J Rehabil Med 2003; 35: 69–75

Taylor & Francis healthsciences

PHYSICAL PERFORMANCE TESTS FOR PEOPLE WITH LONG-TERM SPINAL PAIN: ASPECTS OF CONSTRUCT VALIDITY

Therese Ljungquist,[1,2] Irene B. Jensen,[2] Åke Nygren[2] and Karin Harms-Ringdahl[1]

From the [1]Neurotec Department, Division of Physiotherapy and [2]Department of Clinical Neurosciences, Section for Personal Injury Prevention, Karolinska Institutet, Stockholm, Sweden

Objective: **The aim of this study was to investigate the construct validity of 6 physical performance tests that had already been shown to have acceptable repeatability.**
Design: **Data were collected in a randomized controlled multi-centre study.**
Subjects: **126 women and 105 men sick-listed for spinal pain carried out the tests and provided personal and background data at inclusion in the study.**
Methods: **One test measured stepping up onto and down from a stool, 2 measured lifting ability and 3 walking speed. Construct validity was examined by analysing the influence of some variables on test performance.**
Results: **High-rated pain behaviour and perceived high pain intensity during testing or during the previous 4 weeks were connected with low test performance. Exercise twice a week was connected with high test performance. The test with the highest ability to detect disability in the women with lumbar pain was a lumbar lifting test, while for the men, it was a cervical lifting test. The test with the highest ability to detect disability in the participants with neck pain was the cervical lifting test in addition to a gait test with burden for the women.**
Conclusion: **Back pain hampered the test performance more than neck pain. Impairments and activity limitations expressed by the patient should guide the choice of test.**

Key words: physical performance test, assessment, physical therapy, validity, impairment, activity limitation.

J Rehabil Med 2003; 35: 69–75

Correspondence address: Therese Ljungquist, Section for Personal Injury Prevention, PO Box 127 18, SE-112 94 Stockholm, Sweden.
E-mail: Therese.Ljungquist@kirurgi.ki.se

Submitted January 31, 2002; accepted August 20, 2002

INTRODUCTION

Spinal pain is the most common cause of long-term sick leave in Sweden (1) and is an increasing problem in the Western world. The 1-year prevalence of spinal pain has been reported to be around 66% in Sweden (2) and that of low back pain around 50% in the Netherlands (3). The lifetime prevalence for low back pain has been reported to be around 60% in Japan (4) and 84% in Canada (5), while the lifetime prevalence for neck pain in Canada has been reported as 67% (6). Physiotherapists are often involved in the rehabilitation of persons with long-term back or neck pain. They are expected to assess these persons' disability level, define their problems, plan interventions jointly with the patients and evaluate their progress during and after the interventions (7). Disability can be classified as impairment, activity limitation or participation restriction, according to the WHO classification (8). Opinions differ regarding the relative clinical importance of questionnaires and of "objective" tests for disability due to spinal pain. In agreement with Simmonds et al. (9), Hoeymans et al. (10) and Waddell et al. (11), we argue for the use of physical performance tests as complementary to patient self-reports as rated in questionnaires. There is to our knowledge, and as Riddle & Stratford stated in 1998 (12), still no single golden standard measure for assessing disability. It thus seems valuable to seek information concerning possible clinical assessment instruments as contributions to a future "assessment instrument bank" for physiotherapists.

Studies of the clinimetric properties of test packages including physical performance tests have earlier been conducted by Simmonds et al. (9) and Harding et al. (13). The present tests, originally 11 in number, were assembled for clinical use for evaluating disability due to back or neck pain at impairment level and at activity limitation level according to the WHO classification (8). They were chosen to supplement self-rated questionnaires in a national randomized, controlled, multi-centre study from which the current data were obtained (14).

Discriminative ability establishes one aspect of construct validity, i.e. how far a test measures what it is intended to on the basis of a theoretical framework. Clinically relevant distinctions in test performance, i.e. "cut-off values", between persons with neck or back pain and back-healthy persons were revealed for all 11 tests but 2 by calculation of sensitivity/specificity (15).

The present aim was to investigate the construct validity, as expressed in our hypotheses below, of the 6 tests shown to have acceptable reliability (16) in the current test package.

Hypotheses

- *Pain site* was hypothesized to be related to performance of the different tests. The lumbar lifting test, the stair-climbing test and the step-on-stool test were assumed to be more difficult for persons with back pain. The cervical lifting test and the walking test carrying a burden were assumed to be more difficult for persons with neck pain. The ordinary walking test was assumed to be equally difficult regardless of pain site.
- A lower overall performance was assumed when *pain behaviour* and *pain intensity* were high, when the participant

Table I. *The original 11 physical performance tests in the fixed order of testing*

1	Åstrand test
2	Neck flexor test
3	PILE lumbar test
4	Trunk 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

had *pain from more than 1 pain site*, and when *prolonged pain duration and sick leave* were present.

