Inter-rater reliability of a modified version of Delitto et al.'s classification-based system for low back pain: a pilot study

Adri T. Apeldoorn¹, Hans van Helvoirt², Raymond W. Ostelo³,
Hanneke Meihuizen², Steven J. Kamper⁴, Maurits W. van Tulder⁵,
Henrica C. W. de Vet⁶

¹Department of Epidemiology and Biostatistics and the EMGO⁺ Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands & Rehabilitation Department, Medical Centre Alkmaar, The Netherlands, ²Medical Back Neck Centre, The Hague, The Netherlands, ³Department of Epidemiology and Biostatistics and the EMGO⁺ Institute for Health and Care Research, VU University Medical Centre & Department of Health Sciences, Faculty of Earth and Life Sciences, VU University Amsterdam, The Netherlands, ⁴Department of Epidemiology and Biostatistics and the EMGO⁺ Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands & The George Institute, University of Sydney, Australia, ⁵Department of Health Sciences and the EMGO⁺ Institute for Health and Care Research, Faculty of Earth and Life Sciences, VU University Amsterdam, The Netherlands, ⁶Department of Epidemiology and Biostatistics and the EMGO⁺ Institute for Health and Care Research, VU University Medical Centre, Amsterdam, The Netherlands

Study design: Observational inter-rater reliability study.

Objectives: To examine: (1) the inter-rater reliability of a modified version of Delitto *et al.*'s classification-based algorithm for patients with low back pain; (2) the influence of different levels of familiarity with the system; and (3) the inter-rater reliability of algorithm decisions in patients who clearly fit into a subgroup (clear classifications) and those who do not (unclear classifications).

Methods: Patients were examined twice on the same day by two of three participating physical therapists with different levels of familiarity with the system. Patients were classified into one of four classification groups. Raters were blind to the others' classification decision. In order to quantify the inter-rater reliability, percentages of agreement and Cohen's Kappa were calculated.

Results: A total of 36 patients were included (clear classification $n=23$; unclear classification $n=13$). The overall rate of agreement was 53% and the Kappa value was 0.34 [95% confidence interval (CI): 0.11–0.57], which indicated only fair inter-rater reliability. Inter-rater reliability for patients with a clear classification (agreement 52%, Kappa value 0.29) was not higher than for patients with an unclear classification (agreement 54%, Kappa value 0.33). Familiarity with the system (i.e. trained with written instructions and previous research experience with the algorithm) did not improve the inter-rater reliability.

Conclusion: Our pilot study challenges the inter-rater reliability of the classification procedure in clinical practice. Therefore, more knowledge is needed about factors that affect the inter-rater reliability, in order to improve the clinical applicability of the classification scheme.

Keywords: Agreement, Physical therapy, Subgrouping, Classification, Low back pain, Reliability

Introduction

The identification of homogenous subgroups according to evidence-based classification systems was determined as a priority for primary care research on low back pain (LBP) as early as 1996.¹ Several classification systems for patients with LBP have been proposed. Billis *et al.*² identified 39 different classification systems in nine countries. Most of these systems utilize a biomedical

approach and a minority a psychosocial or biopsychosocial approach. Although it is generally recognized that categorizing patients solely by biomedical characteristics will be insufficient to discriminate between clinically meaningful subgroups,³ the need for a reliable and valid biomedical classification system based on clusters of signs and symptoms, not presumed patho-anatomical causes, seems certain.⁴

One of the biomedical classification systems with preliminary evidence of methodological robustness is a classification-based treatment algorithm initially

Correspondence to: A. Apeldoorn, VU University Medical Centre, van der Boerhorststraat 7, MF D518, 1081 BT Amsterdam, The Netherlands. Email: a.apeldoorn@vumc.nl

developed by Delitto *et al.*⁵ This classification-based treatment approach aims to identify subgroups of patients who are most likely to respond to the following interventions: (1) direction-specific exercises, (2) manipulation, (3) stabilization exercises, and (4) traction. Since being originally proposed, the criteria for allocating a patient to a specific intervention, and also the intervention procedures, have been revised and updated.^{4,6} For example, in 2006, Fritz *et al.*⁷ re-configured the classification system into two parts, excluded the subgroup traction, and introduced a hierarchical structure. The first part describes the criteria that the patient must meet in order to be placed into one of the three subgroups. The hierarchical structure gives precedence to the subgroup directional preference, followed by manipulation and then stabilization exercises. This structure means that patients that meet the criteria for more than one subgroup are assigned to the earlier subgroup. The second part provides additional non-hierarchical criteria (e.g. increasing episode frequency) to classify patients who do not meet the subgrouping criteria in the first part of the algorithm.

Since the proposal of Delitto *et al.*,⁵ the findings of several studies provide evidence in support of this subgrouping approach. The inter-rater reliability of Delitto *et al.*'s classification-based system⁵ has been assessed in at least seven studies (Table 1).^{6–12} The overall picture is that the inter-rater agreement of Delitto *et al.*'s system⁵ seems to be very good when raters assess patients together and discuss examination findings,⁸ moderate to good when only the reliability of the scoring system is tested (vignette studies)^{7,11} and fair to good when each patient is assessed independently (test–retest studies).^{6,9,10,12} Studies using a test–retest design found fair results when raters were unfamiliar with the system (Cohen's Kappa 0.15)¹⁰ and moderate to good results when raters were familiar with the system (Kappa 0.49–0.65).^{6,9,12} This has led to the suggestion that physical therapists who are familiar with the system can obtain a moderate level of inter-rater reliability.¹³ However, this suggestion must be interpreted with caution, because these studies used small patient samples (30–45 patients) with consequently large 95% confidence intervals (CIs).

