

## ORIGINAL REPORT

# TEST-RETEST RELIABILITY OF PSYCHOPHYSICAL LIFT CAPACITY IN PATIENTS WITH NON-SPECIFIC CHRONIC LOW BACK PAIN AND HEALTHY SUBJECTS

Audy P. Hodselmans, PT, MSc<sup>1</sup>, Pieter U. Dijkstra, PT, MT, PhD<sup>2</sup>, Cees van der Schans, PT, PhD<sup>3</sup>, and Jan H. B. Geertzen, MD, PhD<sup>2</sup>

From the <sup>1</sup>Hanze University Groningen School of Health Care Studies, Department of Physiotherapy, <sup>2</sup>Centre for Rehabilitation, Department of Oral and Maxillofacial Surgery, University Medical Centre Groningen, University of Groningen Northern Centre for Health Care Research, University of Groningen, and <sup>3</sup>Centre for Research and Development in Health Care and Nursing, Hanze University Groningen, The Netherlands.

**Objective:** Psychophysical lift capacity tests are lifting tests in which the performance, expressed in Newtons, is divided by the perceived exertion, expressed on a Borg scale. The aim of this study was to analyse test-retest reliability of psychophysical lift capacity tests.

**Subjects:** Patients with non-specific chronic low back pain ( $n=20$ ) and healthy subjects ( $n=20$ ).

**Methods:** Psychophysical lift capacity tests were assessed during a back school intake at the Centre for Rehabilitation of the University Medical Centre Groningen. Patients on the waiting list and healthy subjects were assessed twice, with a 2-week interval between assessments. Intra-class correlation (ICC) was calculated as a measure of reliability. An  $ICC \geq 0.75$  was considered as an acceptable reliability. Limits of agreement as a measure for natural variation were calculated.

**Results:** The psychophysical static and dynamic lift capacity tests showed good reliability ( $ICC \geq 0.75$ ). The limits of agreement are substantial, indicating a considerable natural variation between test-sessions for all psychophysical tests.

**Conclusion:** The psychophysical static lift capacity and dynamic lifting capacity are reliable instruments for patients with non-specific chronic low back pain and healthy subjects. However, a substantial amount of natural variation should be taken into account between 2 test sessions when interpreting the test results clinically.

**Key words:** reproducibility of results, low back pain, psychophysical capacity.

J Rehabil Med 2007; 39: 133–137

*Correspondence address:* Audy P. Hodselmans, Lecturer, Hanze University Groningen School of Health Care Studies, Department of Physiotherapy, Eyseniusplein 18 NL-S9714 CE Groningen, The Netherlands.

*E-mail:* A.P.Hodselmans@pl.hanze.nl

Submitted January 24, 2006; accepted October 11, 2006.

## INTRODUCTION

Traditional clinical assessments of patients with non-specific chronic low back pain (CLBP) are most commonly based on tests of physical disability (e.g. range of motion and strength)

(1–4) and tests of physical capacity that rely on a functional capacity evaluation (FCE). These tests of physical capacity are entrenched in current practice, but reliability and validity are generally weak (1, 5–7). For instance, isokinetic devices are commonly used to quantify trunk strength and motion. However, reliability of patient performance on the equipment is low and these isokinetic test batteries require extensive equipment (1). Furthermore, the constrained manner of testing is functionally unrealistic, implying that correlations between the test results and function (i.e. validity) may be questionable (1, 8).

The Isernhagen FCE also has its limitations in clinical evaluation because of ceiling and criterion effects in some of the tests (9). In these tests, many patients are not limited in performance and meet the maximal test criterion (9). Thus, improvements in performance as a result of an intervention program cannot be determined with these tests due to these criterion effects.

Perceived disability is frequently inferred from physical performance measures, but the outcomes may have little relation to actual functioning (1, 10). Changes in range of motion, muscle strength and physical fitness after a physical exercise program in a rehabilitation centre are not related to changes in the perceived disability measured with the Roland Morris Disability Questionnaire (RMDQ) (1, 11).

