J Rehabil Med 2012; 44: 851–857

ORIGINAL REPORT

INTER-TESTER RELIABILITY OF DISCRIMINATORY EXAMINATION ITEMS FOR SUB-CLASSIFYING NON-SPECIFIC LOW BACK PAIN

Evdokia Billis, PhD[1,2], Christopher J. McCarthy, PhD[3], John Gliatis, MD[4], Matthiew Gittins, PhD[2], Maria Papandreou, PhD[5] and Jacqueline A. Oldham, PhD[2]

From the [1]Department of Physiotherapy, Technological Educational Institute (T.E.I.) of Patras, Patras, Greece, [2]The University of Manchester, Manchester, UK, [3]Imperial College Healthcare NHS Trust, London, UK, [4]Orthopaedic Department, University Hospital of Patras, Patras, Greece and [5]Department of Physiotherapy, Technological Educational Institute (T.E.I.) of Athens, Athens, Greece

Objective: **To investigate the inter-tester reliability of a non-specific low back pain examination procedure, for sub-classifying non-specific low back pain.**
Design: **Reliability study.**
Participants: **Thirty patients with non-specific low back pain (12 males, 18 females, mean age 27.7 years, standard deviation 10.3) and 7 physiotherapists (raters).**
Methods: **Based on a health professionals' consensus via focus groups and a Delphi servey, an examination procedure was developed comprising 206 items discriminatory for non-specific low back pain, 108 of which were from the History (clinical questions) and 98 from the Physical Examination (clinical tests) section. Utilizing this procedure, each patient was examined by a blinded pair of raters.**
Results: **Moderate to excellent agreement was obtained in 125 (61%) items (77 History and 48 Physical Examination items), 47 of which obtained substantial or excellent agreement (kappa > 0.61), 37 moderate agreement (kappa between 0.41 and 0.6), and 41 excellent percentage agreements. Poor reliability (kappa < 0.41) was yielded in the remaining 81 items (31 History and 50 Physical Examination items).**
Conclusion: **Satisfactory reliability was obtained in nearly two-thirds of History and half of the Physical Examination items on a non-specific low back pain assessment list generated through consensus agreement. These findings provide clinicians and researchers with valuable information regarding which items are considered reliable and can be utilized in non-specific low back pain patient evaluation/assessment procedures, classification attempts and clinical trials.**

Key words: reliability; inter-tester; clinical items; history; physical examination; non-specific low back pain.

J Rehabil Med 2012; 44: 851–857

Guarantor's address: Evdokia Billis, Lecturer in Physiotherapy, Department of Physiotherapy, Branch Department of Aigion, Technological Educational Institute (T.E.I.) of Patras, Psaron 6, Aigion, 25100, Greece. E-mail: ebillis@teipat.gr

Submitted February 15, 2011; accepted November 22, 2011

INTRODUCTION

Classification systems are probably the most popular diagnostic approach presently used in non-specific low back pain (NSLBP). Assigning patients with NSLBP into homogenous subgroups based on their clinical presentation is believed to be the optimal way to overcome the diagnostic difficulties encountered with this heterogeneous population. Furthermore, it has been proven that classification systems enhance treatment outcomes more successfully than other management strategies (1–3).

However, despite their advantages (2–4) classification systems have recognized shortcomings in clinical practice, including their subjective nature and their unknown or questionable reliability. Most classification systems include clinical items and subgroups that have not been developed on the basis of a wider professional consensus. This may lead to bias and a system that is not clinically feasible or generalizable (3, 5, 6). In addition, although the reliability of classification systems has been addressed in some studies (7–14), the particular clinical items included are often of unknown or poor reliability. This can result in not knowing whether subgroup classifications are genuine or a product of failure in discriminatory ability resulting from poor reliability. Thus, although there is emerging evidence that classification systems are probably the optimal method for diagnosing and guiding treatment for NSLBP, basic steps towards their evaluation appear to require further elaboration. In particular, reliability is vital for improving the confidence of the clinical items included in a classification system and is also a prerequisite for validity testing (15). Thus, it is important to incorporate reliable items within the sub-grouping process.