Another issue that needs further evaluation is the finding of Stanton *et al.*⁶ that the inter-rater reliability may be lower for patients with an 'unclear classification' (i.e. additional criteria in the second part of the algorithm are needed to classify the patient) compared to those with a 'clear classification' (i.e. patients could be classified using the hierarchical boxes in the first part of the algorithm only). The prevalence of patients with an unclear classification has been estimated at between 26% and 50%.^{6,8,11}

Stanton *et al.*⁶ showed that the inter-rater reliability was fair for patients with an unclear classification (agreement 50%, Kappa 0.23, 95% CI: 0.21–0.66) and good for patients with a clear classification (agreement 79%, Kappa 0.69, 95% CI: 0.42–0.96). Stanton *et al.*⁶ concluded that, if the use of additional criteria in those with unclear classifications results in a lower inter-rater reliability, this would indicate that there is a need to refine the algorithm to reduce the proportion of patients with an unclear classification. However, the CIs in the study of Stanton *et al.*⁶ were large and there is not yet enough evidence to conclude that assessments of clinicians are less reliable when classifying patients with an unclear classification. Thus, the aims of this study were to examine: (1) the inter-rater reliability of a modified version of Delitto *et al.*'s classification-based system;⁵ (2) the influence of different levels of familiarity with the system; and (3) the inter-rater reliability of algorithm decisions in patients with clear and unclear classifications.

Methods

Raters

The raters involved in the present study were three physical therapists trained in Mechanical Diagnosis & Therapy (MDT) with 8–20 years of experience in MDT and 14–26 years of experience in orthopedic settings. The levels of postgraduate education and the degree of clinical experience with the modified version of Delitto *et al.*'s system⁵ differed between the three raters. Principal investigator (AA) holds a Master's degree in Science, and is a credentialed manual therapist and MDT practitioner. The second rater (HvH) holds a Master's degree in Psychology and is a manual therapist and a diplomat in MDT. The experience in using Delitto *et al.*'s system⁵ of the principal investigator and second rater was designated as 'expert' due to their participation in research involving the system and having used the system clinically. This designation is consistent with that in the study of Fritz *et al.*⁷ Relevant research experience involved the development of a randomized controlled trial¹³ to assess the effectiveness of Delitto *et al.*'s system.⁵ In the conduct of the randomized controlled trial,⁸ the first rater (AA) assessed 150 patients according to the system and discussed the outcomes regularly with the second rater (HvH) who had used the system in clinical care for 2 years.

The third rater (HM) is a diplomat in MDT with minimal algorithm experience. Before the present study, the two expert raters examined and classified 10 patients together to minimize variability in examination techniques and algorithm decision-making. The rater with minimal algorithm experience received written instructions regarding operationalization of the system and a 2-hour training course in

Table 1 Studies examining the inter-rater reliability of Delitto et al.'s classification-based treatment approach

Method	Study reference	Participants (n)	Raters (n)	Setting	Number of classifications	Agreement (%)	Kappa value (95% CI)
One examination A pair of raters examined the patient simultaneously and discussed the final classification Vignettes Written patient cases	Apeldoorn ⁸	34 sub-acute and chronic LBP, mean ODI 20%	1 principal investigator and 3 exp. PTs, unfamiliar with the system (2 days of training)	Outpatient physical therapy clinic	3	97	0.95 (0.85–1)
	Fritz ⁷	123 acute and sub-acute LBP, mean ODI 43%	10 exp. PTs familiar with the system, 10 exp. PTs, unfamiliar with the system, 10 novice PTs, unfamiliar with the system	Outpatient physical therapy clinic	3	76	0.60 (0.56–0.64) (No significant differences based on level of experience)
	Henry ¹¹	24 acute and sub-acute LBP, mean ODI 40%	18 mostly exp. PTs and 1 neuroscientist, of which most were unfamiliar with the system	Outpatient physical therapy clinic	3	76	0.57 (0.55–0.69) (No significant differences based on level of experience)
Repeated examinations An average of 2 days between examinations (range 0–6) 15 minutes between examinations	Fritz ⁹	43 acute LBP, mean ODI 40%	7 exp. PTs familiar with the system	Outpatient physical therapy clinic	4	65	0.49
	Heiss ¹⁰	45 acute LBP, mean ODI 40%	4 exp. PTs, unfamiliar with the system (one day of training)	Outpatient rehabilitation centre	4	40	0.15
	Kiesel ¹²	30 acute, sub-acute and chronic LBP	1 experienced principal investigator tested in combination with 8 PTs familiar with the system	Outpatient physical therapy clinic	3	77	0.65 (0.42–0.87)
90% of the second examinations directly after the first examination (range 0–4 days)	Stanton ⁶	32 acute and sub-acute LBP, mean ODI 37%	2 exp. PTs familiar with the system tested in combination with 7 PTs unfamiliar with the system (45 minutes of training)	Different settings (mainly hospitals)	4	67	0.52 (0.27–0.77)

Note: CI, confidence interval; LBP, low back pain; ODI, Oswestry Disability Index; exp., experienced; PTs, physical therapists.

Part 1

Does the patient have a directional preference (DP):

A DP is established if certain postures or repeated end-range movements in one direction (e.g. flexion) cause the symptoms to improve or centralize (pain moves from a distal to a proximal area), and in the opposite direction (e.g. extension) cause the symptoms to worsen or peripheralize (pain moves from a proximal to a distal area)

yes

Direction-specific exercises

no

Does the patient have at least 4 of the following:

1. duration of symptoms <16 days
2. No symptoms distal to the knee
3. Low FABQW score (<19)
4. ≥ 1 hypomobile segment on lumbar segmental mobility testing
5. Hip internal rotation ROM (>35° for at least 1 hip)

yes

Spinal manipulation

or

Does the patient have at least 3 of the following:

1. Greater general flexibility (average SLR ROM >91° or postpartum or high BLLS [≥ 4 out of 9])
2. Positive prone instability test
3. Positive aberrant movements
4. Age <40 years

yes

Stabilization exercises

For postpartum patients (no cut-off value defined):

Positive thigh thrust test, positive ASLR, positive modified Trendelenburg, pain provocation with palpation of the long dorsal sacroiliac ligament or pubis symphysis

or

Does the patient meet the following

Signs and symptoms of a current nerve root compression that do not centralize with a full MDT assessment.

yes

Traction

no

Part 2

Which subgroup does the patient fit best into?

