal clinically important change" (Vandermeuelen et al 2000, p.52). The appropriateness of our choosing an absolute value instead of a proportion of improvement can be discussed. The choice of an absolute value for 'clinically important change' seems more relevant than a percentage of initial performance, since the latter would in some cases be very low,while in other cases it would be very high and hardly achievable. A person with a low capacity would have good chances to achieve an 'improvement', while these chances for a person with a high initial capacity would be low. Since the improvements in PPTs seem to be greater for the 'least fit', this approach seems hazardous.
The figure of 15% was arbitrarily chosen to get a clinically useful common repeatability measure for the three samples in Study III, and should not be interpreted as definitive.

Effect sizes can be calculated in different ways. The original calculation involves the mean of the change divided by the SD of the baseline data (Kazis 1989). Another strategy is to divide the mean change for subjects 'improved' by the SD of the change for the 'stable' individuals. These effect sizes should then be compared with those obtained when dividing the mean change for subjects who had worsened on the outcome measure with the SD of the 'stable' subjects. This strategy is used in this thesis and is referred to as 'the Guyatt responsiveness statistic' by Deyo et al 1991. It has the advantage of taking account of the change that often takes place even in 'stable' subjects.

Internal – external validity

Internal validity refers to the relationship between the dependent and the independent variables, while external validity is concerned with generalisation from the sample to the population of interest (Payton 1988). The threats to internal and external validity which are of current interest for the PPTs in this thesis will be discussed briefly.

Threats to internal validity

Learning effects. Precautions against learning effects affecting PPT results were made in different ways: when the step-on-stool test was introduced to the TP, he or she was allowed to try for a couple of steps due to the unfamiliar technique. Likewise, the TP was allowed to try the technique for the PILE tests once before the test started. These precautions were necessary for correct performance, and for the PILE tests it was important for the TP to find his or her own technique of foot placing and proper distance to the shelves. It was likewise considered valuable to convey a feeling of the load before starting "for real". Even in the gait test with burden, the TP was allowed to lift the

carrier bags to get this load feeling before the stop-watch was started. The gait test and the stair-climbing test were started directly, without precautions other than the simple explanation of the tests, due to the familiarity of the tasks.
When considering the unfamiliarity of the muscular endurance tests, TPs should have been allowed to try all these positions and their respective loading once before the stop-watch was started. Only for the neck flexor test, however, was a learning effect evident (Study III).

Calibration precautions considering loading were regularly done for the Åstrand test and the PILE tests. Height of the stair treads and the total staircase height were measured at the different locations where the tests were performed, and the time taken was then recalculated as m/s in Study IV. In Study V, all m/s measures were standardised too, according to the total height of one of the staircases used.
When it comes to the "calibration" of the PT administrating the tests, there naturally occurred situations which could be biased in different ways. The instructions preceding the tests were standardised according to the manual. Two corrections of technique or speed were allowed before the test was interrupted. The judgement of when such a correction should be done, however, was up to the PT. In the PILE tests, e.g., the heart rate was monitored by a measuring device, but the PT was also to take note of the time taken and to put more weights in the box each 20th second. The judgement of when to stop a test due to lowered speed differed between PTs once in Study III. Such discrepancies in judgement could be easily corrected by regularly exercising simultaneous testing.
Other differences in measurements occurring in Study III were due to practical difficulties with the stop-watch – in the muscular endurance test for neck extensors, the test leader was occupied by placing the weight on the TPs head and thus was one or two seconds late with the stop-watch, while in the stair-climbing test, one of the PTs started the stop-watch before the TP was touching the first step instead of at the moment of touching. In the step-on-stool test, one PT simply lost count in a moment of distraction. Such minor calibration problems within raters were easily revealed in the inter-rater agreement situation with simultaneous testing, but in the normal situation, with one single rater administrating the tests, such calibration problems are hidden.

Regression towards the mean could well have occurred in Study V, as improvements were more obvious in the subjects who had most difficulties performing the tests on inclusion testing. The pattern, though, differed somewhat between PPTs, indicating that supplementary explanations of the improvements may be present.

Bias. A selection bias of subjects could be considered in that the participants in the RCT study were all recruited from a database covering mostly blue-collar workers. The participants were seldom subject to any other

rehabilitation interventions, and their current sick leave period had lasted at most six months. This could be interpreted as if they were not to be considered as having long-term spinal pain. When considering the duration of their current pain condition (median value 9 months), though, the classification 'long-term spinal pain' seems valid. The fact that the RCT participants were recruited from a data base independently could alternatively be seen as a surety for non-bias, the subjects not having sought care at a certain hospital or from a certain physician.
In Studies II and III, samples of convenience were included; consecutive subjects referred to rehabilitation clinics, subjects included as pilot subjects in the RCT study, and staff from rehabilitation clinics and adjacent companies. The matched subjects in Study II, however, were identified by nurses at an occupational health service centre nearby.
The administrating PT knew for practical reasons to which group the participants belonged in Studies II-III. This could somewhat bias PPT results, but the PTs, on the other hand, had no access to previous test results. In Studies IV-V, the PT was blind to group membership. The UAB ratings, however, could theoretically be somewhat biased by the PTs' knowledge of the current PPT results.

Attrition rates differed between the rehabilitation groups and the control group in Study V, but since these groups were not maintained in the analyses, this was considered to be of minor importance.

Threats to external validity to consider in this thesis are:

Contamination of test performance by other tests. Test performance could e.g. be affected by pain occurring during a former test. The TPs were explicitly told that they themselves had the responsibility for choosing the endpoint for the tests, and that they could decline a test completely at any time. This precaution might to some extent have prevented such contamination. Another precaution was built in by the fixed order of the PPTs. In Study III, the within-subject variability was higher for the non-dominant than for the dominant leg. As the dominant leg always was tested first, it is possible that fatigue affected the performance for the non-dominant, usually the left, leg.

The ***sample size*** was small in Studies II and III. The small samples are a possible flaw-back for generalisations from Study II. To some degree though, the matching procedure balance the small sample sizes. The differences between persons of the same age, the same sex and the same or a similar occupation are likely to be smaller than differences between "whoever".
In studies examining large study samples, small changes or differences can be detected, but the clinical relevance of these findings may be low. Using large samples may lead to a false security, believing that we can

interpret all individuals like we do groups described in research reports. The within-subject variability in Study III could be just as large in a larger sample. The small sample sizes in this study can, in the author's opinion, be regarded as a strength rather than a weakness, when considering the fact that in the clinic, you typically do not meet groups of people, you meet individuals, who most likely will vary just as much as our participants did.

Sick listing was an attribute only for the 'patients' in Studies II-III. The possibility of sick-listing influencing test performances cannot be excluded. Being at work means that you do 'the usual things', and use your body in a usual way. Even if your work is sedentary, sick-listing might somewhat lower everyday physical performance. On the other hand, indications of better PPT results for subjects who had been on sick leave more than three monhts compared to others were revealed in Study IV. A possible explanation is that after some time of sick leave, pain and disability level is likely to decrease spontaneously, leading to better PPT results.

Discussion of results

Inter-professional judgements (Study I)

Given the extremely low agreement between different professionals, there seems to be a need for a common basis for evaluation of rehabilitation needs. Hansson and co-workers stresses that different health care professionals have different perspectives, and thus differs in their judgements concerning patients (Hansson et al 2001). Our findings underline a seeming injustice in the judgement process, where indetected priorities and chance rule whether a person is to get rehabilitation. Waddell and co-workers found that persons who showed a large amount of 'inappropriate illness behaviour' (according to pain drawings and Waddell score) had received significantly more treatment than others (Waddell et al 1984). In Sweden, only a minor part, 15-20%, of persons on sick leave for at least three months in 1997 were offered participation in a rehabilitation programme (Jensen 1998, Selander et al 1998). This was despite the fact that not all money budgeted for rehabilitation was used.
A guideline for procedures in the examination and judgement of spinal pain is needed in routine clinical practice. In a study by Binkley and co-workers, low agreement between Canadian physical therapists was revealed for identification of clinical findings related to six of 25 diagnostic classes for LBP patients. One diagnostic class for which agreement was low was 'chronic pain syndrome', a finding that underlines the difficulties associated with judgements of long-term pain conditions. The authors conclude that a standardised system of classification should exhibit a) consistent terminology, b) exclusive categories representing distinct, recognisable clinical conditions, and c) categories specific enough to guide

patient management (Binkley et al 1993). Such a standardised classification system would serve as a basis for clinical decision-making, and would enhance communication among health care professionals, as well as between professionals and payers. In addition, it would be helpful in identifying homogenous subgroups for intervention studies.

The relevance of the patient's owns beliefs in prediction of self-ratings on SF-36 and degree of sick leave in our study was striking. In a recent study, Kalauokalani and co-workers reported that among persons with high expectations of a certain treatment, a higher proportion improved than among persons with low expectations (Kalauokalani et al 2001). Patients' expectations and beliefs are obviously important to include in treatment studies. There are, however, ethical problems to consider when knowledge of predictive factors for outcome increases: should we put all rehabilitation efforts into those most likely to improve from such interventions? And what would be the criteria for 'improvement', to say nothing of - whose perspective? And, lastly, what would become of all the others? We must remember that many factors benefit the participants through a rehabilitation process, many of which will scarcely be revealed in research studies.

Reliability (Study III)

The inter-rater agreement between two simultaneous PTs was lower than expected for three tests; the step-on-stool, the neck extensor test and stair-climbing. These low agreement figures could be understood by the circumstances discussed in the 'Methodological discussion' section, and thus could be adjusted easily.

High variability between test occasions was expected due to the varying nature of spinal disorders often demonstrated in the clinic. Interestingly, the variation turned out to be about as large in our 'back-healthy' persons.

Overall, the inter-rater repeatability could be considered good in back-healthy persons, as differences between test occasions were typically no larger when the tests were run by three different PTs than when administered by one, with exception of the Åstrand test and the stair-climbing test. Rather, the variability was *lower* in the inter-rater repeatability study, probably due to our fixed RPE-strategy described in the 'Methodological discussion' section.