Patient's perceptions of the level of disability, as measured by, for example, the RMDQ (12, 13), may be subject to a perceptual or belief mismatch. There may be differences between how patients actually function and how they perceive their functioning, and there may be differences in what patients report and what practitioners observe and conclude, especially when patients are receiving injury compensation (1, 14). This mismatch may negatively influence the patient's willingness and sense of safety while accomplishing specific physical tasks that cause loading of the lower back (1, 15). Therefore, psychophysical capacity, calculated as the acceptable maximal effort (AME) of the patient divided by perceived effort of the performance, may help to address the sense of safety of the patient (16, 17). The patient's sense of safety exists because patients may continue the tests to their own maximal voluntary

effort without unacceptable pain and the perceived effort is weighted for the result of the psychophysical capacity.

It was found that the most important variable for successful treatment of CLBP is a reduction in patients' perceived disability, and that an appropriate perception of physical improvement contributes to a reduction in perceived disability (18). Following a cognitive-somatic rehabilitation approach for patients with non-specific CLBP, the change in the psychophysical trunk lift capacity was significantly related to the change in the RMDQ ( $r=0.74$ ) (16).

Until now no information has been available concerning the test-retest reliability of psychophysical capacity tests (16). The purpose of this study was to assess the test-retest reliability of the psychophysical capacity test.

## MATERIAL AND METHODS

### Participants

The test-retest reliability and the natural variation of psychophysical tests were evaluated in 20 outpatients (12 women) with non-specific CLBP and in 20 healthy, pain-free controls (10 women). The patients were recruited from the Centre for Rehabilitation of the University Medical Centre Groningen. The mean age of the CLBP group was 33.8 years (SD=8.6), mean height was 1.76 metres (SD=0.1) and mean weight was 73.9 kg (SD=15.1). The healthy subjects comprised students from the Institute for Human Movement Sciences of the University of Groningen. The mean age of the healthy subjects was 22.0 years (SD=1.6), mean height was 1.79 metres (SD=0.1) and mean weight was 72.5 kg (SD=8.5).

### Procedure

Admittance assessment was performed by a rehabilitation physician at the University Medical Centre Groningen before patients entered the study. Inclusion criteria for the waiting list of the back school rehabilitation treatment program are: (i) non-specific CLBP, lasting for more than 3 months and without demonstrable specific origin; (ii) the patient is satisfied with the diagnostic process and outcome, and is motivated to participate in the rehabilitation treatment (16). Exclusion criteria were: (i) small amount of suffering due to low back pain; (ii) conflicts with employer or insurance company; (iii) the patient is dissatisfied with the diagnostic process or outcome of it, or is not motivated to participate in the treatment program (secondary illness profits) (16). Exclusion criteria for patients and controls were: any medical condition that would interfere with psychophysical performance tests, major surgery within the last year, current infectious disease, cancer and neuralgic or cardiovascular disease. Additionally, subjects in the healthy group were excluded if they had a history of low back pain within the previous 6 months that had lasted more than one week, required medical attention, or resulted in absence from work or school. All participants signed an informed consent form. The first author was the physical therapist who assessed the psychophysical performance of the patients and the healthy subjects (16). Patients on the waiting list and healthy subjects were assessed twice; the second test was carried out after a 2-week interval. The time of day, day of the week and place of assessment were held constant for the test sessions. Patients on the waiting list were assessed twice before starting the back school rehabilitation program.

### Measurements

Aerobic capacity was estimated by means of a modified indirect protocol of the Åstrand test on a cycle ergometer (Excalibur Sport, Lode BV, Groningen, The Netherlands) before the psychophysical capacity tests were performed (16). The method for measuring lifting abilities is based on a psychophysical model referred to as AME (17).

### Psychophysical static lift capacity test

The psychophysical *static lift capacity test* was performed by pulling up a horizontal bar connected to a height-adjustable pillar (Fig. 1) (16). The vertical force was measured with a force transducer (EBN 8500-1250, Depx type brosia; GmbH & Co, Tettang, Germany; range 0–2500 N; linearity 0.02%) and an amplifier (Elan-Schaltlemente MBP 6218; Kurt Maecker GmbH, Neuss 1, Germany; range 500–5000  $\mu\text{m/m}$ ) and was registered on a plotter (PM 8043 Eindhoven, The Netherlands; range 2 mV/cm to 1 V/cm). Three lift tests described by Chaffin et al. (19) were performed: the arm lift, the leg lift and the trunk lift. The position of the subject during the arm lift was: standing straight up with the elbow bent at 90 degrees and close to the body. The equipment handles were at elbow height. During the arm lift and trunk lift the horizontal distance was 375 mm. The horizontal distance was for the leg lift was 0 mm. The vertical distance during the leg lift and trunk lift was 500 mm.