- A higher overall performance was assumed for persons who *exercised regularly*.
- The *exertion level* during tests was assumed to possibly be related to either low performance *or* high performance.

MATERIAL AND METHODS

Two hundred and thirty-five persons participating in a randomized, controlled multi-centre study were included (14). They were identified from the AGS insurance register, which is a sick and disability pension insurance scheme agreed between the Swedish Employer's Confederation (SAF) and the Swedish Trade Union Confederation (LO) and which covers 2.5 million employees in Sweden. The inclusion criteria were spinal pain, sick leave for between 1 month and 6 months for spinal pain, fluency in Swedish and age between 18 and 60 years. The participants were all considered to be suitable for a rehabilitation programme with a cognitive-behavioural approach, including physical exercise, according to a physician. Exclusion criteria were pregnancy, co-morbidity that could affect participation or results, involvement of nerve roots indicating surgical treatment, trauma resulting in verified fracture within the previous 6 months, and recent participation (<3 months) in a rehabilitation programme.

The study was approved by the Regional Research Ethical Committe, Karolinska Institutet, Stockholm, Sweden.

Measurements

The participants were asked to provide personal and background data.

They were examined by a physician according to a standardized status formula, including medical history taking, neurological examination and range of motion. Sick-leave data were obtained from the Swedish National Social Insurance Board.

Questionnaires

The participants answered a battery of questionnaires designed to cover the multidimensional problems often associated with long-term spinal pain. The questionnaires used in the present study were the Short-Form 36 (SF-36) (16) and a questionnaire about exercise habits during leisure time (frequency and intensity), modified from Engström et al. (18).

Physical performance tests

The tests were assembled by a group of physiotherapists employed in a rehabilitation company in Sweden, who were all experienced in rehabilitation of patients with musculoskeletal pain.

All participants but 4 performed the 11 tests on inclusion in the randomized study. The 6 tests which had acceptable reliability (16) were:

- *Muscular endurance in the lower extremity*, the "step-on-stool test". The person tested was asked to step up onto and down from a stool at a self-selected speed. The step height was 0.40 metres for women, 0.44 metres for men. The number of steps managed was recorded. The measurements were discontinued after 50 steps. The inclusion of the step-on-stool test was based on the clinical assumption that lower-extremity endurance is important when protecting your back, e.g. in lifting manoeuvres.
- *Self-selected walking speed* was measured in 3 ways. The subjects walked 20 metres at a comfortable speed along a corridor and turned around where 20 metres was marked. Thereafter, they repeated the procedure carrying 1 carrier bag in each hand, containing 4 kg each for women, 8 kg each for men. The time taken was recorded to allow calculation of the walking speed. The tests were discontinued after 50 seconds. The subjects were then asked to walk up and down a flight of stairs at a self-selected speed, preferably without support. The stairs had 18–20 steps, the number differing between clinics. To standardize the measurements, they were converted to metres of height per second (m/s). The test was discontinued after 35 seconds. We considered walking to be an important activity for all people, and chose self-selected speed as most resembling real life. We assumed that "comfortable speed" would be lower when the patient was in pain.
- *Lumbar and cervical lifting tests*, *PILE tests*, developed by Mayer and co-workers (19). The lifting tests were performed standing in front of bookshelves with shelves at 0.76 metres and 1.37 metres from the floor. The subject was asked to lift weights in a plastic box from floor to waist level (0–0.76 metres) for the lumbar lifting test, or from waist to shoulder height (0.76–1.37 metres) for the cervical lifting 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 4 such lifting movements (during approximately 20 seconds), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last "lifting movement" was measured. The tests were discontinued if the heart rate reached 85% of the estimated maximal level, adjusted for age. It was considered important to include lifting tests because many people report difficulties with lifting tasks. The PILE tests were chosen because they were well described and reproducible, and because they had showed sensitivity to change by Mayer et al. (19, 20).

Test procedure

The 11 tests were arranged to be performed consecutively without pausing (Table I). The subjects wore exercise clothing. The test leader asked them to try their hardest, but to take pain and fatigue into account. It was emphasized that they could discontinue at any time or decline a test completely. During the testing, the test leader gave any verbal instructions needed to standardize the test procedure, but did not encourage or distract the subjects in any way. After each test, the participants were asked to rate their pain intensity during the test on Borg's Category Ratio Scale, CR10 Scale, scales from 0 to 11, and the exertion they perceived during the test on Borg's Ratings of Perceived Exertion Scale, RPE Scale, scales from 6 to 20 (21) After completion of the test package, the participant's pain behaviour was rated on the University of Alabama Pain Behavior Scale, UAB Scale, scales from 0 to 10 (22), by the physiotherapist in charge.

Statistical methods

For the analyses we used the SPSS program (23). Only the 6 tests shown to have acceptable reliability were included in the analyses.