Direction-specific exercises		Spinal manipulation	Stabilization exercises	
Factors favouring	Factor against	Factors against	Factors favouring	Factors against
<ul style="list-style-type: none"> • Strong preference for sitting or walking or cycling • <u>Peripheralization</u> in direction opposite to centralisation <p>Flexie DP:</p> <ul style="list-style-type: none"> • >50 years of age • Imaging evidence of lumbar spinal stenosis <p>Lateral DP:</p> <ul style="list-style-type: none"> • Visible lateral shift 	<ul style="list-style-type: none"> • LBP only 	<ul style="list-style-type: none"> • Symptoms below the knee • Increasing episode frequency • No pain with mobility testing • No <u>hypomobility</u> • Less discrepancy in left-to-right hip internal rotation (<10°) • Negative <u>Gaenslen's</u> sign • <u>Peripheralization</u> with motion testing 	<ul style="list-style-type: none"> • Hypermobility with spring testing • Increasing episode frequency • 3 or more prior episodes • Previous severe low back/pelvis incident 	<ul style="list-style-type: none"> • Discrepancy in SLR ROM (>10°) <p>Does the patient have at least 3 of the following:</p> <ul style="list-style-type: none"> • FABQPA score <9 • Negative prone instability test • Absence of aberrant movements • Absence of lumbar hypermobility

Abbreviations: LBP, low back pain; FABQA and FABQW, Fear-Avoidance Beliefs Questionnaire subscales activity and work; ROM, range of motion; SLR, straight leg raise; BLLS, Beighton Ligamentous Laxity Scale; ASLR, active straight leg raise; MDT, Mechanical Diagnosis & Therapy

Figure 1 Modified classification-based treatment system for patients with LBP.

the protocol and examination techniques by the principal investigator. The extensive experience in MDT of the three raters can be considered an advantage for this study since, in MDT and in Delitto *et al.*'s system,⁵ directional preference and

centralization plays an essential role. The study was approved by the Medical Committee of the VU University Medical Centre in Amsterdam and registered into a clinical trial register (<http://www.controlled-trials.com/ISRCTN37135033/>).

Abbreviation: NRS, numerical rating scale

Figure 2 Design of the study.

Patients

Patients were recruited by a research assistant in the Medical Back Neck Centre, an outpatient private practice in The Hague, The Netherlands. Potentially eligible patients had the study procedures explained to them and were sent written information about the study. Inclusion criteria were LBP as the primary complaint, with or without associated leg pain, age

between 18 and 65 years, and the ability to read and write Dutch. Exclusion criteria were known or suspected serious low back pathology, e.g. cauda equina compression, fractures, severe radiculopathy, and spondylolisthesis (grade 2 or more), serious comorbidity (e.g. metastases, AIDS, cerebrovascular accident), psychopathology, currently pregnant or had given birth in the past 3 months, and an inability

Table 2 Baseline patient characteristics

Characteristics	Classification-based group (n=36)
Age (SD)	43.9 (13.1)
Female, n (%)	20 (55.6)
History of LBP	
No. of months since first experience of LBP (range)	123.0 (24.0–240.0)
Previous episodes of LBP, n (%)	28 (77.8)
Lower back surgery, n (%)	3 (8.3)
No. of months duration of current LBP (range)	8.0 (2.1–18.0)
Acute LBP (0–6 weeks), n (%)	4 (11.1)
Sub-acute LBP (7–12 weeks), n (%)	7 (19.5)
Chronic LBP (>12 weeks), n (%)	25 (69.4)
Pain intensity in the past week (NRS: 0–10) (SD)	5.8 (2.4)
Pain is continuous, n (%)	11 (30.5)
Pain radiated into the leg, n (%)	20 (55.6)
ODI (0–100) (SD)	24.4 (18.1)
ODI (0–100) ≥25, n (%)	15 (41.7)
Currently taking medication for LBP, n (%)	6 (16.7)
Smoker, n (%)	7 (19.5)
Employed, n (%)	30 (83.3)
Employed and currently working, n (%)	25 (83.3)
Employed, but currently on sick leave, n (%)	5 (16.6)
FABQ	
Activity (0–24) (SD)	13.1 (5.8)
Work (0–42) (range)	8.0 (0.5–14.5)
Waddell's non-organic signs (0–8) (range)	0.0 (0.0–1.0)

Note: SD, standard deviation; LBP, low back pain; NRS, numerical rating scale; ODI, Oswestry Disability Index; FABQ, Fear-Avoidance Beliefs Questionnaire.

Values are the mean (standard deviation) and the median (inter-quartile range).

to demonstrate any LBP symptoms during a mechanical examination. For the sake of clarity, patients with mild sensory loss and diminished myotomal strength on the ipsilateral lower extremity of not less than 4 with standardized manual muscle testing (0–5 scale),¹⁴ and patients with a diagnosed spondylolisthesis grade 1 were included.

Modifications of the Delitto et al.'s classification-based system

In the present study, Delitto *et al.*'s classification-based system⁵ was modified partly based on a clinical commentary by Fritz *et al.*⁴ They advise to use recently developed clinical prediction rules for the manipulation^{15,16} and stabilization exercises¹⁷ classifications and to use centralization and directional preference, instead of only centralization, to classify patients into the direction-specific exercises group. Directional preference involves a broader operational definition than centralization (Fig. 1). Another modification was the use of a complete MDT assessment to assess centralization and directional preference instead of the simplified version used in previous studies.^{6,7,18} In the latter studies, repeated movement testing was only performed with the patient in the sagittal plane.^{6,7,18} In a complete MDT assessment, symptomatic and mechanical responses are monitored during highly standardized single and repetitive end-range lumbar test movements (e.g. flexion, extension or side-gliding) and/or prolonged positioning (e.g. flexion/rotation and extension in prone) in different directions. In the present study, test movements were repeated in sets of 5 or 10 with a maximum of 40, and terminated if a clear centralization or peripheralization response occurred. Several studies have reported on the inter-rater reliability of centralization and MDT classifications; however, only three studies have used independent and consecutive examinations in patients with LBP.^{7,19,20} In the study by Riddle and Rothstein,¹⁹ 49 raters and 363

patients participated. The agreement value for inter-rater reliability of the MDT syndromes (postural, dysfunction, and derangement) was 60% and the Kappa value 0.33. A limitation of this study was that only a minority of the raters (16 out of 49) had attended one or more postgraduate MDT courses. In the study by Fritz *et al.*,⁷ 57 patients were tested; inter-rater agreement values were 54% to 70% and Kappa values 0.15–0.46 for judgments about centralization following repeated extension, repeated flexion and sustained extension. They did not use a formal MDT procedure and provided no information about the number of examiners, or how centralization was measured. Recently, Werneke *et al.*²⁰ examined the inter-rater reliability of the MDT system (postural, dysfunction, derangement, and other) with 47 raters at different stages of MDT postgraduate training (A–D) and 1662 patients. The agreement values varied between 86 and 91% for MDT subgroups and corresponding Kappa values varied between 0.37 and 0.44. Werneke *et al.*²⁰ used a threshold of 0.60 as a criterion for the acceptable Kappa value and concluded that the reliability of the MDT system was not adequate. Although we take the view that their conclusion is unduly influenced by Kappa values (a relative measure), as opposed to percentages agreement values (an absolute measure),²¹ we conclude that there is room for improvement in the inter-rater reliability of centralization and MDT classifications.