In an attempt to improve the selection of examination items for future classification of NSLBP, a reliability study was conducted. The study investigated the inter-tester reliability of an extensive clinical list comprising items that were considered discriminatory for NSLBP, which were developed by a large consensus of health professionals (16, 17). This process is believed to improve upon previous studies by providing information on the reliability of a set of examination items considered *discriminatory* for NSLBP through consensus. Those items proving reliable could then be further utilized to identify patients with similar (homogenous) characteristics.

This is the first of two studies to investigate the inter-tester reliability of a consensus-agreed list of potentially discrimina-

tory items for the clinical assessment of patients with NSLBP. The second accompanying study (18) assesses whether the clinical items proven to be the most reliable can identify homogenous subgroups in a sample of patients with NSLBP.

METHODS

Sample

A convenience sample of Greek adult patients with NSLBP recruited via local physiotherapy referrals was invited to participate in the study. Patients were excluded if their low back pain (LBP) was due to specific pathology, they had undergone lumbar surgery, were pregnant, or if they had a severe neurological condition (influencing their cognitive and motor performance). Overall, 30 patients consented to participate. In addition, 7 physiotherapists (5 men, 2 women), experienced in treating LBP agreed to perform the assessments, which took place in 3 physiotherapy clinics in Greece, situated in Athens, Patras and Lamia. Ethical approval for reliability testing was obtained from the ethics committees of the Technological Educational Institute (TEI) of Lamia, Greece and the University of Manchester, UK.

Procedure

The process of developing the items being tested for reliability has been described previously (16, 17). Briefly, clinical features considered *discriminatory* for assessing and sub-classifying NSLBP were developed following 3 focus groups and a 2-round Delphi servey involving 23 health professionals (physiotherapists and doctors) and a representative stratified random sample of 150 physiotherapists (PTs), respectively. These clinical features were then transformed into clinically applicable questions and tests. This process was undertaken by 2 manual therapists experienced in spinal problems, utilizing an extensive content analysis procedure. Thus, this led to the development of a comprehensive, clinically applicable and discriminatory clinical examination list for assessing NSLBP. The list was separated into 2 sections, namely History and Physical Examination. The History section comprised 108 items and the Physical Examination 98 items. These clinical features comprising the examination guide are summarized in the results presented in Table II, Tables SI–SII (available from http://www.medicaljournals.se/jrm/content/?doi=10.2340/16501977-0950) . Further details regarding each question are presented in Billis (19). Prior to reliability testing, a training procedure took place to ensure a basic level of standardization and comprehension of operational definitions amongst PTs. Training lasted 4 h and was supported by a booklet, which summarized all important (key) examination points.

For reliability testing, raters were divided into pairs, in which the principal investigator was always one of the people within each pair. Thus, 6 PT pairs were formed, with each pair examining 5 patients. Every patient was simultaneously assessed in random order by the 2 therapists, with a 10–15 min break between the 2 examinations. PTs were instructed to advise their patients to rest should pain became intolerable. In addition, self-reported outcome measures for evaluating pain, physical disability, and psychosocial status were administered (summarized in Table I). Each rater was blind to the other raters' assessments as well as each patient's outcome measure scores. The whole procedure lasted approximately 1.5 h (approximately 30–40 min for each rater).

Data analysis

Data were analysed utilizing the kappa coefficient of concordance for nominal level data and weighted kappa coefficients with equal weighting amongst the point scales for ordinal level data (20, 21). Kappa values usually range from 0 to 1, where 0 accounts for no agreement and 1 for excellent agreement, although negative kappas may also be obtained, representing worse than chance agreement (21). Additional subcategories have been suggested, where 0 implies poor agreement, 0.01–0.2 slight agreement, 0.21–0.4 fair agreement, 0.41–0.6 moderate agreement, 0.61–0.8 substantial agreement, and 0.81–1.0 almost perfect agreement (22).