The patient was instructed to stop performing the test when the AME was reached (17). The AME is the highest level of voluntary effort that a person can achieve without inducing unacceptable pain (12). To measure patients perceived exertion the Borg score (range 0.5–10) was assessed directly after reaching the AME of the static lift capacity tests (20–22). After the test, the subject pointed to the perceived effort on a table where the explanation of the number was described. The Borg score is modified by changing the 0 score in 0.5. The explanation of the 0.5 score is not changed and corresponds to, not at all as the original score 0 from Borg, 1 = very light, 2=light, 3=moderate, 4=slightly heavy, 5=heavy, 6=less than 7, 7=very heavy, 8=less than 9, 9=less than 10, 10= very terrible heavy (almost maximum).

The psychophysical static lift capacity was calculated as the ratio of performance and the Borg score, expressed in Newton/Borg (N/B) (16).

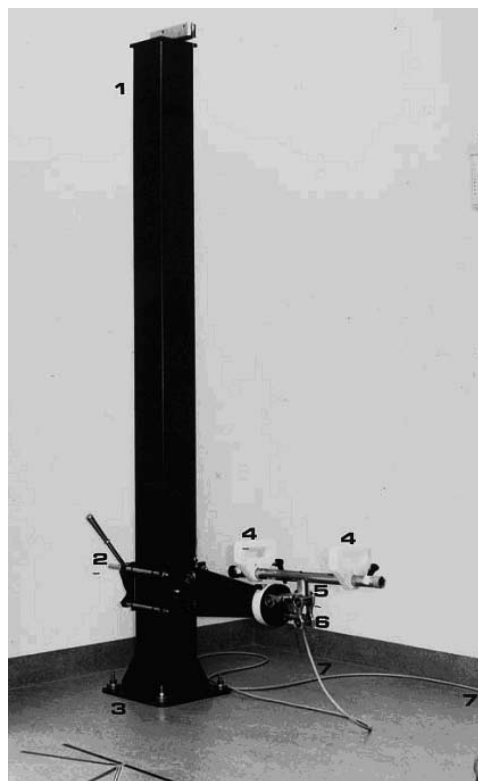


Fig. 1. Static lifting force measurement set-up: (1) pillar, (2) height adjuster, (3) floor, (4) handle, (5) joint, (6) transducer, and (7) connection with amplifier/plotter.

*Psychophysical dynamic lifting capacity test*

To measure the psychophysical dynamic lifting capacity, the standardized Progressive Isoinertial Lifting Evaluation protocol (PILE), described by Mayer et al. (23, 24) was used. For a period of 20 seconds the patient had to lift a box with weights, 4 times from the ground on to a table. Stepwise, after each session, during the 20 second rest, the weight of the box increased by 2.25 kg for women and 4.5 kg for men, respectively. Heart rate was measured with a heart rate equipment (Sport tester PE-3000; Polar Electro, Kempele, Finland (16)). The observer stopped the test when the patient had reached the heart rate safety limit (220 – age×85%) (16, 25). The patient was instructed to stop the performance when the AME was reached (17). Perceived exertion was assessed directly after reaching the AME, the cardiac safety limit, or the end of the test calculated with the Waters formula. The maximum safety load for men is 402 N and for women 202 N (26). The psychophysical dynamic lifting capacity was calculated as the ratio of performance and Borg score, expressed in N/B (16).