Examination procedures

All potentially eligible patients were referred to the principal investigator who evaluated patient eligibility, obtained written informed consent, collected baseline questionnaires, and conducted a clinical examination according to a modified version of Delitto *et al.*'s classification-based system⁵ (Fig. 1). All clinical tests, such as the prone instability test and lumbar segmental mobility testing, were standar-

Table 3 Distribution of patients to classifications made by rater pairs A–B (*n*=28) and A–C (*n*=8). Agreement within pairs of raters in **bolds**

Rater A	Specific exercises		Manipulation		Stabilization exercises		Traction	
	B	C	B	C	B	C	B	C
Specific exercises	6	2	1	1	1	0	2	1
Manipulation	3	0	5	2	0	0	0	0
Stabilization exercises	1	0	2	0	2	0	2	0
Traction	1	1	1	0	0	0	1	1
Total number of patients	11	3	9	3	3	0	5	2
Percent agreement and Kappa (95% CI) by classification category	8/14=57% K=0.14 (0–0.66)		7/12=58% K=0.17 (0–0.72)		2/3=67% K=0 (0–1)		2/7=29% K=could not be calculated	

Note: Rater A [principal investigator (AA)] and rater B (HvH) were designated as experts in the decision-making algorithm. Rater C (HM) had minimal algorithm experience. The three raters were physical therapists trained in Mechanical Diagnosis & Therapy (MDT) with >7 years of experience in MDT and >13 years of experience in orthopedic settings. CI, confidence interval; K, Kappa.

Figure 3 An overview of the inter-rater reliability of Delitto *et al.*'s classification-based system for low back pain. CI, confidence interval. Inter-rater reliability studies: One examination (raters were allowed to discuss the examination): 1, Apeldoorn *et al.*⁸ Using vignettes: 2, Fritz *et al.*⁷; 3, Henry *et al.*¹¹ Repeated examinations: 4, Fritz and George⁹ (no CI available); 5, Heiss *et al.*¹⁰ (no CI available); 6, Kiesel *et al.*¹²; 7, Stanton *et al.*⁶; 8, Apeldoorn *et al.* (current study).

dized. The operational definitions of these tests are summarized in Appendix 1. In cases of uncertainty about the primary classification, extra criteria, such as three or more prior LBP episodes in the medical history and a visible lateral shift during the examination, were used to determine the patient's subgroup (Fig. 1). All patients were classified into one subgroup according to the classification criteria and hierarchical structure of the system. If additional criteria in the second part of the classification algorithm were needed to place the patient in one of the four subgroups, the patient was given the unclear classification label.

Immediately following examination by the principal investigator, each patient was examined independently by one of the two other raters using the same examination protocol. Which of the two raters conducted the second examination depended on availability and was not randomized. The patients were told not to discuss findings of the first examination with the second rater. The duration of each examination was approximately 1 hour. Each patient was classified into one of the following treatment subgroups: direction-specific exercises, spinal manipulation, stabilization exercises, or traction. Before each examination, a score for current LBP and a score for pain in the leg was collected to check stability of the patient's pain between examinations. Unstable patients were defined as those having >2 points of change on an 11-point numerical

rating scale (NRS: 0–10) for current LBP or current pain in the leg. Two points has been regarded as the minimum clinically important difference on the NRS.²² The written evaluations were placed in envelopes and the raters remained blind to the other's classification decision during the study.

Figure 2 shows the design and the clinical assessment procedure of the present study.

Data analysis

The prevalence of each of the four classifications was calculated for the principal investigator and for the other two raters together. The prevalence in the classifications between raters is not an indication for reliability, because this represents a different construct. Therefore, we calculated the inter-rater reliability of the modified version of Delitto *et al.*'s system⁵ for the total group of patients, and separately for rater pairs with different levels of familiarity with the system and for patients with a clear or unclear classification. The principal investigator's decision was used to establish a clear or unclear classification. Percentages of agreement and unweighted kappa, including its 95% CI, served to test inter-rater reliability. Kappa indicates the proportion of observed agreement between raters beyond agreement by chance. For kappa, the following interpretation was used: ≤0.20 poor, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, and 0.81–1.00 very good agreement.²³ Finally, in an additional analysis, the inter-rater reliability was calculated excluding patients who were 'unstable' (>2

points of change on an 11-point NRS) between examinations. The data were analyzed using IBM SPSS Statistics Version 20.0 (IBM Corporation, Armonk, NY, USA) and VasserStats (<http://faculty.vassar.edu/lowry/kappa.html>).

Results

From April 2009 to December 2010, 36 patients were included in the study. We did not count those patients who refused to participate and did not ask for their reasons for declining to participate in the study. Of the 36 participating patients, there were no withdrawals. The majority of the patients (69%) had chronic LBP (>12 weeks), and the mean Oswestry Disability Index (ODI)²⁴ score was 24. Table 2 presents the main clinical and demographic characteristics of the patients. As outlined in the section on 'Methods', all patients were examined first by the principal investigator. The second (HvH) and third rater (HM) conducted 28 and 8 examinations, respectively. The uneven numbers between the second and third rater occurred due to logistic reasons related to availability. The proportion of patients in each classification group for the principal investigator was: specific exercises 39%, manipulation 28%, stabilization exercises 19%, and traction 14%. The proportions for the other two raters together were: specific exercises 39%, manipulation 33%, stabilization exercises 8%, and traction 19%. Percentage of agreement varied between 29 and 67% for the four classification categories (Table 3).