Table I. *Characteristics and outcome measure score for the patients (*n = *30)*

Characteristics	
Sex, *n* (%)	
Male	40 (12)
Female	60 (18)
LBP, *n* (%)	
Acute low back pain (<6 weeks)	70 (21)
Marital status, *n* (%)	
Married/living with partner	26.7 (8)
Single/divorced/widowed	73.3 (22)
Type of occupation, *n* (%)	
Sedentary	87.7 (26)
Active/manual	13.2 (4)
Pain, mean (SD)	
VAS – Present pain intensity	3.03 (2.27)
Disability, mean (SD)	
RMDQ	6.33 (4.58)
ODI	18.96 (13.1)
Psychosocial, mean (SD)	
FABQ – Work	18.93 (10.53)
FABQ – Physical activity	15.17 (5.72)
HADS – Anxiety subscale	8 (4.68)
HADS – Depression subscale	3.9 (2.72)

LBP: low back pain; SD: standard deviation; VAS: visual analogue scale (0–10); RMDQ: Roland-Morris Disability Questionnaire (0–24); ODI: Oswestry Disability Index (0–100); FABQ: Fear-Avoidance Beliefs Questionnaire (FABQ Work: 0–42, FABQ Physical activity: 0–24); HAD: Hospital Anxiety and Depression Scale (HAD subscales: 0–21).

For some items, which presented almost perfect agreement and minimal variability across therapists, their kappa values could not be calculated. In cases where negative (rather than positive) findings were too high for both testers (i.e. "saddle anaesthesia" was reported in only two cases by the one therapist and in no cases by the second), the classic 2 × 2 contingency table was not formed, and subsequently, "meaningful" kappas could not be calculated (23). For such items percentage agreements on all paired ratings were calculated instead. In addition, 95% confidence interval (CI) were calculated for all items. Analysis was performed utilizing SPSS (version 15.0).

RESULTS

Thirty patients with NSLBP (12 males, 18 females) with a mean age of 27.7 years (standard deviation (SD) 10.3, range 19–58) were examined. Seventy percent of patients had back pain of less than 6 weeks' duration. The sample's profile is summarized in Table I.

The 7 clinical physiotherapists (5 men, 2 women) performing the examinations had a mean clinical experience of 11.8 years (range 7–19) in treating patients with LBP, and 4 of them were musculoskeletal specialists.

A total of 206 clinical items were included in the reliability analysis. Kappa values ranged from –0.050 to 1 and weighted kappa values from –0.168 to 0.665. Eighty-eight items demonstrated moderate to perfect agreement (values over 0.41), fair agreement (values between 0.21 and 0.41) was achieved in 34 items, whereas 47 items demonstrated no agreement between the pairs of therapists, 26 of which presented with