Sensitivity and specificity calculations were performed for the stair-climbing test for cut-off values between persons with spinal pain and "back-healthy" persons.

For our first hypothesis, Mann Whitney U tests were used for analysing differences between groups fulfilling conditions, in this case, having *neck pain as main complaint* or *back pain as main complaint*, as shown on a pain drawing (24). Additional analyses were made for persons with *neck pain only, lumbar pain only*, combined *pain in the neck and the back*, and *multiple pain sites*.

For our second hypothesis, we used Mann Whitney U tests to detect differences between groups fulfilling/not fulfilling the following criteria; *pain behaviour* rated as >3 on the UAB Scale (22) by the physiotherapist after conclusion of the tests, *self-rated pain intensity* on Borg's CR10 (21) more than "rather severe" (>4) during a particular test, *pain intensity during the previous four weeks* rated as "severe" or "very severe" (>4 on a 1–6 scale) on SF-36 (17), *pain duration exceeding 3*

Table II. *Demographic and background data for the study participants (*n *= 235)*

	Median (IQR[1])
Age (years)	46.0 (17)
Number of days on sick-leave 1 year before inclusion	144.0 (79.0)
Number of days on sick-leave 3 months before inclusion	90.0 (26.8)
Pain duration main pain site (months)	8.0 (26.0)
SF-36 physical function[2] (8 items, rating scale 1–3)	2.0 (0.5)
SF-36 general health[2] (3 items, rating scale 1–5)	3.0 (1.5)
SF-36 mental health[2] (5 items, rating scale 1–6)	4.0 (2.0)
SF-36 vitality[2] (5 items, rating scale 1–6)	2.5 (1.0)
SF-36 pain during the last 4 weeks[2] (rating scale 1–6)	2.0 (1.0)
SF-36 pain affecting work during the last 4 weeks[2] (rating scale 1–5)	2.0 (1.0)
Women, number (%)	129 (54.9)
Men	106 (45.1)
Main pain site (pain drawing, $n = 234$), number (%)	
Neck	92 (39.3)
Shoulder	1 (0.4)
Thoracic spine	4 (1.7)
Lumbar spine	107 (45.7)
Multiple pain sites	29 (12.4)
Lower extremity	1 (0.4)
Exercise habits in leisure time ($n = 219$), number (%)	
Hardly any exercise (never or irregularly)	37 (16.9)
Exercise at low intensity at least once a week	71 (32.4)
Exercise at least at medium intensity once a week	12 (5.5)
Exercise at least at medium intensity at least twice a week	99 (45.2)

[1] IQR = Inter-quartile range, i.e. the numerical difference between the first and the third quartile, the appropriate distribution measure for data not normally distributed.
[2] The higher value, the better the function/the less the pain.

Table III. *Test results expressed as median (IQR[1]) for persons with neck pain location as their main complaint (*n *= 90) compared with persons with lumbar pain location as their main complaint (*n *= 106). Median values reaching cut-off limit shown in italics.* p*-values of ≤0.016 in bold italics, representing the current limit for statistical significant difference on 5% level between groups*

Variable	Neck pain ($n = 90$)	Lumbar pain ($n = 106$)	*p*-value[4]
Step-on-stool test (steps), cut-off value 30 steps[2]			
Total	29.0 (34.8)	20.0 (26.0)	***0.010***
Women	28.0 (37.0)	18.0 (20.5)	0.047
Men	*30.0*[2] (31.0)	26.0 (42.0)	0.103
Gait test (m/s), cut-off value 1.3 m/s[2]			
Total	*1.31* (.20)	1.21 (0.22)	***0.003***
Women	1.29 (.20)	1.21 (0.19)	0.049
Men	*1.33* (.24)	1.25 (0.29)	0.019
Gait test with burden (m/s), cut-off value 1.3 m/s[2]			
Total	1.29 (.21)	1.18 (0.26)	***0.010***
Women	1.21 (.24)	1.21 (0.25)	*0.469*
Men	*1.33*[2] (.23)	1.18 (0.31)	***0.003***
Stair-climbing (m/s), cut-off value 0.29 m/s[3]			
Total	*0.30* (.09)	0.26 (0.08)	***0.001***
Women	*0.30* (.10)	0.26 (0.07)	***0.006***
Men	*0.31* (.09)	0.27 (0.08)	0.032
Lumbar lifting test (kg)			
Total	12.6 (10.2)	10.4 (9.0)	***0.014***
Women (cut-off value 12.6 kg[2])	10.4 (4.5)	8.1 (4.5)	0.051
Men (cut-off value 23.9 kg[2])	19.4 (7.9)	14.9 (11.3)	0.034
Cervical lifting test (kg)			
Total	5.9 (4.5)	8.1 (7.3)	0.039
Women (cut-off value 8.1 kg[2])	5.9 (4.5)	5.9 (3.9)	0.063
Men (cut-off value 19.4 kg[2])	10.4 (9.0)	10.4 (9.0)	0.218

[1] IQR = Inter-quartile range, i.e. the numerical difference between the first and the third quartile, the appropriate distribution measure for data not normally distributed.
[2] Cut-off values are given for each test, representing earlier shown best cut-off values for distinguishing between persons with spinal pain and those without (15).
[3] Cut-off value based on sensitivity and specificity calculations in the present study.
[4] *p*-values according to the Mann Whitney U tests, values <0.016 shown in bold italics.

months for the current pain condition, *sick leave exceeding 3 months*, and *pain from more than one site*, as shown on a pain drawing (24).