The overall agreement was 53%, with a Kappa of 0.34 (95% CI: 0.11–0.57). The agreement was 50% for the pair of raters considered experts in algorithm use (AA–HvH), with a Kappa of 0.31 (95% CI: 0.05–0.56) and 63% for the pair of raters of whom one had only minimal algorithm experience (AA–HM), with a Kappa of 0.43 (95% CI: 0–0.94). According to the principal investigator, the percentage of patients that met the criteria of a clear classification was 64% ($n=23$), classified as follows: specific exercises 52%, manipulation 17%, stabilization exercises 17%, and traction 13%. Thirty-six percent ($n=13$) had an unclear classification. This group had the following final classifications: specific exercises 15%, manipulation 46%, stabilization exercises 23%, and traction 15%. For patients with a clear classification, the overall agreement was 52%, with a Kappa of 0.29 (95% CI: 0–0.59). For patients with an unclear classification, the overall agreement was 54%, with a Kappa of 0.33 (95% CI: 0–0.72).

In an additional analyses, three patients that reported >2 points of change on an 11-point NRS between the two examinations were excluded. However, excluding these patients did not lead to higher reliability values.

Discussion

Research has shown that most procedures commonly used by clinicians in the examination of patients with back pain demonstrate low reliability.²⁵ In order to improve reliability, the use of a combination of test findings has been advocated.²⁶ Although Delitto *et al.*'s system⁵ uses clusters of signs and symptoms, the present pilot study shows that, for raters experienced in MDT method, the inter-rater reliability of a modified version of Delitto *et al.*'s classification-based treatment system⁵ using a test–retest design was not more than fair. Experience with the classification system of Delitto *et al.*⁵ did not improve the inter-rater reliability. That is, the inter-rater reliability between two raters who were considered experts in use of the algorithm was not better than the inter-rater reliability of two raters of whom one had minimal algorithm experience. Figure 3 presents the reliability scores of the Delitto *et al.*'s classification system⁵ from the current study, as well as previous studies.

Our MDT trained raters clearly had trouble classifying patients in a consistent way, although they regarded the system as easy to use and uncomplicated. The utility of the system is supported by the finding that experience with the system of Delitto *et al.*⁵ did not affect inter-rater reliability, which is in line with two other reliability studies of Delitto *et al.*'s system⁵ that used written patient cases.^{7,11}

A total of 36% of the patients received an unclear classification. This is in concordance with recent studies that found 26%,⁸ 34%,⁶ and 50%¹¹ of patients with unclear classifications. Interestingly, we found almost no difference in the inter-rater reliability between patients with a clear classification and an unclear classification. This is in contrast with Stanton *et al.*⁶ who found a lower inter-rater reliability in patients with an unclear classification compared to patients with a clear classification. Our study and the study of Stanton *et al.*⁶ used small patient samples. In order to draw definitive conclusions about the inter-rater reliability of patients with clear an unclear classification more research is needed. In our study, the most common classification for patients with a clear classification was specific exercises. For patients with an unclear classification, manipulation was the most common subgroup. Stanton *et al.*⁶ did not find the same trend. That group reported a larger proportion of patients with an unclear classification belonging to the specific exercise subgroup.

Four other studies evaluating the inter-rater reliability of Delitto *et al.*'s classification system⁵ have used a similar methodology to this study (Table 1 and Fig. 3).^{6,9,10,12} However, direct comparability of the studies is hampered due to differences in patient populations, characteristics of the raters and the somewhat different algorithms and test procedures

(Table 1). For example, in order to determine if the patient met the subgroup direction-specific exercises, we used the MDT method (usually 10–20 repetitions with a maximum of 40) and the operational definition for directional preference, whereas the four other studies used a maximum of 10 repeated flexion and extension movements and the operational definition of centralization. Because directional preference is a broader category than centralization, more patients will fall into the classification of direction-specific exercises. This modification was based on recommendations of Fritz *et al.*⁴ Also, Long *et al.*²⁷ found that direction-specific exercises matching patients' directional preference significantly and rapidly decreased pain in patients assessed with a standardized MDT assessment. Another difference is that the present study included mainly patients with chronic pain (69%), whereas three of the other four studies included only patients with acute and sub-acute pain.^{6,9,10} The presence of chronic LBP might reduce the reliability of the assessment. It could be argued that patients with chronic LBP are more difficult to classify using this approach, because the higher levels of distress and depression common in these patients may confound the interpretation of biomedical signs and symptoms.²⁸

Delitto *et al.*'s classification-based treatment approach⁵ was originally developed for patients with acute symptoms and disability scores of >30 on the ODI, although Delitto *et al.*⁵ defined acuteness based on the severity of disability rather than days since onset. However, to date, there is no empirical evidence to suggest that this classification system is not suitable for patients with a longer duration of pain and/or less disability. In fact, it can be argued that this system is well suited to patients with chronic pain. Firstly, a large proportion of patients with chronic LBP meet the criteria for the direction-specific exercises, i.e. centralization and directional preference. May and Aina²⁹ showed in a recent review that a centralization response occurs in 50% of patients with sub-acute LBP and 40% of patients with chronic LBP. Secondly, although the prevalence of certain psychological characteristics is higher in patients with chronic LBP, it is unclear if these characteristics are helpful in guiding treatment,^{30,31} and, as such, a mechanically-based classification system may still be useful. This point is reinforced by findings that, in patients with acute, sub-acute and chronic LBP, centralization might be a stronger predictor of outcome than psychosocial factors if treatment is matched to directional preference.^{32–34}

There have been three inter-rater reliability studies that have used a different methodology compared to this study (Table 1 and Fig. 3). In a study of Apeldoorn *et al.*,⁸ raters were allowed to discuss examination findings. They reported a very good inter-rater reliability

(97% agreement, Kappa 0.95). Fritz *et al.*⁴ and Henry *et al.*¹¹ (the latter used a randomly selected subset of the cases by the study of Fritz *et al.*⁷) used written patient cases (vignettes) and obtained similar inter-rater agreement results, 76% (Kappa 0.57) and 76% (Kappa 0.60), respectively. It is not surprising that the study by Apeldoorn *et al.*⁸ showed almost perfect agreement because raters assessed each patient together and could influence each other's opinion. The two vignette studies found somewhat better agreement scores compared to most of the studies that used repeated assessments. This is also not surprising because, in vignette studies, ratings are based on the same signs and symptoms and only concern the reliability of the scoring system. In repeated assessments, there are two further sources of variability; patients may present differently in two separate examinations and testing procedures might differ between raters.