Table II. *Items with substantial or excellent reliability*[a] *(*n = *88)*

History items	Response format	Kappa	Weighted kappa	Percentage agreement	Lower 95% CI	Upper 95% CI
Present symptoms						
Left-sided back pain	Yes/No	0.758			0.440	1.000
Right-sided back pain	Yes/No	0.911			0.741	1.000
Left buttock pain	Yes/No	0.615			0.361	0.869
Left posterior thigh pain	Yes/No	0.627			0.295	0.959
Right posterior thigh pain	Yes/No	0.783			0.374	1.000
Right posterior calf pain	Yes/No			100		
Right foot sole pain	Yes/No			100		
Left upper back pain	Yes/No			100		
Right upper back pain	Yes/No			100		
Abdominal pain	Yes/No	1.000			1.000	1.000
Right anterior leg pain	Yes/No			100		
Left anterior leg pain	Yes/No	1.000			1.000	1.000
Right foot pain in dorsum	Yes/No			100		
Left foot pain in dorsum	Yes/No	1.000			1.000	1.000
Anterior chest pain	Yes/No			100		
Type of pain – Dull ache	Yes/No	0.714			0.457	0.972
Type of pain – Intense pain	Yes/No	0.645			0.379	0.910
Type of pain – Superficial	Yes/No	0.902			0.714	1.000
Type of pain – Sharp/acute	Yes/No	0.733			0.492	0.974
Type of pain – Diffuse	Yes/No	0.706			0.396	1.000
Predominant pain – in the leg	Yes/No	1.000			1.000	1.000
Predominant pain – in the back	Yes/No	0.630			0.158	1.000
Relieving position/motion – Bending	Yes/No	0.783			0.374	1.000
Relieving position/motion – Straightening	Yes/No			96.7		
Relieving position/motion – Sitting	Yes/No	0.714			0.348	1.000
Relieving position/motion – Standing	Yes/No	1.000			1.000	1.000
Relieving position/motion – Lying	Yes/No	0.814			0.566	1.000
Relieving position/motion – Staying still	Yes/No			96.7		
Aggravating position/motion – Sitting	Yes/No	0.648			0.368	0.928
Aggravating position/motion – Walking	Yes/No			96.7		
Aggravating position/motion – Sit to stand	Yes/No			96.7		
Pain status – Getting better	Yes/No	0.730			0.486	0.974
Pain status – Getting worse	Yes/No	0.667			0.319	1.000
24-h pain pattern – Waking at night	Yes/No	0.889			0.676	1.000
24-h pain pattern – Worse in the morning	Yes/No	0.772			0.533	1.000
24-h pain pattern – Worse in the evening	Yes/No	0.722			0.533	1.000
Stiffness	Yes/No	0.675			0.430	0.921
Pins and needles	Yes/No	0.683			0.396	0.969
Clumsiness	Yes/No			100		
Dragging feet	Yes/No	0.634			0.178	1.000
History of condition						
Acute or chronic low back pain	Yes/No	0.823			0.633	1.012
First low back pain episode	Yes/No	0.609			0.208	1.000
Investigations – Radiographs (X-rays)	Yes/No	0.732			0.488	0.976
Investigations – Blood tests	Yes/No	0.684			0.402	0.967
Investigations – MRI	Yes/No	0.889			0.676	1.000
Investigations – Other	Yes/No	0.712			0.335	1.000
Function						
Occupation – Sedentary	Yes/No	0.738			0.507	0.969
Hobbies – Being severely affected	Yes/No	0.714			0.457	0.972
Medical history						
Red Flags – Saddle anaesthesia	Yes/No			93.3		
Red Flags – Bladder/bowel	Yes/No			100		
Red Flags – Anorexia	Yes/No			100		
Red Flags – Unexplained weight loss	Yes/No			100		
Red Flags – Night pain	Yes/No			90		
Red Flags – Intense unremitting pain	Yes/No			97		
Deformity (i.e. scoliosis)	Yes/No	0.689			0.420	0.959
Neck pain	Yes/No	0.865			0.686	1.000
Leg length inequality	Yes/No	0.783			0.374	1.000
Previous surgery	Yes/No	0.714			0.348	1.000
Postnatal backache	Yes/No	0.651			0.021	1.000

Table II. Contd.

Physical Examination items	Response format	Kappa	Weighted kappa	Percentage agreement	Lower 95% CI	Upper 95% CI
Observation						
Posture – Kyphotic	Yes/No			97		
Posture – Sway back	Yes/No			86.7		
Posture – Scoliotic	Yes/No			97		
Posture – Antalgic	Yes/No			97		
Gait – Normal	Yes/No			93.3		
Gait – Antalgic	Yes/No			100		
Gait – Trendelenburg	Yes/No			100		
Gait – Neurological	Yes/No			100		
Gait – Walking aids	Yes/No			100		
Facial expression – Normal	Yes/No			100		
Look in good health	Yes/No			100		
Active movements						
Pain – Lumbar flexion	Yes/No	0.769			0.523	1.000
Neurological examination						
L2 sensation	4-point Likert		0.667		0.048	1.000
L4 sensation	4-point Likert		0.665		0.205	0.933
L3 myotome	Yes/No			93.3		
L4 myotome	Yes/No			100		
L5 myotome	Yes/No			93.3		
S1 myotome	Yes/No			90		
S2 myotome	Yes/No			97		
Passive joint & palpation						
Hip pain – External rotation	Yes/No			93.3		
Hip pain – Internal rotation	Yes/No			97		
SI pain – Distraction test	Yes/No			100		
Postero-anterior pain – T12	Yes/No			97		
Postero-anterior pain – L2	Yes/No			93.3		
Allodynia	Yes/No			100		
Clinical reasoning analysis						
Movement pattern – impairment dysfunction	Yes/No	0.683			0.396	0.969
Primary pain mechanism involved	5-point Likert			100		
Predominant domain	3-point Likert	0.639			0.134	1.000
Prognosis	2-point Likert		0.634		0.178	1.000

[a]Kappa > 0.61.
CI: confidence interval; MRI: magnetic resonance imaging; SI: sacroiliac.

negative kappas, indicating worse than chance agreement. Percentage agreements in the remaining 41 items ranged from 86.7% to 100%, indicating almost perfect agreement. Excellent, moderate and poor reliability results from History and Physical Examination sections are presented in Table II, Tables SI–SII, respectively.