For our third hypothesis, the criterion used was reported *exercise at least twice a week* with at least moderate intensity, defined as talking to someone being still tolerably possible while exercising (18).

For the fourth hypothesis, *perceived exertion* rated as at least "heavy" (≥15) on Borg's RPE scale (21) was the criterion used.

The Mann Whitney U tests were performed first with all subjects included, and, secondly, with women and men separately. The significance levels were adjusted by Bonferroni corrections (25); a *p*-value of at most 0.016 for neck/lumbar pain, 0.006 for the ratings on CR10, RPE and UAB, and 0.003 for the remaining variables examined (sick leave, pain duration, >1 pain site, pain intensity according to SF-36, exercise level) were considered significant at the 5% level.

Regression analyses were used to further analyse the effect of the background factors on performance. Pain site, rated pain intensity on SF-36, pain duration, duration of sick leave and stated physical activity in leisure time were independent variables. The SF-36 ratings were turned into dummy variables to fit in the regression model, as described by Devore & Peck (26). Dummies were likewise used for pain site (neck pain as main complaint vs other pain sites, multiple pain sites vs one single pain site), sick leave, and for reported physical activity. Gender was included as a dummy variable to control for the known differences in performance for the lifting tests and for the step-on-stool test. Age was included to control for age-related differences. Due to the non-linear relation between age and the test measurements, age was turned into a dummy variable too. The test measurements were used as dependent variables.

Ordinal regression, backward method, a variation of logistic regression (27), was used for the step-on-stool test and the gait test with burden, due to distribution of data. Multiple regression, backward method, was used for the other dependent variables. All independent variables with a *p*-value of 0.1 or less were included in the final model.

RESULTS

Demographic and background data for the test participants are presented in Table II. The calculations of sensitivity and specificity revealed a stair-climbing speed of 0.29 m/s as the relevant cut-off point for persons with neck or back pain vs back-healthy persons.

Pain site

The measurements differed significantly for all tests except the cervical lifting test between participants with neck pain as their

Table IV. *Median values for persons reaching the dichotomizing limit vs persons not reaching the dichotomizing limit set for ratings of pain intensity (CR10), perceived exertion (RPE) and pain behaviour (UAB) during testing. Total* n *women = 126,* n *men = 105. The sample size differs between tests due to the differing numbers of persons rating over the dichotomizing limit set for each variable.* p-*values according to the Mann Whitney U tests, values ≤0.006, representing the current limit for statistical significant difference between groups shown in bold italics*

	CR-10			RPE			UAB		
Variable	≤4	>4[2]	*p*	<15	≥15[3]	*p*	≤3	>3[4]	*p*
Step-on-stool test[1] (steps)									
Total	30	10	<***0.001***	21	27	*0.570*	28	7	<***0.001***
Women	20	10	<***0.001***	18.5	19	*0.642*	24	5	<***0.001***
Men	39	11	<***0.001***	24	36	*0.070*	32.5	9	***0.002***
Gait test[1] (m/s)									
Total	1.25	1.14	*0.050*	1.25	0.98	***0.005***	1.29	1.11	<***0.001***
Women	1.25	1.04	*0.016*	1.25	1.00	***0.002***	1.25	1.14	<***0.001***
Men	1.29	1.23	*0.553*	1.29	1.22	*0.896*	1.33	1.10	<***0.001***
Gait test with burden[1] (m/s)									
Total	1.25	1.11	<***0.001***	1.25	1.11	<***0.001***	1.25	1.07	<***0.001***
Women	1.23	1.05	<***0.001***	1.21	1.07	<***0.001***	1.21	1.08	<***0.001***
Men	1.25	1.14	*0.032*	1.25	1.14	*0.129*	1.27	1.05	<***0.001***
Stair-climbing[1] (m/s)									
Total	0.28	0.22	<***0.001***	0.28	0.21	<***0.001***	0.28	0.22	<***0.001***
Women	0.28	0.21	<***0.001***	0.28	0.21	<***0.001***	0.28	0.21	<***0.001***
Men	0.28	0.23	*0.060*	0.28	0.22	*0.010*	0.29	0.23	***0.002***
Lumbar lifting test[1] (kg)									
Total	12.6	10.4	***0.001***	8.1	12.6	<***0.001***	10.4	8.1	<***0.001***
Women	10.4	8.1	*0.017*	8.1	10.4	*0.027*	10.4	5.9	***0.003***
Men	19.4	10.4	***0.001***	10.4	19.4	<***0.001***	19.4	10.4	*0.015*
Cervical lifting test[1] (kg)									
Total	8.1	5.9	***0.003***	8.1	8.1	*0.154*	8.1	5.9	***0.005***
Women	5.9	5.9	***0.003***	5.9	5.9	*0.542*	5.9	5.9	<***0.001***
Men	10.4	10.4	*0.034*	10.4	14.9	<***0.001***	10.4	10.4	*0.038*