It could be argued that the symptoms reported by patients can vary between two test situations and might affect classification judgement. In the present study, only three patients reported more pain and the inter-rater agreement in these patients was 100%. Therefore, in concordance with the findings of Heiss *et al.*,¹⁰ we do not believe that our inter-rater reliability results have been seriously influenced by fluctuations in symptom severity scores.

It has been suggested that the inter-rater reliability might differ for the four categories. For example, Henry *et al.*¹¹ reported that raters disagreed particularly about the stabilization exercises category. However, because of the small number of patients included in our study, we did not have the opportunity to examine specific patterns of mismatches. Concerning the specific exercises category, previous studies assessed repeated movements only in the sagittal plane.^{6,7,18} In these studies, patients were asked to perform repeated (i.e. 10 repetitions) extension in standing, repeated flexion in sitting, and sustained (i.e. 30 seconds) extension in prone. Although this version of the assessment is less complicated than a full MDT assessment, inter-rater reliability regarding centralization was only poor to moderate (Kappa 0.15–0.46, percentage of agreement 54–74%).^{6,7} For the assessment of centralization and directional preference we consider a full MDT assessment more appropriate than repeated movements in the sagittal plane only. It is well accepted that movements other than flexion and extension may be necessary to establish a directional preference.³⁵ For example, patients with a visible lateral shift often respond well to repeated and sustained side-gliding movements.³⁶ It is interesting to note that the original study of Delitto *et al.*⁵ described three different syndromes; extension, flexion, and lateral,⁵ whereas the follow-up studies of Fritz *et al.*⁷ and Brennan *et al.*¹⁸ concentrated on sagittal movements and excluded patients with a visible lateral shift. In previous studies, an incomplete MDT assessment protocol to establish directional preference or

centralization may be responsible for reduced reliability in assigning patients to the specific exercises category. This could negatively impact the reliability of Delitto *et al.*'s classification system⁵ as a whole. We expected to find good reliability results for the classification specific exercises, because we used a full MDT assessment conducted by raters trained in the MDT method. However, the reliability of this classification was poor (Kappa 0.14, agreement 57%) and not better than the other three classifications. Although the current raters used the standard MDT assessment procedure and formal and well described definitions of directional preference and centralization, there is room to interpret signs and symptoms differently. For example, there is no consensus about how large the change in pain location must be to determine centralization. To grade centralization, Donelson³⁷ has proposed using a pain drawing that divides the lower back and leg in six regions (mid-lumbar, lateral lumbar, gluteal, thigh, crus, and foot). Centralization is identified when the location of pain moves to a more central location. For a patient with only mid lumbar pain, the pain must cease in order for them to be categorized as a centralizer. However, Albert *et al.*³⁸ believe that smaller changes in pain location (e.g. distal pain moving proximally from above the ankle to below the knee) must also be recorded as centralization. Another example that can lead to different interpretations is the time that centralization must be retained in response to therapeutic loading strategies. McKenzie and May³⁹ believe that this must be maintained once the patient resumes the upright position and becomes normally active. However, it is unclear how long this upright position must be continued and what normally active is. To categorize the patient as a centralizer, Albert *et al.*³⁸ suggested that the most distal pain must remain abolished for at least 1 minute after positioning the patient in a weightbearing position.

It is difficult to compare our inter-rater reliability findings for the category specific exercises with previously conducted studies that investigate the reliability of MDT. For example, most studies about the reliability of the MDT classification system or the centralization phenomenon used simultaneous assessment procedures,^{40–42} simultaneous subjective history,⁴³ or vignettes.⁴⁴ It is likely that such methodologies will result in higher reliability scores than the independent, consecutive examinations method that was used in our study. There is one reliability study that used a formal MDT procedure, MDT-trained examiners (levels A–D) and independent, consecutive examinations in patients with low back pain.²⁰ In this recent study, agreement percentages for categorizing patients in one of the four MDT subgroups (postural, dysfunction, derangement, and other) were satisfactory, but the Kappa values were low. They suggested several avenues for future research to improve inter-rater reliability of

the MDT system. One suggestion is to evaluate the level of agreement on patient's elements of the history, as well as on physical examination findings.²⁰

Concerning the reliability of Delitto *et al.*'s system,⁵ we must conclude that, to date, it is difficult to define precisely where the specific problems lie. Nevertheless, we suggest that the reliability of the system could be improved by standardizing tests and outlining clear operational definitions with uniform and reproducible descriptions.

Limitations

Owing to organizational constraints, the pairing of raters and the order of testing were not randomized. Instead, the principal investigator was always the first rater. Although we do not think that this has influenced our results essentially, the possibility of systematic bias cannot be excluded. Another limitation was the recruitment of patients from a specific primary care center and the use of only MDT trained raters. Therefore, external validity is limited to this patient population and to physical therapists who are trained in the MDT assessment. Further, a relatively small patient sample and only two pairs of raters were involved in the present study, whereas large patient samples and multiple pairs of raters are recommended to improve the quality of the study.²³ Lastly, Werneke *et al.*⁴⁵ suggested that, for some patients with LBP, more than one examination can produce more reliable final classification decisions. However, the idea that combining several assessments result in better inter-rater reliability has yet to be confirmed in an empirical study.

Conclusion

This pilot study provides additional insight into the impact of clinician expertise on performance of Delitto *et al.*'s classification-based system,⁵ explores the system as applied to patients with chronic symptoms and points toward the need to improve the reliability of the classification before it is suitable for clinical usage. The inter-rater reliability of Delitto *et al.*'s classification-based treatment system,⁵ as used in this study, was not more than fair and did not differ between patients with a clear or unclear classification. Despite the small sample size, our pilot study challenges the utility of the classification procedure. Greater understanding regarding the factors that affect inter-rater reliability is required from a larger study in order to improve the classification procedure.

Disclaimer Statements

Contributors All authors have made a substantial contribution to the concept and design, acquisition of data, or analysis and interpretation of data, drafted the article or revised it critically for important intellectual content and approved the version to be published.

Funding Department of Health.

Conflicts of interest The authors declare that they have no competing interests.

Ethics approval This study, under the alternate name 'A classification algorithm for low back pain:

matching patients to treatments that they are most likely to benefit', has received ethical approval from the Medical Ethics Committee of the VU University Medical Centre in Amsterdam (registration number: 2008\5).