DISCUSSION

This study explored the inter-tester reliability of an examination list of clinical features perceived to be discriminatory in assessing back pain (16, 17). Such a study was considered necessary for identifying which clinical items are considered reliable for inclusion in future NSLBP classification attempts (see accompanying publication). Overall, the sample utilized had comparable demographics with previous reliability studies (7, 24, 25), and can be considered representative of typical NSLBP populations. In addition, all PTs (raters) had previous clinical experience with patients with LBP.

Moderate to excellent agreement was obtained in 125 (61%) out of the 206 items. Of these, 77 (37.4%) were from History section and 48 (23.3%) from Physical Examination. Kappa coefficients were calculated in all except 41 items, which were calculated with percentage agreements. Although high percentage agreements do not automatically assume acceptable reliability, they provide an indication of the consistency achieved amongst testers, are considered an appropriate reliability alternative for categorical data (23), and have been used extensively in previous studies (7, 25–27). Poor reliability results were yielded in 81 items (39.3%); 31 (15%) from history and 50 (24.3%) from physical examination, 26 of which reported negative kappas, indicating worse than chance agreement (21).

Although two-thirds of items obtained from History demonstrated satisfactory reliability, it is interesting to note that nearly one-third were not reliable. One could assume that simple questioning would result in consistent responses; however, this was not always the case. Whether this is attributed to the patient (i.e. lack of consistency, fatigue, change of

presentation between examinations) or to the rater is unknown. However, unlike specific LBP, for which there are studies that have investigated the consistency of history-taking (28, 29), within the NSLBP field there are only a few reports. Waddell et al. (26) found high percentage agreements for pain location and other symptoms, LBP onset and severity, diurnal pattern, function, disability as well as aggravating and easing factors. Pain location in the form of pain drawing was also highly repeatable in another study (30). These findings agree with the current study for the aforementioned items. In addition, in the study by McCarthy et al. (24), in which the reliability of clinical tests and questions included in international LBP guidelines were investigated in a large patient sample, several similarities to the present study were detected, with the exception of psychosocial items, which were found to be reliable in their study and unreliable in this one.

Within the Physical Examination, half of the items reported good reliability. Several previous studies have investigated the reliability of LBP physical examination items. This study's results on postural/gait observation agree with previous research reports indicating good reliability (12, 26, 31, 32). Active movements yielded more reliable results on assessing pain reproduction than range of movement (ROM). Overall, this agrees with most studies indicating that pain provocation testing is a more reliable assessment marker than ROM (23, 25, 27, 31). Centralization (33, 34) in flexion was a more reliable clinical indicator compared with extension. Although centralization is usually considered highly reliable for both flexion and extension (9, 33), the results in the current study were characterized as satisfactory, particularly considering that none of the raters had specialized (McKenzie-type) training.

Neurological examination yielded satisfactory reliability for myotomal testing (most) dermatomal testing and straight leg raise (SLR). However, reflex testing, ROM of SLR and some dermotomal tests were unreliable. Although neurological examination is usually reliable (23, 24, 26, 35), a mixed picture whereby some parts of this examination are reliable and others are not, has also been reported (32). The lower levels of agreement observed in this study may be partly attributed to variability in the Likert-type responses (3-point for SLR, 4-point for dermatomes and 5-point for reflexes), which could have confused the therapists. Another possible explanation could be that PTs were not thoroughly familiar with these tests in clinical practice, as medical doctors are the only "first-line" primary care practitioners screening for LBP in Greece, and are thus the ones predominantly utilizing neurological examination.

Passive examination was unreliable for evaluating ROM, which is in agreement with previous literature (25, 27). Pain provocation testing by palpation and passive joint examination was reliable for some spinal levels (T12, L1, L2 and S1) and lumbosacral areas (upper lumbar and sacroiliac). This, again, agrees with previous reports (25, 32). Lumbar pain provocation tests in the form of passive intervertebral motion testing have been extensively investigated, having an acceptable level of reliability; however, criticism has been levelled at the methodological quality of most studies (36). For sacro-iliac joint testing, this study's results agree with previous literature, indicating that reliability on individual tests is not considered good (37), and that testing should be performed on clusters of different tests (which are more reliable). Muscle testing was unreliable; whether this was attributed to the test, tester or patient (i.e. fatigue) factors is unknown.