The variability between test occasions was large for the Åstrand test. Some earlier studies have suggested high reliability of different ergometry tests (Cox et al 1989, Becque et al 1993), while others have been cautious (Armstrong and Costill 1985, Lockwood et al 1997). Many factors in our study settings could have influenced test results. One contributing factor to the high variability might be our choice of a minimal steady-state heart rate of 120. A minimal steady-state heart rate of 130 is proposed by Åstrand and

Rodahl (1987) as the most appropriate value. The influence of different confounding factors is considered to be lower when choosing this work heart rate. The reason for choosing 120 beats per minute was our clinical experience that persons with long-term spinal pain often have difficulties managing excessive exertion, an assumption confirmed by the high proportion of subjects not being able to reach a test result (Table 10).

The high variability shown for the muscular endurance tests is not surprising. These tests are not representing familiar activities, which have immediate face validity for the participants. Some of the tests, for example the back extensor endurance test, can feel somewhat uncomfortable even for a back-healthy person for different reasons. Also contributing to the endurance time performed are factors as pain tolerance, competitiveness and boredom. These factors could have large influence when only one attempt is recorded. For endurance tests such as these, the best of two attempts perhaps would be preferred. Mannion and co-workers allowed their TPs to become familiar with their tests on a separate day before testing, and to allow "sufficient trials" when testing in order to increase the chances of eliciting the best voluntary effort (Mannion et al 2001).
Moreland and co-workers reported low ICC and high within-subject SD for tests of isometric back extensor endurance and abdominal endurance (Moreland et al 1997). On the other hand, other authors have considered the reliability of similar tests as acceptable. Moffroid and co-workers found that the modified Sörensen test was reliable in persons with long-term LBP who considered themselves physically active (Moffroid et al 1994). Ito and co-workers found both a modified Sörensen test and an isometric curl-up test to be reliable in persons with long-term LBP as well as 'healthy' persons, referring to a high ICC and to Pearson correlation coefficients (Ito et al 1996). Dedering and co-workers found the modified Sörensen test reliable in back-healthy persons from several aspects. The within-subject SD for endurance time was 28.2 s (Dedering et al 2000), thus revealing a repeatability coefficient of 2.77 x 28.2 = 78.1 s, a figure comparable to our repeatability coefficient for persons with spinal pain displayed in this summary. Alaranta and co-workers (Alaranta et al 1994) reported that the SD of the mean difference for two physiotherapists recordings was 42 seconds, thus meaning that 95 % of all differences would lie in between +/- 84 seconds (Bland and Altman 1986). The authors concluded that the reliability coefficient was 'fairly good'. Notably, there are still no generally accepted rules for what methods to use in reliability studies, nor what constitutes 'good reliability', thus leaving this to each author to decide.
The stability desired for a measure may depend on the purpose. It seems important to state clearly the context in which the measure is intended to be used, and to explain what

interpretations can be made from the study results.
Differences in reliability results between studies can, apart from differences in statistical methods and interpretations, be due to different test set-ups, different equipment and different verbal instructions. The latter was obvious in our study, where the inter-rater repeatability figures, as mentioned above, in some cases turned out to be better than the intra-rater figures. An approach with a predefined end-point, chosen by the TP, e.g. at a time point when the pain intensity reach '5', i.e. 'strong' on the CR10 Scale, has been suggested by Dedering and co-workers (Dedering et al 1999) for people with painful conditions, thus avoiding difficulties in distinguishing between muscle fatigue and mental fatigue due to pain. Some of our participants had difficulties with such distinctions, as did those in an earlier study (Schmidt 1985).

The PILE cervical test showed low variability between test occasions. One probable contributory factor is that this test is demanding with regard to arm muscle strength and endurance, and therefore the range of test results was narrow. The test soon became physically demanding when more weight was added.

Despite the high variability for the Åstrand test and the PILE lumbar test, these tests are arguably valuable in a screening procedure, when an overall evaluation of the disability level of a person with long-term spinal pain is of interest. Some authors have proposed that tests showing high variability should be performed at least twice, and that the mean should be recorded (Simmonds et al 1998). This approach cannot be recommended on the same test occasion for the Åstrand test, nor for the PILE lumbar test, due to the high overall exertion levels connected with them. On the other hand, it could be valuable to use these tests if there was a possibility of repeating the tests on at least two separate days. The PILE lumbar test results were best on the first test occasion, indicating some adjustment due to e.g. impairments such as fear of pain or avoidance of excessive exertion. Lockwood and co-workers found a learning effect of the Åstrand test in their study (Lockwood et al 1997). No such effect was revealed in our study, however, perhaps due to the small samples. General cardiovascular fitness is essential for the overall activity level (Moffroid 1997). The PILE lumbar test is informative for observing the patient in whole-body motion: body awareness, fear avoidance behaviour, and working techniques are examples of body functions and activities revealed.

Construct validity (Studies II and IV)

All PPTs except the trunk flexor test showed discriminative ability between persons with spinal pain and back-healthy persons (Study II). The reasons for this may include the following;

- The persons with spinal pain (SPP) could exhibit impairments and activity limitations not present

in the back-healthy persons (BHP). Several other authors likewise have found significant differences in physical performance between persons with long-term spinal pain and back-healthy persons; e.g. Mayer and co-workers comparing 100 persons with LBP to 92 industrial workers performing the PILE tests (Mayer et al 1988 a + b); Ito and co-workers comparing isometric muscular endurance in trunk flexion and extension in 100 LBP patients and 90 'healthy' persons (Ito et al 1996), and Simmonds comparing 44 persons with LBP and 48 pain-free controls using different PPTs (Simmonds et al 1998).

- Intra-individual factors, such as pain, fear-avoidance behaviour, or fear of failure, in the SPP could have affected the performances negatively. Pain is naturally a part of the problem, but the causal relationship between disability and pain is still unclear. Fear-avoidance behaviour includes termination of physical performance tasks beforehand, 'just in case'. Fear of failure in persons with long-term LBP was shown by Schmidt to affect physical performance negatively (Schmidt 1985). In our Study IV, persons with high self-rated pain intensity as well as high rated pain behaviour performed worse in the PPTs.
- The SPP could have had lower performance due to sick-listing. As earlier stated, we cannot exclude the possibility of sick-listing being one of the factors influencing test performance negatively. There could well be a correlation between sick-listing and a more pronounced disability, thus making lower PPT results in the 'Patient' group logical.
- Our study persons were not representative of the populations SPP and BHP, respectively. The small sample sizes are, as mentioned earlier, considered to be partly balanced by the matching procedure.The circumstance that a large majority was female is of course a flaw, but on the other hand more women then men seek care for long-term spinal pain (Unruh 1996).

The cut-off points for sensitivity and specificity are likely to be revised if the calculations are replicated for a larger sample.
We found no clinically meaningful distinction between persons with long-term spinal pain and back-healthy persons for the Åstrand test (Study II). The latter was included in part because we assumed that persons with long-term spinal pain would have lower fitness levels. This has also been assumed by other authors, e.g. McQuade and co-workers 1988. In several studies, though, evidence showing that persons with chronic low back pain have fitness levels comparable to others has been revealed (Schmidt 1985, Wittink et al 2000). This is perhaps not primarily evidence of the absence of decreased fitness for persons with long-term spinal pain, but rather evidence of poorer cardiovascular fitness for many people in the Western world. A test of overall fitness level is a good opening for general discussions with the

persons tested about the impact of pain on physical activity, lifestyle habits, and body weight. The recognition of these factors is important for the individual in the long-term effects of rehabilitation. Increasing cardiovascular capacity also, certainly, has beneficial effects for persons with long-term spinal pain, just as for others, thus making a baseline measurement adequate.

In the regression analyses, at most 27 % of the variance in test results was explained by the background factors included in the model. Thus most of the variance remained unexplained. Many factors influence physical performance, some listed in Figure 4. PPTs show the subject's performance ability in the relevant test on the relevant occasion. Crombez and co-workers showed that high expectations of increased pain intensity and high ratings of pain-related fear predicted poor physical performance in trunk flexion-extension in a Cybex machine for persons with long-term LBP (Crombez et al 1999). Lackner and co-workers showed that 'functional self-efficacy' beliefs together with gender and average pain intensity during the previous week explained up to as much as 63% of the variation in PPTs, similar to the PILE tests and the gait test with burden, for persons with long-term LBP (Lackner et al 1996). Grönblad and co-workers found differing correlations between PPTs and self-reported disability and pain between gender (Grönblad et al 1997). On our results, PPT results were affected by e.g. age, gender, pain site, rated pain intensity and perceived exertion during testing; rated pain behaviour and rated pain intensity during the previous 4 weeks. We did not examine the impact of self-efficacy of any kind, or of pain expectations; but these concepts could be among the possible factors contributing to the unexplained proportion of the variance in PPT results, as revealed in the regression analyses. Rated pain intensity, perceived exertion and pain behaviour were not included in the regression models, as being variables related to the PPTs. These measures of impairments and activity limitations are likely to explain some of the remaining variance. It seems natural to suggest that other impairments and activity limitations within the individual contribute to the variance, too, e.g. muscular deficits and problems with performing activities .
In sports, the view of physical performance as a multifactorial concept is established (Wormgoor and Björholt 1994). It is generally accepted that physical capacity, as well as constitutional, psychomotor and psychological factors are all important for outcome. Perhaps we as physiotherapists should stop believing that we can measure explicit 'physical performance'.

The findings that persons with LBP performed worse than persons with neck pain accorded with those in a study by Jette and Jette 1996, where persons with LBP rated their physical function on SF-36 as worse than persons with neck pain did. In a recent study, persons with LBP rated more fear of work-related activities

than persons with neck pain did, both groups being exposed to work injuries (George et al 2001). In a study by Kjellman and co-workers, though, persons who had been sick-listed for neck disorders had more complaints than persons sicklisted for LBP, as rated on a questionnaire at a twelve-year follow-up (Kjellman et al 2001). As these studies all concerned self-rated complaints and beliefs, comparisons with PPTs seem unsure. The moderate differences between persons with *neck pain only* or *lumbar pain only* could be due to the small numbers in each group. These analyses were done for checking tendencies, and should be interpreted with cautiousness. The tendencies, however, were in the expected direction, that is, persons with *lumbar pain only* had lower performance on most PPTs. The main pain site, thus, seemed to be adequate for predicting the PPT results. Interestingly, the median values for persons with *neck pain only,* as those for all the gait tests, were higher for the gait test with burden than for persons with neck pain as their main complaint. This finding appears to refute our hypothesis that the gait test with burden affects persons with neck pain. The actual state of the neck pain might be of importance – neck pain in an acute phase may be more disabling than long-term neck pain. In the author's clinical experience, persons with lumbar pain manage the cervical lifting test much better than they do the lumbar lifting test, while persons with neck pain have problems with both levels alike. These observations were partly confirmed; persons with lumbar pain as their main complaint performed better on the cervical lifting test than those with neck pain as their main complaint, and for those with *lumbar pain only*, the difference was significant. The median value was also higher for women with *lumbar pain only* on the cervical lifting test.