Appendix 1: Operational definitions for the physical measurements

Patient standing

Test	Procedure and criteria
<ul style="list-style-type: none"> Lateral shift (list) Flexion, extension and lateral flexion mobility of the lumbar spine Aberrant movements 	<p>Visible asymmetrical alignment of the spine in the frontal plane to the right or left</p> <p>The patient is asked to flex, extend, and sidebend as far as possible. The examiner assesses the mobility as hypermobile, normal, or hypomobile</p> <p>The patient is asked to flex the trunk forward as far as possible and bend back to the erect position. The existence of any of the following five abnormalities is noted. The score is positive if one or more aberrant movements are present</p> <p>Transient mid-range pain felt during full flexion</p> <p>Transient mid-range pain felt on the way up from full trunk flexion</p> <p>Using the hands for assistance during return from the flexed to the erect position</p> <p>Any sudden acceleration, deceleration or lateral deviation of the trunk during full flexion or from the flexed to the erect position</p>
<ol style="list-style-type: none"> Painful arc in flexion Painful arc on return Gower sign ('thigh climbing') Instability catch 	<p>Transient mid-range pain felt during full flexion</p> <p>Transient mid-range pain felt on the way up from full trunk flexion</p> <p>Using the hands for assistance during return from the flexed to the erect position</p> <p>Any sudden acceleration, deceleration or lateral deviation of the trunk during full flexion or from the flexed to the erect position</p>
<ol style="list-style-type: none"> Reversal of lumbopelvic rhythm 	<p>The patient bends the knees and shifts the pelvis anteriorly before returning to erect position</p>
<ul style="list-style-type: none"> Trendelenburg test 	<p>The patient is instructed to raise one leg at 90° hip flexion. The test is positive if the pelvis is descending on the flexed side</p>
<ul style="list-style-type: none"> General ligamentous mobility on the Beighton Ligamentous Laxity Scale (BLLS) 	<p>The test consists of four bilateral tests: passive hyperextension of the elbow greater than 10°; passive hyperextension of the fifth finger greater than 90°; passive abduction of the thumb to contact the forearm; and passive hyperextension of the knees greater than 10°. The last test consists of placing both hands flat on the floor while bending forward without flexing the knees. A point is given for each test the subject can perform. The scores range from 0 to 9. With a test result of ≥ 4, hypermobility is generally considered to be present</p>
Patient lying supine	
<ul style="list-style-type: none"> Straight leg raise range of motion 	<p>The examiner raises the patient's leg passively at the ankle with one hand, while the other hand maintains the knee in extension. The leg is raised slowly to the maximum tolerated (not until the onset of pain) straight leg raise. Mobility is recorded</p>
<ul style="list-style-type: none"> Pain provocation test of the pubis symphysis Pelvic torsion or Gaenslen's test 	<p>The examiner gently palpates the front side of the pubic symphysis. The test is positive if the pain persists for at least five seconds after the examiner's hand is removed</p> <p>One leg is bent in full hip and knee flexion and the other extended with the shank over the edge of the table. The examiner simultaneously applies over-pressure to the flexed knee in a superior/posterior direction, and a downward force on the extended leg. The test is positive if the patient experiences familiar pain in the sacroiliac region. The test is performed on both sides</p>
<ul style="list-style-type: none"> Thigh thrust or posterior shear test 	<p>The examiner flexes the patient's hip on the painful side to 90°. The examiner places one hand under the sacrum and applies a downward (posterior) force through the line of the femur with the other hand. The test is positive if the patient experiences their familiar pain in the sacroiliac region</p>
<ul style="list-style-type: none"> Active straight leg raise test 	<p>The patient lies with legs straight and feet 20 cm apart, and is instructed to raise one leg 5 cm above the couch without bending the knee. The test is positive if the patient notices (slight or severe) heaviness in the leg that is raised. The test is performed on both sides</p>
Patient lying prone	
<ul style="list-style-type: none"> Posteroanterior tests for pain provocation and mobility 	<p>Segmental mobility testing of the lumbar spine is performed by the examiner placing the hypothenar eminence of the testing hand over the spinous process of the segment to be tested. This hand is reinforced with the other hand. The elbows and wrists are extended. The examiner gradually moves the body weight forwards and applies a gentle, but firm, anteriorly directed pressure on the spinous process. The examiner assesses the mobility of the five different lumbar levels as hypermobile, normal, or hypomobile, and notes whether pain is provoked</p>
<ul style="list-style-type: none"> Hip internal rotation range of motion 	<p>The patient lies in a prone position with legs straight and feet 20 cm apart. The examiner, supporting the patient's ankle on the side being examined, bends the patient's knee to 90° and passively rotates the hip internally until the pelvis starts to move. Mobility is recorded with a goniometer which is placed on the outside of the patient's ankle on the side being measured</p>
<ul style="list-style-type: none"> Test of the long dorsal sacroiliac ligament 	<p>The examiner palpates the areas above both of the patient's sacroiliac joints. The test is positive if at least on one side the pain persists for a minimum of 5 seconds after the examiner's hand is removed</p>

Appendix 1 continued

Test	Procedure and criteria
Patient lying half-prone • Prone instability test	The patient lies prone on the examination table, with the legs over the edge of the table and the feet resting on the floor. While the patient rests in this position, the examiner applies a postero-anterior pressure on the lumbar spine to identify a painful segment as previously described. Any provocation of pain is reported. The patient then lifts the legs off the floor to a height of 10–20 cm with straight knees, while holding the table to maintain position. In this position, passive intervertebral motion testing is applied again to the segments that were identified as painful. The test is positive if pain is present in the first position but subsides in the second position