[1] Cut-off values for the tests, which represents earlier shown best cut-off values for distinguishing between persons with spinal pain and those without (15), are given in Table III.
[2] *n* = depending on test; step-on-stool test: 28 women (W) 30 men (M), gait test: 9 W 9 M, gait test with burden: 34 W, 28 M, stair-climbing test: 18 W 13 M, cervical lifting test: 61 W 46 M, lumbar lifting test: 51 W 39 M.
[3] *n* = depending on test; step-on-stool test: 77 W 47 M, gait test: 7 W 2 M, gait test with burden: 24 W 8 M, stair-climbing test: 15 W 3 M, cervical lifting test: 50 W 30 M, lumbar lifting test: 74 W 61 M.
[4] *n* = 19 W, 16 M.

main complaint and those with their main complaint in the lumbar spine. For all those tests, the participants with mainly neck pain performed significantly better than those with mainly lumbar pain, women and men considered together (Table III). Their performance, as expressed in median values, in some tests reached the cut-off value distinguishing “patients” from “back-healthy persons” derived from the sensitivity/specificity figures presented in our earlier study (15). However, when considering individuals, at most 57% reached these cut-off values (data not shown). Women with neck pain had most problems with the gait test with burden and the cervical lifting test, while the women with lumbar pain had most problems with the lumbar lifting test. For the men, the cervical lifting test was most challenging for both pain sites.

Additional Mann Whitney U analyses

Persons with *lumbar pain only* ($n = 56$, 33 men and 23 women), had significantly lower performance on the step-on-stool test, the gait test and the stair-climbing test compared with persons with *neck pain only* ($n = 34$, 16 men and 18 women). Persons with *neck pain only* had significantly lower performance on the cervical lifting test. 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* had 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* reached the cut-off value distinguishing women with spinal pain from those without (15).

Neither persons who had combined pain in the neck and the back nor persons having multiple pain sites performed significantly different from others.

Other variables hypothesized to have significant effects on test measurements

For participants rated as showing relatively high *pain behaviour* (>3 on the UAB Scale), the measurements were significantly lower in all tests, women and men considered together, while for rated pain intensity more than “rather strong” on the CR10 Scale, this was true for all tests but the gait test.

The women rating *perceived exertion* during testing as at least “heavy” on the RPE Scale had significantly slower walking

Table V. *Background factors explaining part of the variation in test results* (n = *231) according to the linear multiple regressions*[1] *and ordinal regressions*[2] *are marked with a cross. The proportion of variation explained for each dependent variable in brackets after the variable name (adjusted* R^2 *for the multiple regressions and Nagelkerke's pseudo* R^2 *for the ordinal regressions). The direction of the effects on test performance is presented before each cross-marking, a minus sign indicating an inverse relationship*

Dependent variables	Independent variables: Age	Woman	Sick leave	Neck pain	>1 pain site	High exercise level	Pain previous 4 weeks >4
Step-on-stool test (20%)[2]	− X	− X	+ X	+ X		+ X	− X
Gait test (12%)[1]	− X	− X	+ X	+ X			
Gait test with burden (15%)[2]	− X			+ X			− X
Stair-climbing (20%)[1]	− X			+ X	− X		
Lumbar lifting test (27%)[1]	− X	− X		+ X	+ X		− X
Cervical lifting test (25%)[1]	X*	− X		− X			

* Participants between 32 and 38 years of age performed better than both younger and older.

speeds for all 3 gait tests, while the men managed more weight in both lifting tests (Table IV).

Participants with *more than one pain site* ($n = 141$, 56 men and 85 women) did not perform significantly worse than others.

The women that had rated their *pain intensity on SF-36* ($n = 146$, 65 men, 81 women) during the previous 4 weeks as "severe" or "very severe", had significantly poorer performance on the step-on-stool test and in lumbar and cervical lifting, while the men showed no significantly different performance.

The participants with *duration of pain* ($n = 211$, 95 men, 116 women) *or sick leave* ($n = 183$, 86 men, 97 women) exceeding 3 months performed comparably to the others on all tests.

The participants who reported *exercise at least twice a week* ($n = 99$, 40 men, 59 women) with at least moderate intensity managed significantly more steps on the step-on-stool test.