Former studies examining differences in PPT results between persons with LBP and neck pain have been difficult to find.

Sensitivity to change / Responsiveness (Study V)

In general, the number of subjects who improved 15% of possible range was lower for the PPTs than for the self-rated outcome measures examined. At most 40% of the subjects (women in the stair-climbing test) who possibly could improve did so during the 6-month period. The PPT showing most sensitivity to change in the sense of 'most frequently showing better result' were the step-on-stool test, the gait test with burden and the stair-climbing test.

Indications of improvements in working technique after a rehabilitation intervention have been reported by Haldorsen and co-workers (Haldorsen et al 1998) and by Magnusson 1992. This is an interesting point of view for further studies. The quality of movement could well be of interest, especially in a potentially pain-provoking task. In a study by Piela and co-workers, persons with LBP had low ability to predict their lifting capacity on a subsequent lifting test (Piela et al

1996). This suggests the inclusion of a PPT for evaluating lifting ability in persons who are obliged to lift in their daily life.
In a study of lifting techniques in elderly persons, those who had significantly lower muscular strength in the knee and hip extensors used a lifting technique loading primarily the back (Puniello et al 2001). This indicates that it might be important to evaluate strength and endurance in the lower extremities to prevent potential pain-provoking lifting. The step-on-stool test, however, cannot be recommended for these purposes due to high variability between test occasions and only minor sensitivity to change.

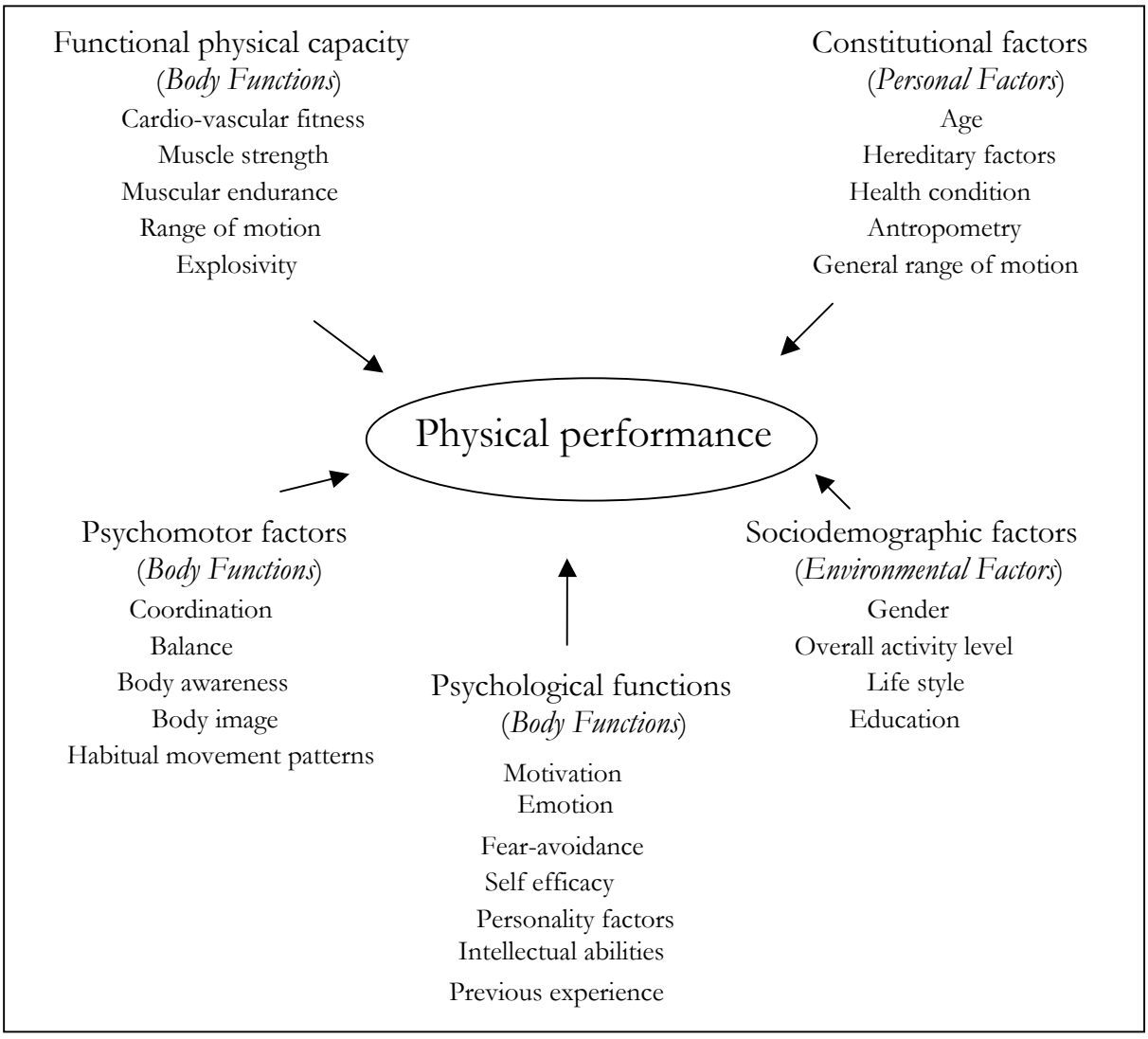

Figure 4. Factors affecting physical performance. Modified after Wormgoor and Björholt 1994. Terms from the ICF Classification (WHO 2001) are added to illustrate the multiplicity covered by the physical performance tests. Impairments are problems with body functions, e.g. muscular endurance, coordination or motivation. Activity limitations are problems such as fear avoidance behaviour.

Most physical performance improvements that could be related to the outcome measures appeared at the six-month follow-up. This suggests that a subjective experience of general health, pain intensity and self-efficacy, as rated by questionnaires, might be easier to change, while changes in physical performance tests requires more time to achieve. It could well be that attitudes are more easily changed than 'real' behaviour (Åberg 1984). In questionnaires, subjects can relate to their overall perception over at least a couple of days, in their own environment with adaptations and possible support; but in a physical performance test, what counts is specific performance on that very occasion, in that particular spot.

Manniche and co-workers concluded that a training period of at least two to three months is necessary for improvement in physical performance for persons with long-term spinal pain (Manniche et al 1988, Randlöv et al 1998). In other studies reporting large improvements in physical performance, the intervention was 12 weeks (Kuukkanen et al 1996, Ljungkvist 2000). The rehabilitation programmes evaluated in our RCT were not exclusively designed for improving physical performance, but rather aimed at encouraging overall healthy behaviour, training of different coping strategies including physical exercise, and convincing the patient of his or her own capability and responsibility for health. Moreover, the programmes lasted only for four weeks. These factors could contribute to the modest improvements in the PPTs. However, large improvements have been achieved by other investigators, examining interventions of comparable length (Alaranta et al 1994, Williams et al 1996, Strand and Moe-Nilsson 2001).
For those improved on the PPTs, though, several effect sizes were moderate or high at the six-month follow-up. According to the effect sizes, all our six PPT's examined showed sensitivity to change to differing degrees. The effect sizes were in most cases in the expected direction, i.e. positive for subjects rating improvement on the outcome measures, and negative for subjects rating deterioration. The PILE lumbar test seemed to be sensitive both to improvements and deterioration for men, but only to deterioration for women. The PILE cervical test was, according to effect sizes, only responsive to improvements.

The lack of a more exact classification than 'long-term spinal pain' of participants in the RCT could contribute to relatively few improving on the PPT results. The higher

sensitivity to change for subjects with poorer results on the PPTs at inclusion supports this. The PPTs may be most responsive for subgroups of patients; as indicated in Study IV, the PILE cervical test had high ability to detect impairments and activity limitations in persons with neck pain from both genders. The PILE lumbar test had high ability to detect lumbar problems for women, but for men with lumbar pain, the PILE cervical test was still the most detective test.

The moderate responsiveness shown for these PPTs for persons with long-term spinal pain as a group over a six-month period could be explained partly by the fact that these measures were not chosen for their individual clinical implications, but rather were administered as a package to all the individuals irrespective of their stated problem areas. This assumption is supported by the fact that among the subjects who had most problems in performing the tests, more persons were improved. Functional disabilities expressed by the patient should guide the choice of test.

The high attrition rate, especially in the control group, for the PPTs was a drawback for the power to detect degree of responsiveness. The reasons for non-compliance in the follow-up PPTs were lack of time, transport inconvenience, difficulties in getting time off work, and personal or motivational problems. The latter were more likely in the control group, where the subjects had had little attention from the project group. There is, though, no reason to believe that the controls were in worse shape and therefore did not attend at follow-ups, since they were comparable to the rehabilitation subjects on inclusion with regard to sick leave, pain duration, ratings on SF-36 and PPT results. The relatively larger improvement rates at the six-month follow-up might conceivably be because many control group subjects were absent. This seems, however, unlikely, as the rehabilitation subjects were not more improved, as rated on the selected outcome measures, than the controls were. Other authors in Scandinavia have reported high attrition rates in the control group. Thus Haldorsen and co-workers reported that only 60 % of their control group turned up to the 12-month follow-up (Haldorsen et al 1998). From England, Williams and co-workers reported 30 % attrition rates in the control group at the one-year follow-up (Williams et al 1996).

The gait tests were sensitive to change in a number of ways, but showed no clinically important changes according to our definition. Large effect sizes, though, were found for the gait test and the stair-climbing test in women rating deteriorated general health. Maybe a longer distance would be more responsive for persons with long-term spinal pain. A modification of the Bag and Carry test, where time and maximum weight carried were both recorded, could be worthwhile testing for persons with long-term spinal pain (Noonan and Dean 2000).

In several studies, the weak correlations between PPTs and other measures, such as pain, self-rated

disability and return to work have been revealed. Grönblad and co-workers note that the relationship between perceived disability and physical status are unclear for persons with LBP (Grönblad et al 1997). Moderate correlations between self-reported activity limitation and PPTs in persons with LBP was shown by Lee and co-workers. The author´s conclusion was that each method by their respective perspective appears useful for understanding activity limitations in LBP (Lee et al 2001). This inconsistent relationship makes the comparison between changes in PPTs and changes in the chosen self-rated opinions hazardous. Conclusions concerning the apparently low responsiveness on the PPTs should be drawn with some caution.