*Data analysis*

Descriptive statistics were calculated for the scores of the 2 test-sessions. Test-retest reliability was determined by means of a paired *t*-test and intra-class correlation coefficient (ICC, 1-way random model). To quantify variation over time, limits of agreement were calculated as ± 2 times ×SD<sub>difference</sub> (27, 28). In case of significant differences between the first and second test the limits of agreement were not calculated, because the test-retest reliability is unacceptable (28). An ICC ≥0.75 was considered as an acceptable reliability (30, 31). No criteria for interpretation of the limits of agreement are available. However, smaller limits of agreement indicate more stability over time, because it indicates that the natural variation is small (10). Plots were made of the individual difference between sessions against the individual average of the 2 sessions, as recommended to determine if the size of the difference is related to the mean performance (28). Comparisons between groups were calculated by using the mean of the 2 test occasions for each group and thereafter a between group *t*-test analysis.

Data analyses were performed using the Statistical Package of Social Sciences (SPSS 12.0).

RESULTS

Outcomes of psychophysical lift capacity tests did not differ significantly between the 2 sessions (Table I). In patients, AME

performance of physical strength only, in Newtons, was significantly lower in the second session than the first. The standard deviations of all measurements were considerable.

No funnel shape was present in the plots (Fig. 2), indicating that the difference between the 2 sessions is not related to the average performances on the 2 sessions. The other 3 figures are similar.

The ICC for the psychophysical lift capacity ranged from 0.82 to 0.93 in patients and from 0.75 to 0.94 in healthy subjects (Table II). Limits of agreements ranged for the psychophysical static lift capacity tests in patients from ± 26 N/B to ±106 N/B and in healthy subjects from ±42 N/B to ±170 N/B, respectively. The psychophysical PILE capacity had a limit of agreement in patients of ±14.2 N/B and ±19.6 N/B in healthy subjects.

Psychophysical lift capacity tests (N/B) of healthy subjects showed significantly higher values than patients in leg lift (*p*=0.007) trunk lift (*p*=0.002) and PILE (*p*=0.000), except for the psychophysical arm lift capacity tests (*p*=0.227).

DISCUSSION

A significant decrease was found in patients between session 1 and 2 of only strength in Newtons of the AME of static leg lift, trunk lift and arm lift and dynamic lifting. This significant decrease in patients with non-specific CLBP, confirms the inadequate reliability of traditional clinical assessments for these patients based on tests of physical strength alone (1). The SDs of all measurements are considerable. The cause can be the significant differences between men and women. However, the values are normally distributed calculated with the 2-tailed Kolomogorov-Smirnov test.

In all measurements of the psychophysical capacity of patients and healthy subjects, the ICC was >0.75. This means that the psychophysical static lift and dynamic lifting capacity test in patients and in healthy subjects are reliable. However,

Table I. Results of 20 patients and 20 healthy subjects in the first and second measurement sessions of the psychophysical capacity in Newton/Borg, the acceptable maximal effort in Newtons and the perceived exertion in Borg

Variable	Patients n=20			Healthy n=20		
	T1 (SD)	T2 (SD)	ΔT (SD)	T1 (SD)	T2 (SD)	ΔT (SD)
Leg lift Newton/Borg	119.1 (76.3)	108.1 (59.7)	11.1 (53)	185.1 (104.1)	183.1 (90.1)	1.73 (85)
Newton	468.1 (226.0)	395.1 (182.0)	73.3 (120)*	566.1 (117.1)	610.1 (208.1)	-44.3 (107)
Borg	4.4 (1.5)	4.0 (1.9)	0.4 (0.9)	3.6 (1.6)	3.8 (1.4)	-0.2 (1.3)
Trunk lift Newton/Borg	71.6 (40.4)	65.1 (29.7)	6.4 (20)	117.0 (54.8)	123.0 (61.9)	-6.5 (21)
Newton	301.2 (128.0)	266.2 (106.1)	34.6 (72)*	401.1 (112.0)	395.1 (142.0)	5.2 (74)
Borg	4.7 (1.6)	4.4 (1.3)	0.3 (1.1)	3.7 (1.0)	3.5 (1.1)	0.2 (0.7)
Arm lift Newton/Borg	58.8 (32.1)	54.9 (28.4)	3.8 (13)	67.3 (32.4)	71.5 (41.5)	-4.2 (26)
Newton	215.1 (68.1)	195.1 (80)	19.8 (39)*	262.1 (89.9)	252.1 (94)	10.8 (31)
Borg	4.2 (1.4)	3.8 (1.2)	0.4 (0.9)	4.3 (1.3)	4.0 (1.3)	0.3 (1.2)
PILE Newton/Borg	33.9 (17.6)	33.6 (18.9)	0.3 (7.1)	65.2 (27.4)	68.3 (29.2)	-3.1 (9.8)
Newton	170.1 (81.6)	154.1 (82.2)	15.7 (30)*	287.1 (96.5)	284.1 (98.4)	2.2 (22.9)
Borg	5.3 (1.5)	4.9 (1.4)	0.4 (0.7)*	4.6 (1.1)	4.3 (0.7)	0.3 (0.7)