Regression analyses

The regression analyses revealed background factors explaining 11–27% (= adjusted R^2 for the multiple regressions and Nagelkerkes pseudo R^2 for the ordinal regressions) of the variation in the measurements. The highest explanation figures were revealed for the lifting tests (Table V).

Age, gender and neck pain as main complaint had most impact on test performance. Test performance was generally positively affected by young age (<45 years), but in the cervical lifting test, persons aged between 32 and 38 performed better than both younger and older persons. Being a woman negatively affected performance in 4 of the tests. Having neck pain as main complaint compared to having other pain sites as main complaints negatively affected measurements on the cervical lifting test, but positively affected all other test results.

A rating of a minimum of severe pain intensity on the SF-36 questionnaire, compared to at most moderate pain intensity, negatively affected measurements on the step-on-stool test, the gait test with burden and the lumbar lifting test. Sick leave more than 4 months had a positive effect on test measurements for the step-on-stool test, and sick leave for 6 months had a positive effect on the ordinary gait test. The only background factor included in the regressions but without significant influence on the measurements was pain duration.

DISCUSSION

Construct validity has been defined by Angoff (28) as reciprocal verification of the measuring instrument and the theory underlying the phenomenon it is meant to measure. The construct validity of an instrument or a test must be demonstrated within a particular theoretical context, as stated by Sim & Arnell (29). When constructing an instrument or a test, one naturally has ideas concerning what kind of data it should produce, and how these data should be interpreted. We decided to use each test for each person, regardless of main pain site, due to our clinical experience that pain in one area often affects other areas, too. Nevertheless, we believed the tests would be able to discriminate to some extent between persons with different pain sites. Our hypothesis was essentially confirmed in that the participants with predominantly neck pain had significantly higher stair-climbing speed and performed significantly more in the step-on-stool test than those with lumbar pain as their main complaint. The participants with neck pain as their main complaint also had less lifting ability from waist to shoulder than those with predominantly lumbar pain, although not significantly less.

What we did not expect was that the neck-pain participants would have significantly higher walking speed on level ground, without as well as with burden (Table III). The relationship between low-back pain and slow walking speed has, however, been described earlier by Cheng et al. (30) and Arendt-Nielsen et al. (31).

Even if these results suggest the performance tests to be less discriminative for neck pain than for lumbar pain, neck pain *did* affect the test measurements for many participants. The inter quartile ranges presented in Table III clearly show the great variation in test results. The finding that low-back pain affects physical function more than does cervical pain has been described earlier by Jette & Jette (32) and Toomey et al. (33).

The analyses of differences in test performance between persons with *neck pain only* and persons with *lumbar pain only* were done for checking tendencies, and should be interpreted with some cautiousness due to the limited sample size. The tendencies, however, were in the expected direction, that is, persons with *lumbar pain only* had lower performance on all tests but the cervical lifting test, on which persons with *neck pain only* had significantly lower performance. The main pain site, thus, seemed to be adequate for predicting the test performances in our sample.

Pain behaviour rated as more than "3" on the UAB Scale turned out to be connected with a low test performance. As can be seen in Table IV, the ratings of pain behaviour exceeded 3 points for only 35 persons, and the median value, as well as the mean, was 2.0. In an earlier Swedish study by Lindström and co-workers, the mean value for 49 persons sick-listed for low back pain participating in a rehabilitation programme was 2.3 (34), which supports our UAB ratings being as high as could be expected in this kind of sample. The finding that pain behaviour was related to physical performance is in accordance with earlier studies by Fishbain et al. (35) and Hirsch et al. (36), and is not surprising from a clinical point of view.

Regarding rated pain intensity during testing, the results showed that women who rated pain intensity as more than "rather strong" had significantly lower performance on most tests, and men on the step-on-stool test and the PILE lumbar test. A study by Solem-Bertoft et al. (37) showed a strong association between pain rating on Borg's CR10 scale (21) and disability as measured by physical performance tests in people with a humerus fracture. Kuukkanen & Mälkiä (38) showed in a controlled treatment study of 90 persons with chronic low back pain that muscular endurance was lower in subjects with high pain intensity (Borg >4) during testing.

High perceived exertion was assumed to possibly be related to either low performance *or* high performance: *if* a test was overwhelmingly physically demanding for a participant, performance was assumed to be low and perceived exertion high; *if*, on the other hand, a participant was capable of a good performance on a physical demanding test, perceived exertion was assumed to be high connected with a *high* test result. High-rated perceived exertion had, as hypothesized, a positive relationship to performance in the more physically demanding lifting tests for men. Jacobs and co-workers have earlier noted that subjects managing many lifts are more likely to perceive high exertion (39). In a study by Dedering et al. (40), perceived exertion correlated positively to endurance time in the Sörensen test (41) for "back-healthy" persons. For the gait tests, high perceived exertion, instead, associated negatively to speed, indicating that for persons experiencing even a "simple" gait test as physically demanding, the speed was lowered. We conclude that ratings of perceived exertion could contribute to the overall understanding of the test participants – they reveal information about a person's motivation and overall energy.