Practicality

Floor and ceiling effects were common for the muscular endurance tests, suggesting that they did not adequately capture the range of performance. For the neck extensor test, 2% declined to perform the test at all, 1% tried but failed to get a test result, and as high a proportion as 37% managed to reach the maximum limit: 180 s. Such a high ceiling effect is of course unacceptable, but even for the other muscular endurance tests, the ceiling effects were high. Swiontkowski and co-workers suggested a limit of maximum 5% floor and ceiling effects for their examined questionnaire to be considered useful (Swiontkoski et al 1999). Others have suggested as much as 20% as a relevant limit for clinical tests (Klässbo, unpublished data). Forty-three percent stopped the PILE cervical test because of pain and 14 % because of fatigue. The corresponding figures for the PILE lumbar test were 48% and 11%, respectively. Five percent declined the PILE cervical test because of pain or fear of pain, while the corresponding figure for the PILE lumbar tests was 4 %. Thus, the differences between the two PILE tests were not as large as expected considering the differences in repeatability coefficients (Study III). When considering all PPTs, the proportion of decliners were among the highest for the PILE tests. High rates of attrition for lifting tests have been described also by Fishbain and co-workers, who state that “a significant percentage” refused to attempt lifting even minor weights (Fishbain et al 1994). The reason for declining was in our study pain or fear of pain, the TP referring to former painful experience of lifting, often stressing that the pain would come delayed after the lifting task.

Gender differences

When assembling the test package, precautions were taken concerning the expected differences in performance between men and women. In several tests, but not all, men and women had differing test loading; the neck extensor endurance, the step-on-stool test, the PILE tests, and the gait test with burden. In the Åstrand test, the load was always guided by the TP’s heart rate, and the achieved values were interpreted according to the sex-

specific nomogram described by Åstrand and Rhyming 1956. In the step-on-stool test, the men seemed to perform better than the women in spite of the higher stool used for them, indicating that the height might still be too low to be discriminating for men. The proportion of men performing 'best possible test value' on the step-on-stool test was 31%, versus 20% of women, indicating that the step height should have been slightly higher for both genders. The same pattern was obvious for the trunk flexor test, and, in particular, for the neck flexor test, where 44% of the men reached the 'ceiling', but only 5% of the women. These latter tests should no doubt have been designed differently between genders. In the neck extensor test, 47% of the men and 27% of the women reached the 'ceiling' value, indicating that the chosen loading and/or discontinuing time set was too small for both genders.

Men are generally stronger than women, differences varying according to muscle group, equipment and sample. Explanations have been suggested to be found in larger overall body size, in larger muscle fibre sizes (Miller et al 1993), and larger body mass (Bäckman et al 1995) for men in general. Fothergill and co-workers found that men were stronger than women in dynamic lifting, but not in static (Fothergill et al 1996). Barnekow-Bergkvist and co-workers found that men were stronger than women in two-hand lift, hand grip and sargent jump both at the age of 16 and 34 years. Men also had higher muscular endurance in dynamic bench-press and curl-ups at both ages, but in the back extensor muscle endurance only at the age of 16 (Barnekow-Bergkvist et al 1996). Neither Moffroid and co-workers nor Dedering and co-workers found any significant differences between the endurance times for back extensors of adult men and women (Moffroid et al 1994, Dedering et al 1999). Sunnerhagen and co-workers found that self-selected walking speed (30 m) was comparable between genders (Sunnerhagen et al 2000), as did Ralston (1958).

The proportion improvers was higher among the men than among the women in Study V. This was particularly obvious in the PILE tests; 28% of the men but only 10% (lumbar test)/8% (cervical test) of the women who possibly could improve in the tests did so. Only in the stair-climbing test at the 5-week measurement did the women improve proportionally more often. This pattern is interesting, since the improvements according to SF-36 were earlier revealed to be most prominent for women participating in the rehabilitation groups in the RCT (Jensen et al 2001). In the outcome measures used in Study V, however, no significant differences in improvement rate were seen between gender.

Differences between genders were also revealed in Study IV, considering e.g. the relations to the ratings on CR10, RPE and UAB for PPT results.

It is obviously important to analyse PPT results for men and women separately, as otherwise important

information will not be revealed, being hidden by the common analyses.

Age differences

In Study IV, the regression analyses revealed that age had impact on test performance in all tests. Test performance decreased with higher age for all tests but for the PILE cervical test, where subjects aged 32-38 years performed better than younger as well as older subjects. The reason for this is not possible to detect from this thesis.
In general, capacity in PPTs has been shown to decrease over the years. The strength in trunk extension as well as flexion was reported to be decreasing after the age of 40 years by Hasue and co-workers 1980. Alaranta and co-workers found that dynamic sit-ups and repetitive squatting performance decreased with advancing age (Alaranta et al 1994). Sunnerhagen and co-workers found that self-selected walking speed decreased slightly with age, and that knee flexor and extensor strength decreased with age (Sunnerhagen et al 2000).

Concluding comments on the physical performance tests

The Åstrand test

The repeatability coefficients were too high for the test to be considered reliable, but if there is a possibility of repeating the test on at least two test occasions, it can be valuable for evaluating overall cardiovascular capacity and as a basis for health-related discussions. In our study, the proportion of persons who did not manage to get a test result was high (13%). Perhaps a treadmill test would be preferable, such as the Single-Stage Submaximal Treadmill Walking test, described by Noonan and Dean 2000. This test is, however, to the author's knowledge not tested for persons with long-term spinal pain.

The isometric endurance tests

All the isometric endurance tests had very high variability between test occasions. Theoretically, they should be able to contribute valuable information to an overall evaluation. As performed in this thesis, they are not recommended due to these high repeatability coefficients and the high proportions of ceiling effects. The latter could be removed by increasing the external applied load for males, or by simply letting the TPs continue the tests until exhaustion. This, however, could be very time-consuming, and would often not be possible due to lack of time in the clinic. Besides, the variability between test occasions would presumably increase even further.

The step-on-stool test

The step-on-stool test showed high repeatability coefficients. Valid primarily for persons with lumbar pain. Moderate to large effect sizes for changes in self-efficacy in men. However, the test is not recommended due to the high variability between test occasions and deficits in practicability: the proportion of persons reaching the

maximum value was high, and special equipment is needed. The stool heights used in this thesis did not discriminate sufficiently between men and women.

The PILE tests

The PILE tests constitute good opportunities for observation of the TP in whole-body movement. Much 'soft' information is revealed through the performances of the tests. Likewise, much 'hard' information is revealed, of which in this thesis mainly maximal weight lifted was used. Both tests have high practicality. Special equipment is needed, but is easy to find and practical for other purposes, too.

The PILE lumbar test has high variability between test occasions, but may be considered if repeated at least twice. It is valid for persons with long-term spinal pain, in the lumbar as well as in the neck region. Moderate to high effect sizes shown for both genders, for improvements as well as deterioration in men, but only for deterioration in women.

The PILE cervical test. The PILE cervical test is highly reliable, and is valid for persons with long-term spinal pain. It showed sensitivity to change for both genders, particularly in women, plus signs of responsiveness for 'least fit' men. Effect sizes moderate for improvements in both genders.

The gait tests

Walking is a necessary part of most people's lives, and therefore a valid measure of impairments and activity limitations. The tests all had high overall practicality, and rather low variability.

The gait test is reliable, and is valid primarily for persons with lumbar pain. Some sensitivity to change, and large effect sizes for deterioration in women. It is very quickly performed and can serve as a 'warm-up' for the gait test with burden. Using both these tests also allows for clinically interesting comparisons.

The gait test with burden is reliable, and it is valid for women with long-term spinal pain, and for men with lumbar pain. There is a somewhat high attrition rate. Sensitive to change in all examined ways. Effect sizes moderate-to-high for both improvement and deterioration in both genders, and for all examined outcome measures.

The stair-climbing test is reliable, and valid primarily for persons with lumbar pain. It is sensitive to change in all examined ways. Moderate effect sizes for women improving on the examined outcome measures, and large effect size for women rating deterioration in general health.

The Borg scales

The ratings of perceived pain intensity and exertion during the PPTs contributed to the overall clinical interpretation of the TPs' performance. As expected, the PPT results were related to these ratings in different ways. The different experiences perceived during bodily activities can probably not be distinguished from the pure bodily exertion in a clinical setting. The inter-relations between PPTs and

ratings like these should be studied further. The Borg Scales are being revised continuously, and later versions are being published (Borg 1998).

The UAB Pain Behavior Scale

The rated pain behaviour was consistently related to the PPT results; the more overt the pain behaviour, the less PPT result. For the individual patient, ratings of pain behaviour could contribute to the interpretation of PPT results; if the TP e.g. exhibits great pain behaviour, but nevertheless showes good results on PPTs, the interpretation may be that the TP has minor impairments and activity limitations. The UAB Scale is easy to use and requires little time. It seems recommendable to control for intra- and inter-rater reliability in the clinical setting before introducing the scale (Öhlund et al 1994). Simultaneous ratings by two raters are recommended, since discrepancies between raters then become obvious and could be adjusted for.

Further research

The work on developing and, most important, evaluating existing measurements is crucial for the physiotherapy profession. We must, as a body, be aware of limitations of the measurements we use, and use the measurements judiciously. It seems reasonable that the information from the person seeking care should be complemented by caregiver assessments. In the clinical situation, this is always the case anyhow, so we should be careful about what assessments to use to form our opinions.
Informal reliability studies in the clinic are strongly recommended before the introduction of a new measurement method, as reliability is situation- and population-specific. Not only should reliability be established once, but ideally on a regular basis. The aspects of validity examined for a measurement method by no means cover all situations or diagnoses, so caution must be exercised when deciding what tools to use.
It will be some years before we work entirely 'evidence-based', but we are not alone here – Rothstein stated in 1996 that no health care occupations had sufficient data for universal evidence-based practice (Rothstein 1996), and that is likely to be true even now, in 2002.