\*Paired *t*-test significant (*p*<0.05).

T1: mean value of the first assessment; T2: mean value of the second assessment; SD: standard deviation; ΔT: difference between T1 and T2; PILE: Progressive Isoinertial Lifting Evaluation.

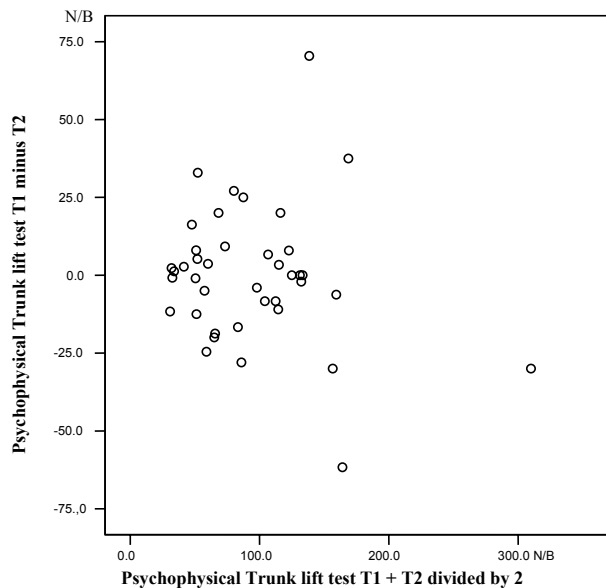


Fig. 2. Psychophysical trunk lift patients and healthy subjects. Difference between Psychophysical trunk lift T1 and T2 scatter plotted against T1+T2 divided by 2. N/B = Newton/Borg.

the 95% confidence interval (CI) are large, which suggests that, as in other studies, ICCs lower than 0.75 may be found. On the other hand, the ICC expresses how well 2 observations are likely to classify a patient consistently relative to other patients (32). The ICC value provides no indication of the magnitude of the disagreement between 2 observations (33). To determine the magnitude of disagreement on an individual level, the limits of agreement were calculated (29, 30). In other words, in an individual patient with non-specific CLBP, the changes due to treatment should exceed limits of agreement before deciding that true changes have occurred. Despite the acceptable reliability (ICC) the limits of agreement were substantial for the psychophysical lift capacity tests in patients and healthy subjects. This means that the "noise" was relatively large. The limits of agreement of the psychophysical leg lift

were large ( $\pm 106$  N/B) in patients. This means that true change can be detected only when an individual changes at least 93% in comparison with 113.6 N/B (Table II). When the change is less than 93% after a treatment program, it may be due to natural variation. The limits of agreement of the psychophysical trunk lift capacity were  $\pm 40.0$  N/B in the patients. This means that true change can only be detected when an individual changes at least 59% in comparison with 68.3 N/B (Table II). When the psychophysical static trunk lift capacity in the patients increases by 59% they will almost reach the value of the healthy subjects. In a previous study we have shown that patients with CLBP increased by 87% in the psychophysical trunk lift and 98% in the psychophysical leg lift capacity test after a cognitive-somatic rehabilitation approach (16). For the psychophysical dynamic lifting capacity test this individual change has to be 42% in patients, which is also substantial. For the psychophysical dynamic lifting test the difference between patients and healthy subjects was relatively larger. When the patients increase by 42% the patients will not have reached the value of the healthy subjects. This may be because the psychophysical endurance of abdominal and back muscles is also important within the psychophysical dynamic lifting test, compared with the psychophysical static lift test. This is supported by the result of the PILE, that 17 of the healthy subjects and none of the patients ended the test due to the ceiling effect according to the study result of Brox et al. (34).