Finally, reliability was examined in some clinical reasoning items, half of which obtained good results. From the clinical judgements on the patients' active movements, the existence of a closing pattern[1] and impairment dysfunction[2] were considered reliable. It could be argued that these two movement patterns are commonly encountered (38, 39), thus PTs are confident in their assessment. The primary pain mechanism and the predominant domain of influence and prognosis for recovery were also reliable amongst the judging therapists. These outcomes agree with a previous large-scale reliability study (24), thus providing some confidence in the PTs' clinical reasoning processes. Interestingly, behavioural signs were not deemed reliable, and this is in disagreement with McCarthy et al. (24), where similar items yielded moderate reliability.

This study has made an effort to provide a good standard reliability design by attempting to follow most of the suggestions for improving reliability studies that have been recommended by May et al. (27). However, a limitation regarding the sequential type of assessments performed by each therapist pair must be acknowledged. However common and acceptable this procedure may be (7, 24, 25, 31), it is a confounding factor in relation to the consistency of the clinical findings, in that it may either overestimate the consistency of the findings, or it may result in changes in the patient's clinical presentation (between examinations) should the patient become fatigued or aggravated by symptoms. This "biasing" effect was, however, reduced by randomizing the order of examination between therapists and by ensuring that the gap between examinations was not too short or too long (considering the patients' pain/disability levels) so as to dramatically change their presentation. In addition, although the sample's clinical profile was comparable to samples used in previous reliability studies (7, 24, 25), it should be acknowledged that it consisted of relatively acute, minimally disabled patients.

The similarities obtained between this and other reliability reports provide some confidence in the study's outcomes. In particular, the more "straightforward" aspects of the examination (i.e. history items, aggravating and easing factors, pain location, pain provocation testing, etc.) demonstrated higher reliability, similar to the findings of a number of previous studies (23–25,

[1]A closing (or compressive) pattern is evident when the patient's pain or symptoms are reproduced from the same side the movement is directed (i.e. left-sided pain with left-side flexion) (37).

[2]Impairment dysfunction refers to the loss of physiological motion (active or passive) due to pain. In such cases, motion is usually characterized by muscle guarding and co-contraction of the lumbopelvic muscles during the painful movement (38).

27, 31). Therefore, such clinical items with established reliability across studies could be recommended for use as a standardized approach, forming the basis for future clinical evaluation and clinical trials involving patients with NSLBP. The items that were proven to be unreliable across studies (i.e. motion palpation, muscle activation tests, sacroiliac testing, etc.), seem to form more "complex" examination processes, and should either be discarded or used with caution before further research indicates whether their reliability can be improved. Continuing to apply unreliable items in clinical practice/research may distort or compromise the true outcome value of the undertaken process. It is, however, interesting that some items that presented with poor reliability in this study, are considered significant markers or valuable prognostic indicators in other studies, such as several psychosocial and functional items (40, 41). Further research is required carefully to evaluate the overall value of these items as well as considering alternative approaches.

A strength of this study is that this examination list was developed on the basis of a large consensus of experienced clinicians, thus improving the generalizability of the examination process. In addition, consensus assures a degree of face validity of the involved questions and tests, making the examination process a practical, easily applicable tool for clinicians dealing with NSLBP. This approach enhances the clinical significance of the discriminatory items for patients with NSLBP, and the accompanying reliability statistics presented in this study facilitate health professionals to adopt this evaluation approach in their practice. However, it must be acknowledged that good reliability in examination testing does not guarantee validity in developing clinical subgroups. Further research should explore subgroup analysis. The accompanying study (Billis et al.) presents an exploration of the development of homogenous subsets in NSLBP by the use of a cluster analysis approach that utilizes the clinical items deemed reliable in the current study.