The most powerful outcome measures for physiotherapy purposes may be individually goal-related assessments: At the beginning of a rehabilitation intervention, individually related goals, short-term and long-term, are usually listed in co-operation with the patient. These goals could be used as outcome measures much more than is common today. Thoughts like these, which probably have long been current in the clinics though perhaps not always in a standardised way, have been formulated in scales developed for individual goal assessments. Kiresuk and Sherman developed the Goal Attainment Scaling method (Kiresuk and Sherman 1968, Rockwood et al 1997), and Stratford and co-workers reported on the

Patient-Specific Functional Scale (Stratford et al 1995, Westaway et al 1998). It is interesting, like Rothstein argued 1994 that we tend to find it natural to start an intervention based on the complaints from a patient, but have not found it natural to end the intervention based on the patient's opinion; nor to conclude that the intervention was effective (Rothstein 1994).

The profession of physiotherapy is to be congratulated when we can, like rheumatologists all over the world, agree on a pocketful of measurements, which should always be included in clinical outcome studies. The optimal approach would be that of the rheumatologists; a) consensus regarding what measures constitutes the minimally important "test kit" to include in all clinical studies (Felson et al 1993, Boers et al 1995), and b) consensus regarding what would be considered a clinically important change in these measures, in order to be able to interpret the intervention as successful (Felson et al 1995). The challenges for achievement of such a consensus in spinal pain are considerable, but not unobtainable. The importance of a thourough classification of our study participants should also be stressed. When we as physiotherapists apply such an approach, we can conduct clinical studies with much less effort, and with much more valid results, leading to a higher quality care for persons with long-term spinal pain.

Conclusions

There is an obvious need for a common basis for treatment recommendations to establish clinical consensus between professionals involved in rehabilitation regarding the evaluation and judgement of persons with long-term spinal pain. Thus, commonly accepted measurements are needed.

Clinical implications

The clinical implications of this thesis are:

- that four of the physical performance tests examined can be recommended without reservation for screening purposes, when the purpose is to evaluate impairments and activity limitations in people with spinal pain. These are: the PILE cervical lifting test, the gait test, the gait test with burden, and the stair-climbing test, all measures of activity limitations;

- that these tests, together with the PILE lumbar lifting test, could also be used before a treatment intervention, for evaluation of base-line performance as well as for collecting other ”soft data” such as pain behaviour, fear-avoidance behaviour and attitudes to the pain and to bodily activities. The PILE lumbar test should then be performed at least twice, on two separate test occasions, given the high variability between occasions shown for this test;

- that for use as outcome measures, the PILE cervical test, the gait test with burden and the stair-climbing test can be of most interest. They are all sensitive to change. However, only the PILE cervical lifting test was shown responsive to any clinically important change as defined in this thesis. If using the PILE lumbar lifting test as outcome measure, the test should be performed twice, on two separate test occasions,

- that the individual’s perceived impairments and activity limitations should guide the choice of outcome measures, and

- that physical performance tests and self-rated measures of disability complement each other, and might be used both as tools for describing disability and as outcome measures for persons with long-term spinal pain.

It is suggested that the PILE lifting tests and the gait tests are incorporated in an ‘assessment instrument bank’ for physiotherapists.

Acknowledgments

I wish to express my thanks to everyone who, in different ways, helped and supported me during these years of excitement, confusion, frustration and a lot of hard work. In particular, I wish to thank the following people:

Karin Harms-Ringdahl, my super-supervisor: you have been my guru for many years, and you have succeeded in balancing freedom and responsibility in your relations with me. Sometimes, I wished for more 'supervision', but on the other hand, that probably would not have been a good idea. Thank you for being you, a truly generous person with a 'super-computer mind'!

Åke Nygren, my boss and co-supervisor, skilled at absorbing the essence of a message in a minimum of time, and at giving wise comments and suggestions. I am really glad to have got to know you, as a person of great warmth and intellectual capacities,

All you skilled and patient statisticians, to whom I went with my numerous questions and 'weird' data; Åke Björnham (co-author in Study II), Lennart Bodin (co-author in Study I), Mathias Nilsson, Stefan Stark, and Anna Wiklund – you are all, in different ways, fabulous!

My best friend and colleague Carina Boström – your never-failing ambitions and eagerness to learn and to always discuss everything - even when I just wanted to eat my lunch or talk about everyday life struggles - inspired me to start this work. Thank you for being there, after all these years and despite the sometimes hard words we have exchanged,

Irene Jensen, co-author in Studies I, and III-V, for giving me the opportunity to be part of a large multicentre study, allowing me access to numerous scientific resources, and for sound advice on research,

Gunnar Bergström, co-author in Study I, for sharing doctoral studentship and many late evenings collecting data, and for always being able to keep things on track,

Britt Fransson, my colleague, co-author and data-collector in Study II, for friendship and support during the years,

Malou Eliason and Karin Lind, PT colleagues, for good co-operation in the HUR Project, for many nice chats, and for participation in the inter-rater agreement study,

My colleagues at HälsoInvest, in particular Sonya Mellqvist and Katarina Jakobsson, for participating in the inter-rater repeatability study, and Alice Kvåle for introducing the PILE tests to me,

Lena Hurtig, for good co-operation and friendship in the HUR Project,

Eva Nilsson, for sharing ‘bad old times’, and for being a truly trustworthy, nice person,

Christian Garheden, Anders Grahn, Anders Hägg, Walis Ludvigsson, and Gull-Britt Norrgård, for help in different ways during the HUR Project,

All staff at the former HälsoInvest Medborgarplatsen, Stockholm, at the former Yrkesinriktad Rehabilitering, MAS, Malmö, and at Förekom Projektet, Helsingborg, for participation, help and patience during the HUR Project,

The staff at HälsoInvest Ramlösakliniken, for good co-operation during the years, in the HUR Project as well as other projects, in particular Görel Rietz,

The staff at Rygginstitutet Växjö, in particular Gunilla Palm, and at Rygginstitutet Sundsvall, in particular Ingrid Käck, at STRONG in Haninge, Stockholm, and at the former LOKA Rehab, Hällefors, for good co-operation during the HUR Project,

Tim Crosfield, who with never-failing professionalism and oceans of patience has helped me revise my ‘swenglish’ into harmonious English,

My skilled PT colleagues Nina Buer, Britt Elfving, Maria Klässbo, Lena Nilsson-Wikmar, and Eva Rasmussen, for generously reading and revising my manuscripts,

Lisbet Broman, for help with editing the text (even late into the night!) and with practical issues before the completion of this thesis,

All the test participants, whose contribution was essential for the studies, especially all you who participated in the repeatability study,

My husband Pierre, who has never questioned my choosing to be a doctoral student, despite late evenings and travel, a house full of papers and heavy files, and a garden full of weeds. Thank you also for helping me with all ‘nasty’ figures,

Our sons Markus and Tobias, for being the essentials in my life. A special thanks to you Markus for drawing of the cover illustration, and to both of you for having had to accept a mother who occupied the computer these last months,

My late mother Inez and my father Martin, for inherited ambition, hard-working skills and stubbornness, properties essential for this work.

Financial support for the studies was gratefully received from AFA Insurance Company, Sweden, from the Swedish Medical Research Council (project number 5720), and from the Research Committee for Health and Caring Sciences, Karolinska Institutet, Stockholm, Sweden.

References

Abenheim L, Rossignol M, Valaat J-P, Nordin M, Avouac B, Blotman F et al for the Paris Task Force. The role of activity in the therapeutic management of back pain. Report of the International Paris Task Force on Back Pain. Spine 2000;25:1S-33S.

Alaranta H, Rytökoski U, Rissanen A, Talo S, Rönnemaa T, Puukka P et al. Intensive physical and psychosocial training program for patients with chronic low back pain. Spine 1994 a;19(12):1339-1349.

Alaranta H, Hurri H, Heliövaara M, Soukka A, Harju R. Non-dynamic trunk performance tests: reliability and normative data. Scandinavian Journal of Rehabilitation Medicine 1994 b;26:211-215.

Alricsson M, Harms-Ringdahl K, Schuldt K, Ekholm J, Linder J. Mobility, muscular strength and endurance in the cervical spine in Swedish Air Force pilots. Aviation Space & Environmental Medicine 2001;72:336-342.

Angoff W. Validity: An evolving concept. In: Wainer H and Braun HI (Editors). Test Validity. LEA Publishers, Hillsdale New Jersey 1988, Chapter 2.

Armstrong LE, Costill DL. Variability of respiration and metabolism: Responses to submaximal cycling and running. Research Quartely 1985;56:93-96.

Bandura A. Self-efficacy: toward a unifying theory of behavior change. Psychological Review 1977;84:475-488.

Barnekow-Bergkvist M, Hedberg G, Janlert U, Jansson E. Development of muscular endurance and strength from adolescence to adulthood and level of physical capacity in men and women at the age og 34 years. Scandinavian Journal of Medicine and Science in Sports 1996;6:145-155.

Basmajian J (Ed.) Physical Rehabilitation Outcome Measures. Canadian Physiotherapy Association, Williams & Wilkins, Baltimore, 1995 (third printing).

Baumgartner TA. Stability of physical performance test scores. Research Quarterly 1969;40:257-261.

Becque MD, Katch V, Marks C, Dyer R. Reliability and within subject variability of VE, VO_2, heart rate and blood pressure during submaximum cycle ergometry. International Journal of Sports Medicine 1993;14:220-223.

Biering-Sörensen F. Physical measurements as risk indicators for low back trouble over a one-year period. Spine 1984;9:106-119.

Binkley J, Finch E, Hall J, Black T, Gowland C. Diagnostic classification of patients with low back pain: Report on a survey of physical therapy experts. Physical Therapy 1993;73:138-155.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; February:307-310.

Bland JM, Altman DG. Measurement error and correlation coefficients. British Medical Journal 1996;313:41-42.

Bland JM, Altman DG. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 1999;8:135-160.

Bland M. An Introduction to Medical Statistics. Third edition. Oxford Medical Publications, Oxford University Press, New York 2000.

Boers M, van Riel PL, Felson DT, Tugwell P. Assessing the activity of rheumatoid arthritis. Baillieres Clinical Rheumatology 1995;9:305-317.

Bohannon RW. Objective measures. Physical Therapy 1989;69:590-593.

Borg G. Perceived exertion as an indicator of somatic stress. Scandinavian Journal of Rehabilitation Medicine 1970;2:92-98.

Borg G. A category scale with ratio properties for intermodal and interindividual comparisons. Proceedings from 22:nd International Congress of Psychology, Leibzig, GDR, 1982, pp.25-34.

Borg G. Borg´s Perceived Exertion and Pain Scales. Champaign, IL: Human Kinetics, 1998.