Large limits are the result of a substantial within-subject variation. This variation can be attributed to the testing procedure, or to other factors, such as the patient or healthy subjects, differences in test behaviour such as fear of pain, motivation or within-subject random errors. However, perceived exertion or an appropriate perception of physical activities is important in the reduction of the perceived disability (18). Table I shows that the Borg score is systematically higher in patients in both measurements compared with the healthy subjects, except for arm lift capacity tests.

No significant difference between patients and healthy subjects was found for the first psychophysical static arm lift capacity ( $p = 0.227$ ), which might be expected because

Table II. Intra-class correlation (ICC) (1-way random model) for patients and healthy subjects of the psychophysical capacity tests N/B

Psychophysical capacity test In N/B	ICC	95% CI	LOA in N/B	Mean of 2 measurements in N/B	LOA % of the mean of 2 measurements
Patients $n = 20$					
Leg lift	0.82	0.56–0.93	$\pm 106.0$	113.6	93
Trunk lift	0.83	0.62–0.93	$\pm 40.0$	68.3	59
Arm lift	0.89	0.79–0.96	$\pm 26.0$	56.9	46
PILE	0.93	0.83–0.97	$\pm 14.2$	33.8	42
Healthy $n = 20$					
Leg lift	0.78	0.44–0.91	$\pm 170.0$	184.1	92
Trunk lift	0.93	0.82–0.97	$\pm 42.0$	120.0	35
Arm lift	0.75	0.48–0.89	$\pm 52.0$	69.4	75
PILE	0.94	0.85–0.97	$\pm 19.6$	66.8	29

CI: confidence interval; LOA: limits of agreement; N/B: Newton/Borg; PILE: Progressive Isoinertial Lifting Evaluation.

patients did not experience arm pain and the lower back was held straight during lifting. The psychophysical dynamic lifting ( $p = 0.000$ ) and static trunk lift ( $p = 0.002$ ) and leg lift capacity ( $p = 0.007$ ) were significantly lower in patients with non-specific CLBP compared with healthy subjects, which might be expected.

In conclusion, the present study shows that the psychophysical static lift capacity and dynamic lifting capacity are reliable instruments for patients with non-specific CLBP and healthy subjects. However, a substantial amount of natural variation should be taken into account between 2 test sessions when interpreting the test results clinically.