In conclusion, this study explored the inter-tester reliability of an examination procedure for NSLBP obtained from a consensus amongst Greek health professionals. Satisfactory reliability was obtained in nearly two-thirds of clinical questions obtained from the History section and in nearly half of the clinical tests obtained from the Physical Examination section. These findings provide clinicians' and researchers' insights into the clinical items considered reliable and that are recommended for inclusion in future NSLBP assessment procedures, sub-classification processes and clinical trials.

ACKNOWLEDGEMENTS

We thank physiotherapists Panagiotis Trigkas, Savvas Spanos, George Krekoukias and Ioannis Stathopoulos as well as all patients for participating in this study. This research was funded with the MACP's Elsevier Science and Doctoral Awards for Research in Manipulative Physiotherapy.

REFERENCES

1. Fritz JM, Cleland JA, Childs JD. Subgrouping patients with low back pain: evolution of a classification approach to physical therapy. J Orthop Sports Phys Ther 2007; 37: 290–302.
2. Fritz JM, Delitto A, Erhard RE. Comparison of classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain: a randomized clinical trial. Spine 2003; 28: 1363–1371.
3. Riddle DL. Classification and low back pain: a review of the literature and critical analysis of selected systems. Phys Ther 1998; 78: 708–737.
4. Brennan GP, Fritz JM, Hunter SJ, Thackeray A, Delitto A, Erhard RE. Identifying subgroups of patients with acute/subacute "non-specific" low back pain: results of a randomized clinical trial. Spine 2006; 31: 623–631.
5. Billis EV, McCarthy CJ, Oldham JA. Subclassification of low back pain: a cross-country comparison. Eur Spine J 2007; 16: 865–879.
6. McCarthy CJ, Arnall F, Strimpakos N, Freemont A, Oldham JA. The biopsychosical classification of non-specific low back pain: a systematic review. Phys Ther Rev 2004; 9: 17–30.
7. Harris-Hayes M, Van Dillen LR. The inter-tester reliability of physical therapists classifying low back pain problems based on the movement system impairment classification system. PM R 2009; 1: 117–126.
8. Heiss DG, Fitch DS, Fritz JM, Sanchez WJ, Roberts KE, Buford JA. The interrater reliability among physical therapists newly trained in a classification system for acute low back pain. J Orthop Sports Phys Ther 2004; 34: 430–439.
9. Kilpikoski S, Airaksinen O, Kankaanpaa M, Leminen P, Videman T, Alen M. Interexaminer reliability of low back pain assessment using the McKenzie method. Spine 2002; 27: E207–E214.
10. Paatelma M, Karvqnen E, Heinqnen A. Inter-tester reliability in classifying acute and subacute low back pain patients into clinical subgroups: a comparison of specialists and non-specialists. A pilot study. J Man Manip Ther 2009; 17: 221–229.
11. Petersen T, Olsen S, Laslett M, Thorsen H, Manniche C, Ekdahl C, et al. Inter-tester reliability of a new diagnostic classification system for patients with non-specific low back pain. Aust J Physiother 2004; 50: 85–94.
12. Razmjou H, Kramer JF, Yamada R. Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low-back pain. J Orthop Sports Phys Ther 2000; 30: 368–383.
13. Van Dillen LR, Sahrmann SA, Norton BJ, Caldwell CA, McDonnell MK, Bloom NJ. Movement system impairment-based categories for low back pain: stage 1 validation. J Orthop Sports Phys Ther 2003; 33: 126–142.
14. Vibe Fersum K, O'Sullivan PB, Kvale A, Skouen J S. Inter-examiner reliability of a classification system for patients with non-specific low back pain. Man Ther 2008; 14: 555–561.
15. Sim J, Wright C. Validity, reliability and allied concepts. In: Sim J, Wright C, editors. Research in health care. Concepts, designs and methods. Cheltenham: Stanley Thornes; 2000, p. 123–139.
16. Billis EV, McCarthy CJ, Stathopoulos I, Kapreli E, Pantzou P, Oldham JA. The clinical and cultural factors in classifying low back pain patients within Greece: a qualitative exploration of Greek health professionals. J Eval Clin Pract 2007; 13: 337–345.
17. Billis E, McCarthy CJ, Gliatis J, Stathopoulos I, Papandreou M, Oldham JA. Which are the most important discriminatory items for subclassifying non-specific low back pain? A Delphi study among Greek health professionals. J Eval Clin Pract 2010; 16: 542–549.
18. Billis E, McCarthy CJ, Roberts C, Gliatis J, Papandreou M, Gioftsos G, et al. Sub-grouping non-specific low back pain patients based on cluster analysis of discriminatory clinical items. J Rehabil Med (in press).
19. Billis, E. Important clinical features in non-specific low back pain in Greece. An international perspective with emphasis on the Greek healthcare setting [PhD dissertation]. Manchester: University of Manchester; 2009.
20. Cohen J. Weighted kappa-nominal scale agreement with provi-