Borkovec TD, Sidney DN. Credibility of analogue therapy rationales. Pergamon Press 1972;3:257-260.

Broberg C. Physiotherapy and classification. 1997 LSR, Swedish Association of Registered Physiotherapists, Stockholm, Sweden (in Swedish).

Bäckman E, Johansson V, Häger B, Sjöblom P, Henriksson KG. Isometric muscle strength and muscular endurance in normal persons aged between 17 and 70 years. Scandinavian Journal of Rehabilitation Medicine 1995;27:109-117.

Cady LD, Bischoff MPH, O´Connell MS, Thomas BA, Allan MD. Strength and fitness and subsequent back injuries in firefighters. Journal of Occupational Medicine 1979; 21:269-272.

Cherkin DC, Deyo RA, Wheeler K, Ciol MA. Physician variation in diagnostic testing for low back pain. Who you see is what you get. Arthritis Rheumatology 1994;37:15-22.

Cohen J. Statistical Power Analysis for the Behavioral Sciences. Revised Edition. Academic Press, New York 1977.

Cox NJM, Hedriks JCM, Binkhorst RA, Folgering HThM, van Herwaarden CLA. Reproducibility of incremental maximal cycle ergometer tests in patients with mild to moderate obstructive lung diseases. Lung 1989;167:129-133.

Crombez G, Vlaeyen JWS, Heuts PHTG, Lysens R. Pain-related fear is more disabling than pain itself: evidence on the role of pain-related fear in chronic back pain disability. Pain 1999;80:329-339.

Dedering Å, Németh G, Harms-Ringdahl K. Correlation between electromyographic spectral changes and subjective assessment of lumbar muscle fatigue in subjects without pain from the lower back. Clinical Biomechanics 1999;14:103-111.

Dedering Å, Roos af Hjelmsäter M, Elfving B, Harms-Ringdahl K, Németh G. Between-days reliability of subjective and objective assessments of back extensor muscle fatigue in subjects without lower-back pain. Journal of Electromyography and Kinesiology 2000;10:151-158.

Dehlin O, Berg S, Hedenrud B, Andersson G, Grimby G. Muscle training, psychological perception of work and low-back symptoms in nursing aids. Scandinavian Journal of Rehabilitation Medicine 1978;10:201-209.

Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clinical Trials 1991;12:142S-158S.

Deyo RA, Andersson G, Bombardier C, Cherkin DC, Keller RB, Lee CK, Liang MH et al. Outcome measures for studying patients with low back pain. Spine 1994;19:2032S-2036S.

Deyo RA, Battié M, Beurskens AJ, Bombardier C, Croft P, Koes B, Malmivaara A, et al. Outcome measures for low back pain research. A proposal for standardised use. Spine 1998;23:2003-2013. Published erratum appears in Spine 1999;15:418.

Di Fabio RP, Mackey G, Holte JB. Disability and functional status in patients with low back pain receiving worker´s compensation: A descriptive study with implications for the efficacy of physical therapy. Physical Therapy 1995;75:180-193.

Dijkers M. Measuring quality of life: methodological issues. (Review). American Journal of Physical medicine & Rehabilitation 1999;78:286-300.

Dunbar CC. Practical use of ratings of perceived exertion in a clinical setting. Sports Medicine 1993;16:221-224.

Feinstein AR. Clinical Judgement. The Williams & Wilkins Company, Baltimore 1967.

Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. Annals of Internal Medicine 1983;99:843-848.

Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, Furst D, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis & Rheumatism 1993;36:729-740.

Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldschmith C, Katz LM, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis & Rheumatism 1995;38:727-735.

Fishbain DA, Abdel-Moty E, Cutler R, Khalil TM, Sadek S, Rosomoff RS, Rosomoff HL. Measuring residual functional capacity in chronic low back pain patients based on the dictionary of occupational titles. Spine 1994;19:872-880.

Fordyce WE. Behavioral Methods for Chronic Pain. St. Louis, Mosby 1976.

Fordyce WE, Lansky D, Calcyn DA, Shelton JL, Stolov WC, Rock DL. Pain measurement and pain behavior. Pain 1984;18:53-69.

Fothergill DM, Grieve DW, Pinder AD. The influence of task resistance on the characteristics of maximal one- and two-handed lifting exertions in men and women. European Journal of Applied Physiology 1996;72:430-439.

George SZ, Fritz JM, Erhard RE. A comparison of fear-avoidance beliefs in patients with lumbar spine pain and cervical spine pain. Spine 2001;26:2139-2145.

Gordon NF, Kohl HW, Pollock ML, Vaandrager H, Gibbons LW, Blair SN. Cardiovascular safety of maximal strength testing in healthy adults. The American Journal of Cardiology 1995;76:851-853.

Grönblad M, Hurri H, Kouri J-K. Relationships between spinal mobility, physical performance tests, pain intensity and disability assessments in chronic low back pain patients. Scandinavian Journal of Rehabilitation Medicine 1997;29:17-24.

Haldorsen EMH, Kronholm K, Skouen JS, Ursin H. Multimodal cognitive behavioral treatment of patients sicklisted for musculoskeletal pain. A randomized controlled study. Scandinavian Journal of Rheumatology 1998;27:16-25.

Hansson E, Hansson T. Medicinska åtgärder för sjukskrivna med rygg- och nackbesvär (p.74). Rygg och nacke 3, Stockholm, Riksförsäkringsverket och Sahlgrenska sjukhuset 1999 (in Swedish).

Hansson M, Boström C, Harms-Ringdahl K. Living with spine-related pain in a changing society – a qualitative study. Disability and Rehabilitation 2001;23:286-295.

Hansson T and Westerholm P (Editors). Arbete och besvär i rörelseorganen. En vetenskaplig värdering av frågor om samband. Arbete och hälsa, Vetenskaplig skriftserie. ISBN 91-7045-610-0. Arbetslivsinstitutet 2001 (in Swedish).

Harding VR, de C Williams AC, Richardson PH, Nicholas MK, Jackson JL, Richardson IH, et al. The development of a battery of measures for assessing physical functioning of chronic pain patients. *Pain* 1994;58(3):367-375.

Harms-Ringdahl K, Schuldt K, Ekholm J, Lannersten L, Kosek E, and Stockholm Music I Study Group. Subjektiv ansträngningsgrad vid isometrisk belastning av halsryggsextensorer i Stockholmsundersökningen 1. In: Hagberg M, Hogstedt C (Ed:s). Music Books 1991. ISBN 91-971497-1-3. Spånga Tryckeri, Sweden (in Swedish).

Hasue M, Fujiwara M, Kikushi S. A new method of quantitative measurement of abdominal and back muscle strength. Spine 1980;5:143-148.

Hazard RG, Haugh LD, Green PA, Jones PL. Chronic low back pain. The relationship between patient satisfaction and pain, impairment, and disability outcomes. Spine 1994;19:881-887.

Hirsch G, Beach G, Cooke C, Menard M, Locke S. Relationship between performance on lumbar dynamometry and Waddell Score in a population with low-back pain. Spine 1991;16:1039-1043.

Hoeymans N, Feskens EJM, van der Bos AM, Kromhout D. Measuring functional status: Cross-sectional and longitudinal associations between performance and self-report (Zuthpen Elderly Study 1990-1993). Journal of Clinical Epidemiology 1996;49:1103-1110.

Holten O. Medisinsk treningsterapi. Fysioterapeuten 1976;43(J an):9-14 (in Norwegian).

Hultman G, Nordin M, Saraste H, Ohlsén H. Body composition, endurance, strength, cross-sectional area, and density of mm erector spinae in men with and without low back pain. Journal of Spinal Disorders 1993;6:114-123.

Härkäpää K. Psychosocial factors as predictors for early retirement in patients with chronic low back pain. Journal of Psychosomatic Research 1992;36:553-559.

Ito T, Shirado O, Suzuki H, Takahashi M, Kaneda K, Strax TE. Lumbar trunk muscle endurance testing: An inexpensive alternative to a machine for evaluation. Archives of Physical Medicine and Rehabilitation 1996;77:75-79.

Jensen I, Bergström G, Ljungquist T, Bodin L, Nygren ÅL. A randomized controlled component analysis of a behavioral medicine program for chronic spinal pain: are the effects dependent on gender? Pain 2001;91:65-78.

Jensen I. Kartläggning av rehabiliteringsinsatser för långtidssjukskrivna/-förtidspensionerade arbetare och tjänstemän med besvär från ryggkotpelaren. Rapport 3, Sektionen för Personskadeprevention, Karolinska Institutet, Stockholm, Sweden 1998 (in Swedish).

Jette AM. Outcomes research: Shifting the dominant research paradigm in physical therapy. Physical Therapy 1995;75:965-970.

Jette DU, Jette AM. Physical therapy and health outcomes in patients with spinal impairments. Physical Therapy 1996;76:930-945.

Kalaoukalanai D, Cherkin DC, Sherman KJ, Koepsell TD, Deyo RA. Lessons from a trial of acupuncture and massage for low back pain. Spine 2001;26:1418-1424.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interepreting changes in health status. Medical Care 1989;27:S178-S189.

Khodadadeh S, Eisenstein SM. Gait analysis of patients with low back pain before and after surgery. Spine 1993;18:1451-1455.

Kiresuk TJ, Sherman RE. Goal attainment scaling: a general method for evaluating comprehensive community mental health programs. Common Mental health 1968;4:443-453.

Kirshner B, Guyatt G. A methodological framework for assessing health indices. Journal of Chronic Diseases 1985;38:27-36.

Kjellman G, Alexandersson K, Hensing G, Öberg B. A 12-year follow-up of subjects initially sick-listed with neck/shoulder or low back diagnosis. Physiotherapy Research International 2001;6:61-73.

Krause N, Ragland DR. Occupational disability due to low back pain: A new interdisciplinary classification based on a phase model of disability. Spine 1994;19:1011-1020.

Krout RM, Anderson TP. Role of anterior cervical muscles in production of neck pain. Archives of Physical Medicine and Rehabilitation 1966; Sept:603-611.

Kuukkanen T. Muscular performance after a 3 month progressive physical exercise program and 9 month follow-up in subjects with low back pain. A controlled study. Scandinavian Journal of Medicine and Science in Sports 1996;6:112-121.

Lackner JM, Carosella AM, Feuerstein M. Pain expectancies, pain, and functional self-efficacy expectancies as determinants of disability in patients with chronic low back disorders. Journal of Consulting and Clinical Psychology 1996;64:212-220.