## REFERENCES

1. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance test in patients with low back pain. *Spine* 1994; 19: 2759–2765.
2. Battie MC, Cherkin DC, Dunn R, Ciol MA, Wheeler KJ. Managing low back pain: attitudes and treatment preferences of physical therapists. *Phys Ther* 1994; 74: 219–226.
3. Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs. *Arch Phys Med Rehabil* 2002; 83: 44–47.
4. Stratford PW, Binkley J, Solomon P. Defining the minimal level of change for the Roland Morris Questionnaire. *Phys Ther* 1996; 76: 359–365.
5. Magee DM, editor. *Orthopedic physical assessment*. 2nd edn. Philadelphia: WB Saunders, 1992.
6. Potter NA, Rothstein JM. Interrater reliability for selected clinical tests of the sacroiliac joint. *Phys Ther* 1985; 65: 1671–1675.
7. Simmonds MJ, Kumar S. Health care ergonomics: part 2 location of body structures by palpation. A reliability study. *Int J Ind Ergonomics* 1993; 11: 145–151.
8. Newton M, Thow M, Somerville D, Henderson I, Waddell G. Trunk strength testing with iso-machines: part 2 experimental evaluation of the Cybex 2 back testing system in normal subjects and patients with chronic low back pain. *Spine* 1993; 18: 812–824.
9. Brouwer S, editor. *Disability in chronic low back pain. Psychometric properties of ADL- and work-related instruments*. Dissertation. Groningen: University of Groningen; 2004.
10. Verbunt J, editor. *Disuse and physical deconditioning in chronic low back pain*. Dissertation. Maastricht: University of Maastricht; 2004.
11. Jillings-Rohaam JJ, Gorter MJ, Buurke JH, Baten CTM, Hermens HJ. [Rugscholing in een revalidatiecentrum]. *Back Education in Rehabilitation Centre*. *Ned Tijdschr voor Fysiotherapie* 1997; 1: 14–18. (In Dutch).
12. Roland M, Morris R. A study of the natural history of low back pain. *Spine* 1983; 8: 141–150.
13. Gommans IHB, Koes BW. [Validiteit en responsiviteit Nederlandse Roland Disability Questionnaire]. *Validity and Responsiveness of the Dutch Roland Disability Questionnaire*. *Ned Tijdschr voor Fysiotherapie* 1997; 2: 28–33. (In Dutch).
14. Fordyce WE, Lansky D, Calshyn DA, Shelton JL, Stolov WC, Rock DL. Pain measurement and pain behaviour. *Pain* 1984; 18: 53–69.
15. Cunha IT, Simmonds MJ, Protas EJ, Jones S. Back pain, physical function and estimates of aerobic capacity: what are the relationships among methods and measures? *Am J Phys Med Rehabil* 2002; 81: 913–920.
16. Hodselmans AP, Jaegers SM, Göeken LNH. Short-term outcomes of a back school program for chronic low back pain. *Arch Phys Med Rehabil* 2001; 82: 1099–1105.
17. Khalil TM, Goldberg ML, Asfour SS, Moty EA, Rosomoff RS, Rosomoff HL. Acceptable Maximum Effort (AME). A psychophysical measure of strength in back pain patients. *Spine* 1987; 12: 372–376.
18. Hildebrandt J, Pflingsten M, Sauer P, Jansen J. Prediction of success from a multidisciplinary treatment program for chronic low back pain. *Spine* 1997; 22: 990–1001.
19. Chaffin DB, Herrin GD, Moroe Keyserling W. Pre-employment strength testing; an updated position. *J Occup Med* 1978; 20: 403–408.
20. Genaidy AM. Spinal compression tolerance limits for the design of manual material handling operations in the workplace. *Ergonomics* 1993; 36: 415–434.
21. Borg G. Psychophysical bases of perceived exertion. *Med Sci Sports Exercise* 1982; 14: 377–381.
22. Troup JDG, Foreman T, Baxter C, Brown D. The perception of back pain and the role of psychophysical tests of lifting capacity. *Spine* 1987; 12: 645–657.
23. Mayer TG, Barnes D, Kishino ND, Nichols G, Gathcel RJ, Mayer H, et al. Progressive isoinertial lifting evaluation: 1. A standardized protocol and normative database. *Spine* 1988; 13: 993–997.
24. Mayer TG, Barnes D, Nichols G, Kishino ND, Coval K, Piel B, et al. Progressive isoinertial lifting evaluation: 2. A comparison with isokinetic lifting in a disabled chronic low-back pain industrial population. *Spine* 1988; 13: 998–1002.
25. Fox S. Physical activity and the prevention of coronary heart disease. *Ann Clin Res* 1971; 3: 404–432.
26. Waters TE. Revised NIOSH equation for the design and evaluation of manual tasks. *Ergonomics* 1993; 7: 749–776.
27. Bland JM, Altman DG. Measurement error. *BMJ* 1996; 313: 744.
28. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 8: 307–310.
29. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; 32: 307–317.
30. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring quantitative variables. *Com Biol Med* 1989; 19: 61–70.
31. Tammemagi MC, Frank JW, LeBlanc M, Artsob H, Streiner DL. Methodological issues in assessing reproducibility – a comparative study of various indices of reproducibility applied to repeat ELISA serological tests for Lyme disease. *J Clin Epidemiol* 1995; 48: 1123–1132.
32. Altman DG, editor. *Practical statistics for medical research*. London: Chapman and Hall; 1991.
33. Evans WJ, Cayten CG, Green PA. Determining the generalizability of rating scales in clinical settings. *Med Care* 1981; 19: 1211–1220.
34. Brox JI, Storheim K, Holm I, Friis A, Reikeras O. Disability, pain, psychological factors and physical performance in healthy controls, patients with sub-acute and chronic low back pain: a case-control study. *J Rehabil Med* 2005; 37: 95–99.