From our data, we could not state a causal relationship between these ratings and test performance. The interrelations between rated pain behaviour, self-rated pain intensity and perceived exertion on one hand and performance on the other need further investigation.

Concerning pain duration, the absence of effect on performance could be because most participants' pain had persisted for at least 3 months. A lack of relationship between pain duration and other illness characteristics has been described by Toomey et al. (33), even though findings of a positive relationship would clinically seem more logical.

The hypothesis that participants with sick leave exceeding 3 months would perform worse was not confirmed. On the contrary, participants on sick leave for more than 4 months performed better on the step-on-stool test, and participants sick-listed for more than 6 months performed better on the gait tests on plain ground. One explanation might be an acquired coping ability learned from the protracted spinal pain – these people could have learned to cope, realizing that they could be active despite their pain. Sick leave is a composite variable, affected by many factors inside and outside the individual.

Since the variation explained by the independent variables in the regression models was only between 11 and 27%, other factors must have been involved. Our interpretation is that the individual's real physical capacity in combination with experienced pain intensity and exertion, and rated pain behaviour are among these factors.

Our theoretical framework, leading to the construction of the current test package, was essentially confirmed. However, the hypotheses concerning the effects of multiple pain sites, pain duration and duration of sick leave were not confirmed.

The interpretations of the Mann Whitney U tests together with the regression analyses are summarized as follows:

- Lumbar pain hampered performance on the physical performance tests more than neck pain did. This was confirmed in the regression analyses, where neck pain as the main problem negatively affected performance on the cervical lifting test, but positively affected all other tests compared with other main problem sites.
- High-rated pain behaviour was connected with low performance.
- High ratings of pain intensity during testing were connected with low performance in most tests.
- Pain intensity ratings on SF-36 >4 on a 1–6 scale were for women associated with low performance on the more physically demanding tests: the step-on-stool test and the 2 lifting tests. Regression analyses showed that high ratings on SF-36 also affected the gait test with burden.
- Those who exercised at least twice a week performed significantly better on the step-on-stool test.
- Sick leave for more than 3 months did not affect performance negatively, but on the contrary affected some tests positively.
- Most of the variability in our measurements could not be explained by the background factors analysed in the regression analyses.

When assessing body functions and activities for people with long-term spinal pain, all the tests proposed here seem to be useful. The test with the highest ability to detect a low performance for the participants with *neck pain* was the cervical lifting test, and in addition the gait test with burden for the women. The test with the highest ability to detect a low performance for the women with *lumbar pain* was the lumbar lifting test, while for the men, it was the cervical lifting test. Impairments and activity limitations expressed by the patient should guide the choice of test.

ACKNOWLEDGEMENTS

Many thanks to Mathias Nilsson and Stefan Stark for valuable advice on statistics. Financial support was given by the AFA Insurance Company, Sweden, and the Research Committee for Health and Caring Sciences, Karolinska Institutet, Stockholm, Sweden.