Lee CE, Simmonds MJ, Novy DM, Jones S. Self-reports and clinician-measured physical function among patients with low back pain: A comparison. Archives of Physical Medicine and Rehabilitation 2001;82:227-231.

Lin CC. Comparison of the effects of perceived self-efficacy on coping with chronic cancer pain and coping with chronic low back pain. Clinical Journal of Pain 1998;14:303-310.

Lindström I, Öhlund C, Eek C, Wallin L, Peterson L-E, Nachemsson A. Mobility, strength, and fitness after a graded activity program for patients with subacute low back pain. Spine 1992,17:641-652.

Lindström I, Öhlund C, Nachemsson A. Validity of patient reporting and predictive value of industrial physical demands. Spine 1994;19:888-893.

Linton SJ, Halldén K. Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. The Clinical Journal of Pain 1998;14:209-215.

Linton SJ. A review of psychological risk factors in back and neck pain. Spine 2000;25:1148-1156.

Linton SJ, van Tulder MW. Preventive interventions for back and neck pain problems. What is the evidence? Spine 2001;26:778-787.

Ljungkvist I. Short- and long-term effects of a 12-week intensive functional restoration programme in individuals work-disabled by chronic spinal pain. Scandinavian Journal of rehabilitation Medicine 2000;32:1-14.

Lockwood PA, Yoder JE, Deuster PA. Comparison and cross-validation of cycle ergometry estimates of VO 2max. Medicine & Science in Sports & Exercise 1997;29:1513-1520.

Lomi C. Evaluation of a Swedish version of the Arthritis Self-efficacy Scale. Thesis for degree of Licentiate, ISSN 1102-8491, Lunds University, Sweden 1995.

LSR, Swedish Association of Registered Physiotherapists: Description of physiotherapy, and physiotherapy as a field of practice. Sjukgymnasten 1998;4:31.

Luoto S, Heliövaara M, Alaranta H. Static back endurance and the risk of low-back pain. Clinical Biomechanics 1995;10:323-324.

Lurie JD. Point of view. Spine 2001; 26:77.

McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. Oxford University Press, New York, Oxford 1987.

McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-item Short Form Health Survey (SF-36). III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Medical Care 1994;32:40-66.

McQuade KJ, Turner JA, Buchner DM. Physical fitness and chronic low back pain.: An analysis of the relationships among fitness, functional limitations, and depression. Clinical Orthopaedics and Related Research 1988; 233:198-204.

Magnusson T. PILE – ett lyfttest för byggnadsarbetare? Högskolan i Växjö 1992 (in Swedish).

Manniche C, Hesselsöe G, Bentzen L, Christensen I, Lundberg E. Clinical trial of intensive muscle training for chronic low back pain. Lancet 1988;2:1473-1476.

Mannion AF, Taimela S, M ntener M, Dvorak J. Active therapy for chronic low back pain. Part I. Effects on back muscle activation, fatigability, and strength. Spine 2001;26:897-908.

Marklund S (ed.). RFV Report 1997:6, Stockholm 1997 (Risk-friskfaktorer, in Swedish).RFV, Publikationsservice, S-103 51 Stockholm, Sweden.

Mayer T, Barnes D, Kishino N, et al. Progressive Isoinertial Lifting Evaluation – I. A Standardized Protocol and Normative Database (published erratum appears in Spine 1990;15:5). Spine 1988 a;13(9):993-997.

Mayer T, Barnes D, Nichols G, et al. Progressive Isoinertial Lifting Evaluation – II. A Comparison with Isokinetic Lifting in a Disabled Chronic Low-Back Pain Industrial Population. Spine 1988 b;13(9):998-1002.

Mellin G, Hurri H, Härkäpää K, Järvikoski A. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part II. Effects on physical measurements three months after treatment. Scandinavian Journal of Rehabilitation Medicine 1989;21:91-95.

Merskey H, Bogduk N (ed:s). Classification of Chronic Pain. Description of Chronic Pain Syndromes and Definitions of Pain Terms. 2:nd ed. IASP Press, Seattle, USA, 1994.

Miller AEJ, MacDougall JD, Tarnopolsky MA, Sale DG. Gender differences in strength and muscle fiber characteristics. European Journal of Applied Physiology 1993;66:254-262.

Moffroid MT, Haugh LD, Henry SM, Short B. Distinguishable groups of musculoskeletal low back pain patients and asymptomatic control subjects based on physical measures of the NIOSH Low Back Atlas. Spine 1994;19:1350-1358.

Moffroid M, Reid S, Henry SM, Haugh LD, Ricamato A. Some endurance measures in persons with chronic low back pain. JOSPT 1994;20:81-87.

Moffroid MT. Endurance of trunk muscles in persons with chronic low back pain: Assessment, performance, training. Journal of Rehabilitation research and Development 1997;34:440-447.

Mooney V. Functional capacity testing: Its role in assessing and treating back pain. Pain Management 1990;March/April:107-113.

Moreland J, Finch E, Stratford P, Balsor B, Gill C. Interrater reliability of six tests of trunk muscle function and endurance. Journal of Orthopaedic & Sports Physical therapy (JOSPT) 1997;26:200-208.

Nachemsson A, Jonsson E, Carlsson C.-A, Englund L, Goossens M, Harms-Ringdahl K, Linton SJ et al. Ont i ryggen, ont i nacken. En evidensbaserad kunskapssammanställning. ISBN 91-87890-60-7. Report no 145/1.The Swedish Council on Technology Assessment in Health Care (SBU), Stockholm, 2000 (in Swedish). A summary is available in English from SBU, tel. +46-8-412 32 00.

National Board of Health and Welfare. Peoples Health. Report 1997:18 (Folkhälsorapporten, in Swedish), Stockholm, Sweden.

Németh G, Ekholm J, Arborelius UP, Harms-Ringdahl K, Schuldt K. Influence of knee flexion on isometric hip extensor strength. Scandinavian Journal of Rehabilitation Medicine 1983;15:97-101.

Noonan V, Dean E. Submaximal exercise testing: Clinical application and interpretation. Physical Therapy 2000;80:782-807.

Ohnmeiss DD. Pain drawings in the evaluation of lumbar disc-related pain. Thesis, ISBN 91-628-4069-X, Division of Rehabilitation Medicine, Karolinska Institutet, Stockholm, Sweden 2000.

Payton OD. Research: The Validation of Clinical Practice. Edition 2. F.A. Davis Company, Philadelphia, 1988, pp. 74-79.

Piela CR, Hallenberg KK, Geoghegan AE, Monsein MR, Lindgren BR. Prediction of functional capacities. Work 1996;6:107-113.

Puniello MS, McGibbon CA, Krebs DE. Lifting strategy and stability in strength-impaired elders. Spine 2001;26:731-737.

Ralston HJ. Energy-speed relation and optimal speed during level walking. International Zeitung angew. Physiol einschl. Arbeitsphysiology 1958;Bd.17:277-283.

Randlöv A, Östergaard M, Manniche C, Kryger P, Jordan A, Heegard S, Holm B. Intensive dynamic training for females with chronic neck/shoulder pain. A

randomized controlled trial. Clinical Rehabilitation 1998;12:200-210.

Reid S, Hazard RG, Fenwick JW. Isokinetic trunk-strength deficits in people with and without low-back pain: a comparative study with consideration of effort. Journal of Spinal Disorders 1991;4:68-72.

Reilly K, Lovejoy B, Williams R, Roth H. Difference between a supervised and independent strength and conditioning program with chronic low back pain syndromes. Journal of Occupational Medicine 1989;31:547-550.

RFV. Risk-friskfaktorer – sjukskrivning och rehabilitering i Sverige. 1997, Riksförsäkringhsverket, Publikationsservice, 103 51 Stockholm,Sweden (in Swedish).

Richards JS, Nepomuceno C, Riles M, Suer Z. Assessing pain behavior: The Pain Behavior Scale. Pain 1982;14:393-398.

Rockwood K, Joyce B, Stolee P. Use of Goal Attainment Scaling in measuring clinically important change in cognitive rehabilitation patients. Journal of Clinical Epidemiology 1997;50:581-588.

Rodriquez AA, Bilkey WJ, Agree JC. Therapeutic exercise in chronic neck and back pain. Review article. Archives of Physical Medicine and Rehabilitation 1992;73:870-875.

Rothstein J (ed). Measurement in Physical Therapy. Churchill Livingstone 1985.

Rothstein JM. Disability and our identity. Editor's Note. Physical Therapy 1994;74:375-378.

Rothstein JM. Outcomes and survival. Editor's Note. Physical Therapy 1996;76:126-127.

Salén BA, Spangfort EV, Nygren ÅL, Nordemar R. The Disability Rating Index: An instrument for the assessment of disability in clinical settings. Journal of Clinical Epidemiology 1994;47:1423-1435.

Sandström J, Esbjörnsson E. Return to work after rehabilitation. The significance of the patient´s own prediction. Scandinavian Journal of Rehabilitation Medicine 1986;18:29-33.

Schmidt AJM. Cognitive factors in the performance level of chronic low back pain patients. Journal of Psychosomatic research 1985;29:183-189.

Selander J, Marnetoft S-U, Bergroth A, Ekholm J. The process of vocational rehabilitation for employed and unemployed people on sick-leave: employed people vs unemployed people in Stockholm compared with circumstances in rural Jämtland, Sweden. Scandinavian Journal of Rehabilitation Medicine 1998;30:55-60.

Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, Radwan H. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine 1998;23:2412-2421.

Silverman JL, Rodriquez AA, Agre JC. Quantitative cervical flexor strength in healthy subjects and in subjects with mechanical neck pain. Archives of Physical Medicine and Rehabilitation 1991;72:679-681.

Skargren EI, Öberg BE. Predictive factors for 1-year outcome of low-back and neck pain in patients treated in

primary care: comparison between the treatment strategies chiropractic and physiotherapy. Pain 1998;77:210-207.
Spangfort E. Clinical aspects of neck-and-shoulder pain. Scandinavian Journal of Rehabilitation Medicine 1995;Suppl 32:43-46.

Spitzer WO, LeBlanc FE, Dupius M. Scientific approach to the assessment and management of activity-related spinal disorders: Report on the Quebec Task Force on Spinal Disorders. Spine 1987;12(7S):S9-S54.

SPSS Base 9.0. User's Guide, 1999, SPSS Inc., USA, ISBN 0-13-020390-4.