sion of scaled disagreement or partial credit. Psychol Bull 1968; 70: 213.
21. Sim J, Wright C. Research in health care. Concepts, designs and methods. Cheltenham: Stanley Thornes Ltd; 2000, p. 294–360.
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–174.
23. Strender LE, Sjoblom A, Sundell K, Ludwig R, Taube A. Interexaminer reliability in physical examination of patients with low back pain. Spine 1997; 22: 814–820.
24. McCarthy CJ, Gittins M, Roberts C, Oldham JA. The reliability of the clinical tests and questions recommended in international guidelines for low back pain. Spine 2007; 32: 921–926.
25. Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. Arch Phys Med Rehabil 2003; 84: 1858–1864.
26. Waddell G, Main CJ, Morris EW, Venner RM, Rae PS, Sharmy SH, et al. Normality and reliability in the clinical assessment of backache. Br Med J (Clin Res Ed) 1982; 284: 1519–1523.
27. May S, Littlewood C, Bishop A. Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. Austral J Physiother 2006; 52: 91–102.
28. Vroomen PC, de Krom MC, Knottnerus JA. Consistency of history taking and physical examination in patients with suspected lumbar nerve root involvement. Spine 2000; 25: 91–96.
29. Henschke N, Maher CG, Refshauge KM. A systematic review identifies five "red flags" to screen for vertebral fracture in patients with low back pain. J Clin Epidemiol 2008; 61: 110–118.
30. Ohnmeiss DD. Repeatability of pain drawings in a low back pain population. Spine 2000; 25: 980–988.
31. Van Dillen LR, Sahrmann SA, Norton BJ, Caldwell CA, Fleming DA, McDonnell MK, et al. Reliability of physical examination items used for classification of patients with low back pain. Phys Ther 1998; 78: 979–988.
32. McCombe PF, Fairbank JC, Cockersole BC, Pynsent PB. 1989 Volvo Award in clinical sciences. Reproducibility of physical signs in low-back pain. Spine 1989; 14: 908–918.
33. Aina A, May S, Clare H. The centralization phenomenon of spinal symptoms – a systematic review. Man Ther 2004; 9: 134–143.
34. McKenzie R. The lumbar spine: mechanical diagnosis and therapy. Waikanae: Spinal Publications Ltd; 1981.
35. Bertilson BC, Bring J, Sjoblom A, Sundell K, Strender LE. Interexaminer reliability in the assessment of low back pain (LBP) using the Kirkaldy-Willis classification (KWC). Eur Spine J 2006; 15: 1695–1703.
36. van TE, Anderegg Q, Bossuyt PM, Lucas C. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. Man Ther 2005; 10: 256–269.
37. Arab AM, Abdollahi I, Joghataei MT, Golafshani Z, Kazemnejad A. Inter- and intra-examiner reliability of single and composites of selected motion palpation and pain provocation tests for sacroiliac joint. Man Ther 2009; 14: 213–221.
38. McCarthy CJ. Introduction to combined movement theory. In: McCarthy CJ, editor. Combined movement theory. Rational mobilization and manipulation of the vertebral column. Edinburgh: Churchill Livingstone; 2010, p. 3–8.
39. O'Sullivan P. Diagnosis and classification of chronic low back pain disorders: maladaptive movement and motor control impairments as underlying mechanism. Man Ther 2005; 10: 242–255.
40. Kent PM, Keating JL. Can we predict poor recovery from recent-onset nonspecific low back pain? A systematic review. Man Ther 2008; 13: 12–28.
41. Flynn T, Fritz J, Whitman J, Wainner R, Magel J, Rendeiro D, et al. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. Spine 2002; 27: 2835–2843.