REFERENCES

1. National Board of Health and Welfare. Sweden's Public Health Report 1997. ISBN 91-7201-240-4.
2. Linton SJ, Hellsing A-L, Halldén K. A population-based study of spinal pain among 35-45-year-old individuals. Prevalence, sick leave, and health care use. Spine 1998; 23: 1457–1463.
3. Picavet HS, Schouten JS, Smit HA. Prevalence and consequences of low back problems in the Netherlands, working vs non-working population, the MORGEN-Study. Monitoring project on risk factors for chronic disease. Public Health 1999; 113: 73–77.
4. Matsui H, Maeda A, Tsuji H, Naruse Y. Risk indicators of low back pain among workers in Japan. Association of familial and physical factors with low back pain. Spine 1997; 22: 1242–1248.
5. Cassidy JD, Carroll LJ, Coté P. The Saskatchewan Health and Back Pain Survey. The prevalence of low back pain and related disability in Saskatchewan adults. Spine 1998; 23: 1860–1867.
6. Coté P, Cassidy JD, Carroll L. The Saskatchewan Health and Back Pain Survey. The prevalence of neck pain and related disability in Saskatchewan adults. Spine 1998; 23: 1689–1698.
7. WCPT. The Description of Physical Therapy, 1999. http://www.wcpt.org/briefing/index.html
8. ICF. International Classification of Functioning, Disability and Health. Geneva, World Health Organisation; 2001.
9. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine 1998; 23: 2412–2421.
10. 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). J Clin Epidemiol 1996; 49: 1103–1110.
11. Waddell G, Newton M, Henderson I, Somerville D. Letter. Spine 1993; 18: 938–939.
12. Riddle DL, Stratford PW. Use of generic versus region-specific functional status measures in patients with cervical spinal disorders. Phys Ther 1998; 78: 951–963.
13. Harding VR, de 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: 367–375.
14. Jensen I, Bergström G, Ljungquist T, Bodin L, Nygren ÅL. A randomized controlled component analysis of a behavioral medicine rehabilitation program for chronic spinal pain: are the effects dependent on gender? Pain 2001; 91: 65–78.
15. Ljungquist T, Fransson B, Harms-Ringdahl K, Björnham Å, Nygren Å. A physiotherapy test package for assessing back and neck dysfunction – discriminative ability for patients versus healthy control subjects. Physiother Res Int 1999; 4: 123–140.
16. Ljungquist T, Harms-Ringdahl K, Nygren A, Jensen I. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. Physiother Res Int 1999; 4: 214–232.
17. Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form Health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30: 473–483.
18. Engström L-M, Ekblom B, Forsberg A, von Koch M, Seger J. LIV 90, Rapport 1. Livsstil-Prestation-Hälsa. Folksam 1993 (in Swedish). ISBN 91-7044-174-X, Suppl A1.
19. Mayer TG, Barnes D, Kishino ND, Nichols G, Gatchel RJ, Mayer H, et al. Progressive isoinertial lifting evaluation – I. A standardized protocol and normative database. Spine 1988; 13: 993–997 (Published erratum Spine 1990; 15: 5).
20. Mayer TG, Barnes D, Nichols G, Kishino ND, Coval K, Piel B, et al. Progressive isoinertial lifting evaluation – II. A comparison with isokinetic lifting in a disabled chronic low-back pain industrial population. Spine 1998; 13: 998–1002.
21. Borg G. Borg's Perceived Exertion and Pain Scales. Human Kinetics, Campaign, Ill, USA; 1998.
22. Richards JS, Nepomuceno C, Riles M, Suer Z. Assessing pain behavior: The UAB Pain Behavior Scale. Pain 1982; 14: 393–398.
23 SPSS Base 9.0. User's guide. 1999, SPSS Inc., USA, ISBN 0-13-020390-4.
24. Mann NH III, Brown MD, Hertz DB, Enger I, Tompkins J. Initial-impression diagnosis using low-back pain patient pain drawings. Spine 1993; 33: 49–51.
25. Bland M. An Introduction to Medical Statistics. 3rd edition. Oxford: Oxford University Press; 2000.
26. Devore J, Peck R. Statistics: the exploration and analysis of data. 2nd edition. Belmont, California: Duxbury Press; 1993, p. 669–671.
27. Agresti A. Categorical Data Analysis. New York: John Wiley & Sons Inc.; 1990.
28. Angoff W. Validity: An evolving concept. In: Wainer H, Braun HI, eds. Test validity. Chapter 2, Hillsdale New Jersey, LEA Publishers; 1988, p. 28.
29. Sim J, Arnell P. Measurement validity in physical therapy research. Phys Ther 1993; 73: 102–115 (including commentaries).
30. Cheng CK, Chen HH, Chen CS, Lee SJ. Influences of walking speed change on the lumbosacral joint force distribution. Biomed Mater Engin 1998; 8: 155–165.
31. Arendt-Nielsen L, Graven Nielsen T, Svarrer H, Svensson P. The influence of low-back pain on muscle activity and coordination during gait: a clinical and experimental study. Pain 1995; 64: 231–240.
32. Jette DU, Jette AM. Physical therapy and health outcomes in patients with spinal impairment. Phys Ther 1996; 76: 930–945.
33. Toomey TC, Gover VF, Jones BN. Site of pain: relationship to measures of pain description, behavior and personality. Pain 1984; 19: 389–397.
34. Lindström I, Öhlund C, Nachemsson A. Physical performance, pain, pain behavior and subjective disability in patients with subacute low back pain. Scand J Rehabil Med 1995; 27: 153–160.
35. Fishbain DA, Abdel-Moty E, Cutler R, Khalil TM, Sadek S, Rosomoff RS, et al. Measuring residual functional capacity in chronic low back pain patients based on the dictionary of occupational titles. Spine 1994; 19: 872–880.
36. 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.
37. Solem-Bertoft E, Lundh I, Westerberg C-E. Pain is a major determinant of impaired performance in standardized active motor tests. A study in patients with fracture of the proximal humerus. Scand J Rehabil Med 1996; 28: 71–78.
38. Kuukkanen T, Mälkiä E. Muscular performance after a 3 month progressive physical exercise program and 9 month follow-up in subjects with low back pain. A controlled study. Scand J Med Sci Sports 1996; 6: 112–121.
39. Jacobs I, Bell DG, Pope J. Comparison of isokinetic and isoinertial lifting tests as predictors of maximal lifting capacity. Eur J Appl Physiol 1988; 57: 146–153.
40. Dedering A, 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. Clin Biomech 1999; 14: 103–111.
41. Biering-Sörensen F. Physical measurements as risk indicators for low-back trouble over a one-year period. Spine 1084; 9: 106–119.