Stenström CH. Home exercise in rheumatoid arthritis functional class II: Goal setting versus pain attention. Journal of Rheumatology 1994;21:627-634.

Stenström CH. Home exercise in rheumatoid arthritis functional class II: Goal setting versus pain attention. Journal of Rheumatology 1994;21:627-634.

Strand LI and Moe-Nilssen R. Back Performance Scale (BPS) for the assessment of mobility-related activities in back pain. Submitted 2001.

Stratford P, Gill C, Westaway M, Binkley J. Assessing disability and change on individual patients: a report of a patient specific measure. Physiotherapy Canada 1995;47:258-263.

Stratford PW, Riddle DL, Binkley JM, Spadoni G, Westaway MD, Padfield B. Using the Neck Disability Index to make decisions concerning individual patients. Physiotherapy Canada 1999; Spring:107-112(119).

Stratford PW, Spadoni G, Kennedy D, Westaway MD, Alcock GK. Seven points to consider when investigating a measure's ability to detect change. Physiotherapy Canada 2002;Winter:16-24.

Sunnerhagen Stibrant K, Hedberg M, Henning G-B, Cider Å, Svantesson U. Muscle performance in an urban population sample of 40- to 79.year-old men and women. Scandinavian Journal of Rehabilitation Medicine 2000;32:159-167.

Swiontkowski MF, Engelberg R, Martin DP, Agel J. Short Musculoskeletal Function Assessment Questionnaire: Validity, reliabiliuty, and responsiveness. The Journal of Bone and Joint Surgery 1999;81-A:1245-1260.

Söderlund A. Physiotherapy management, coping and outcome prediction in whiplash associated disorders (WAD). Thesis, ISBN 91-554-4948-4, Uppsala University, Sweden 2001.

Task Force on Standards for Measurement in Physical Therapy. Standards for tests and measurements in physical therapy practice. Physical Therapy 1991;71:589-622.

Toomingas A. Characteristics of pain drawings in the neck-shoulder region among the working population. International Archives of Occupational and Environmental Health 1999;72:98-106.

Troup JDG, Foreman TK, Baxter CE, Brown D. The perception of back pain and the role of psychophysical tests of lifting capacity. Spine 1987;12:645-657.

Turk DC, Melzack R (Ed:s). Handbook of Pain Assessment. ISBN 0-89862-883-0. The Guilford Press, New York 1992.

United Nations. World Programme of Action Concerning Disabled Persons. United Nations, New York 1982.

Unruh AM. Gender variations in clinical pain experience. Review article. Pain 1996;65:123-167.

Vandermeulen DM, Birmingham TB, Forwell LA. The test-retest reliability of a novel functional test: The lateral hop for distance. Physiotherapy Canada;Winter:50-55.

Waddell G, Bircher M, Finlayson D, Main CJ. Symptoms and signs: physical disease or illness behaviour? British Medical Journal 1984;289:739-741.

Waddell G, Newton M, Henderson I, Somerville D. Letter. *Spine* 1993;18:938-939.

Wade DT. Editorial. Clinical Rehabilitation 1998;12:1-2.

Westaway MD, Stratford PW, Binkley JM. The Patient-Specific Functional Scale: validation of its use in persons with neck dysfunction. JOSPT 1998;27:331-338.

WHO 2001. International Classification of Functioning, Disability and Health. Final Draft. Geneve: World Health Organisation, December 2000. url:http://who.int/icidh.

Williams RM, Myers AM. A new approach to measuring recovery in injured workers with acute low back pain: Resumption of Activities of daily Living Scale. Physical Therapy 1998;78:613-623.

Williams AC de C, Richardsson PH, Nicholas MK, Pither CE, Harding VR, Ridout KL, Ralphs JA, Richardssson IH, Justins DM, Chamberlain JH. Inpatient vs. outpatient pain management: results of a randomised controlled trial. Pain 1996;66:13-22.

Williams JG, Eston RG. Determination of the intensity dimension in vigorous exercise programmes with particular reference to the use of the Rating of Perceived Exertion. Review article. Sports Medicine 1989;8:177-189.

Wittink H, Hoskins Michel T, Wagner A, Sukiennik A, Rogers W. Deconditioning in patients with chronic low back pain. Fact or fiction? Spine 2000;25:2221-2228.

Wormgoor MEA, Björholt PG. Fysisk funksjonsnivå ved kronisk vond rygg. Metoder for en objektiv vurdering som ledd i en diagnostikk og behandling. Tidskrift for den Norske Laegeforeningen 1994;114:1301-1305 (in Norwegian).

Åberg J. Evaluation of an advanced back pain rehabilitation program. Spine 1984;9:317-318.

Åstrand P-O, Ryhming I. A nomogram for calculation of aerobic capacity (physical fitness) from pulse rating during submaximal work. Journal of Applied Physiology 1954;7:218.

Åstrand P-O, Rodahl K. Textbook of Work Physiology. Physiological bases of exercise. Third edition. McGraw-Hill Book Company. 1986.

Öhlund C, Lindström I, Areskoug B, Eek C, Peterson L-E, Nachemsson A. Pain behavior in industrial subacute low

back pain. Part I. Reliability: concurrent and predictive validity of pain behavior assessments. Pain 1994;58:201-209.

Appendix 1

Testprotokoll sjukgymnasten

Namn: ..... Pers.nr .....

Tillfälle:	**1**		**2**		**3**		**4**	
Datum:								
Testare:								
1. Konditionstest (Åstrand)								
Stand.krav uppfyllda								
Sadelhöjd								
Belastning								
Arbetspuls								
VO2 l/min	*nivå*							
VO2 ml/kg x min	*nivå*							
Smärtintensitet								
Smärtlokalisation								
Ansträngningsgrad								
Kommentar								
2. Uthållighet halsens ventrala muskler - kvarhåll flex. 10°. Vikt 0.5 kg								
Antal sekunder (normalt 60)								
Smärtintensitet								
Smärtlokalisation								
Ansträngningsgrad								
Kommentar								
3. Lyfttest (Pile lumbaltest) – se separat protokoll								
4. Uthållighet ryggextensorer - kvarhåll uträtad thorakal kyfos *								
Antal sekunder (normalt 3 min)								
Smärtintensitet								
Smärtlokalisation								
Ansträngningsgrad								
Kommentar								
5. Funktionellt test nedre extremiteten. Uppklivning på pall 40 cm kvinnor, 44 cm män								
	Hö	Vä	Hö	Vä	Hö	Vä	Hö	Vä
Antal gånger (normalt 50)								
Smärtintensitet								
Smärtlokalisation								
Ansträngningsgrad								

* **Modifierat till: Kvarhåll så nära horisontalplanet som möjligt utan att accentuera den lumbala lordosen (Ljungquist T. april 1995)**

	Tillfälle 1	Tillfälle 2	Tillfälle 3	Tillfälle 4
6. Uthållighet nackextensorer - kvarhåll 0°. Vikt 1.5 kg kvinnor, 2 kg män				
Antal sekunder (normalt 3 min)				
Smärtintensitet				
Smärtlokalisation				
Ansträngningsgrad				
Kommentar				
7. Lyfttest (Piletest cervikalt) – se separat protokoll				
8. Uthållighet bukmuskler - kvarhåll flex. när nedre scapulakanten lämnat britsen				
Antal sekunder (normalt 90)				
Smärtintensitet				
Smärtlokalisation				
Ansträngningsgrad				
Kommentar				
9. Gång 2x20 m, självvald takt				
Antal sekunder				
Smärtintensitet				
Smärtlokalisation				
Ansträngningsgrad				
Kommentar				
10. Gång enl. ovan med tyngder, 2x4 kg kvinnor, 2x8 kg män				
Antal sekunder				
Smärtintensitet				
Smärtlokalisation				
Ansträngningsgrad				
Kommentar				
11. Gång i trappa, ...trappsteg upp och ner igen, självvald takt				
Antal sekunder				
Smärtintensitet				
Smärtlokalisation				
Ansträngningsgrad				
Kommentar				

Appendix 2

PILE-testformulär – sjukgymnasterna

Namn:

Ålder: ..... **Längd (cm): .....**

PILE – lumbaltest **(Lyftsträcka per vikt 6,08 m)**

Tillfälle:	**1**	**2**	**3**	**4**		
Datum:						
Testare:						
Personvikt (kg)						
Justerad vikt (kg)						
Vilopuls						
Tid (s)						
Stopporsak					Normalvärde	
Puls					Kvinnor	Män
Slutvikt (kg)					19.3*	36.5*
Slutvikt/justerad vikt					0.35	0.50
Total work (J)					5 620	9 972
TW/justerad vikt (J/kg)					101.8	136.4
Smärtintensitet						
Smärtlokalisation						
Ansträngningsgrad						
Kommentarer						

PILE – cervikaltest **(Lyftsträcka per vikt 4,88 m)**

Personvikt (kg)						
Justerad vikt (kg)						
Vilopuls						
Tid (s)						
Stopporsak					Normalvärde	
Puls					Kvinnor	Män
Slutvikt (kg)					13.8*	29.2*
Slutvikt/justerad vikt					0.25	0.40
Total work (J)					2 414	5 380
TW/justerad vikt (J/kg)					43.8	73.4
Smärtintensitet						
Smärtlokalisation						
Ansträngningsgrad						
Kommentarer						

Lyft vikt (back + innehåll)

	Antal flaskor	1	2	3	4	5	6	7	8
Kvinnor	Vikt (kg)	3,6	5,9	8,1	10,4	12,6	14,9	17,1	19,4
	Summa vikt	3,6	9,5	17,6	27,9	40,5	55,4	72,5	91,8
	Tid (s)	**20**	**40**	**60**	**80**	**100**	**120**	**140**	**160**
	Antal flaskor	2	4	6	8	10	12	14	16
Män	Vikt (kg)	5,9	10,4	14,9	19,4	23,9	28,4	32,9	37,4
	Summa vikt	5,9	16,3	31,2	50,6	74,5	102,9	135,8	173,2

T. Ljungquist, HUR-projektet 950914. Källa: Mayer et al 1988

* Cut-off-värden enligt Ljungquist et al 1999 a:PILE lumbal: Kvinnor 12.6 kg, män 23.9 kg
PILE cervikal: kvinnor 8.1 kg, män 19.4 kg.

Bryt upp, bryt upp!
Den nya dagen gryr.
Oändligt är vårt stora äventyr.

Ur Karin Boyes ”I rörelse”, 1927