References

- Borkan JM, Cherkin DC. An agenda for primary care research on low back pain. *Spine*. 1996;21:2880–4.
- Billis EV, McCarthy CJ, Oldham JA. Subclassification of low back pain: a cross-country comparison. *Eur Spine J*. 2007;16:865–79.
- O'Sullivan P. It's time for change with the management of non-specific chronic low back pain. *Br J Sports Med*. 2012;46:224–7.
- 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.
- Delitto A, Erhard RE, Bowling RW. A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative treatment. *Phys Ther*. 1995;75:470–85.
- Stanton TR, Fritz JM, Hancock MJ, Latimer J, Maher CG, Wand BM, *et al.* Evaluation of a treatment-based classification algorithm for low back pain: a cross-sectional study. *Phys Ther*. 2011;91:496–509.
- Fritz JM, Brennan GP, Clifford SN, Hunter SJ, Thackeray A. An examination of the reliability of a classification algorithm for subgrouping patients with low back pain. *Spine*. 2006;31:77–82.
- Apeldoorn AT, Ostelo RW, van Helvoirt H, Fritz JM, Knol DL, van Tulder MW, *et al.* A randomized controlled trial on the effectiveness of a classification-based system for subacute and chronic low back pain. *Spine*. 2012;37:1347–56.
- Fritz JM, George S. The use of a classification approach to identify subgroups of patients with acute low back pain. Interrater reliability and short-term treatment outcomes. *Spine*. 2000;25:106–14.
- 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;34:430–9.
- Henry SM, Fritz JM, Trombley AR, Bunn JY. Reliability of a treatment-based classification system for subgrouping people with low back pain. *J Orthop Sports Phys Ther*. 2012;42:797–805.
- Kiesel KB, Underwood FB, Mattacola CG, Nitz AJ, Malone TR. A comparison of select trunk muscle thickness change between subjects with low back pain classified in the treatment-based classification system and asymptomatic controls. *J Orthop Sports Phys Ther*. 2007;37:596–607.
- Apeldoorn AT, Ostelo RW, van Helvoirt H, Fritz JM, de Vet HC, van Tulder MW. The cost effectiveness of a treatment-based classification system for low back pain: design of a randomised controlled trial and economic evaluation. *BMC Musculoskelet Disord*. 2010;11:58.
- Kendall FP, McCreary EK, Provance PG. *Muscles: testing and function*. 4th ed. Baltimore, MD: Williams & Wilkins; 1993.
- Childs JD, Fritz JM, Flynn TW, Irrgang JJ, Johnson KK, Majkowski GR, *et al.* A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. *Ann Intern Med*. 2004;141:920–8.
- 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;27:2835–43.
- Hicks GE, Fritz JM, Delitto A, McGill SM. Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. *Arch Phys Med Rehabil*. 2005;86:1753–62.
- Brennan GP, Fritz JM, Hunter SJ, Thackeray A, Delitto A, Erhard RE. Identifying subgroups of patients with acute/subacute 'nonspecific' low back pain: results of a randomized clinical trial. *Spine*. 2006;31:623–31.
- Riddle DL, Rothstein JM. Interrater reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine*. 1993;18:1333–44.
- Werneke MW, Deutscher D, Hart DL, Stratford P, Ladin J, Weinberg J, *et al.* McKenzie lumbar classification: inter-rater agreement by physical therapists with different levels of formal McKenzie postgraduate training. *Spine*. 2014;39:E182–90.
- de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's κ . *BMJ*. 2013;346:f2125.
- Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, von Korf M, *et al.* Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine*. 2008;33:90–4.
- Altman DG. *Practical statistics for medical research*. London: Chapman & Hall; 1991.
- Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine*. 2000;25:2940–52.
- May S, Littlewood C, Bishop A. Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. *Aust J Physiother*. 2006;52:91–102.
- Laslett M. Evidence-based diagnosis and treatment of the painful sacroiliac joint. *J Man Manip Ther*. 2008;16:142–52.
- Long A, Donelson R, Fung T. Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine*. 2004;29:2593–602.
- Waddell G. *The back pain revolution*. 2nd ed. Edinburgh: Churchill Livingstone; 2004.
- May S, Aina A. Centralization and directional preference: a systematic review. *Man Ther*. 2012;17:497–506.
- Smeets RJ, Maher CG, Nicholas MK, Refshauge KM, Herbert RD. Do psychological characteristics predict response to exercise and advice for subacute low back pain? *Arthritis Rheum*. 2009;61:1202–9.
- Underwood MR, Morton V, Farrin A. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset [ISRCTN32683578]. *Rheumatology*. 2007;46:1297–1302.
- George SZ, Bialosky JE, Donald DA. The centralization phenomenon and fear-avoidance beliefs as prognostic factors for acute low back pain: a preliminary investigation involving patients classified for specific exercise. *J Orthop Sports Phys Ther*. 2005;35:580–8.
- Long A, May S, Fung T. The comparative prognostic value of directional preference and centralization: a useful tool for front-line clinicians? *J Man Manip Ther*. 2008;16:248–54.
- Werneke M, Hart DL. Centralization phenomenon as a prognostic factor for chronic low back pain and disability. *Spine*. 2001;26:758–64.
- Wetzel FT, Donelson R. The role of repeated end-range/pain response assessment in the management of symptomatic lumbar discs. *Spine J*. 2003;3:146–54.
- Laslett M. Manual correction of an acute lumbar lateral shift: maintenance of correction and rehabilitation: a case report with video. *J Man Manip Ther*. 2009;17:78–85.
- Donelson R, Grant W, Kamps C, Medcalf R. Pain response to sagittal end-range spinal motion. A prospective, randomized, multicentered trial. *Spine*. 1991;16:S206–12.

- 38 Albert HB, Hauge E, Manniche C. Centralization in patients with sciatica: are pain responses to repeated movement and positioning associated with outcome or types of disc lesions? *Eur Spine J.* 2012;21:630–6.
- 39 McKenzie R, May S. *The lumbar spine. Mechanical diagnosis & therapy*, 2nd ed. Waikanae: Spinal Publications New Zealand Ltd; 2003
- 40 Kilby J, Stigant M, Roberts A. The reliability of back pain assessment by physiotherapists, using a McKenzie algorithm. *Physiotherapy.* 1990;76:579–83.
- 41 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;30:368–83.
- 42 Clare HA, Adams R, Maher CG. Reliability of McKenzie classification of patients with cervical or lumbar pain. *J Manipulative Physiol Ther.* 2005;28:122–7.
- 43 Kilpikoski S, Airaksinen O, Kankaanpää M, Leminen P, Videman T, Alen M. Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine.* 2002;27:E207–14.
- 44 Fritz JM, Delitto A, Vignovic M, Busse RG. Interrater reliability of judgments of the centralization phenomenon and status change during movement testing in patients with low back pain. *Arch Phys Med Rehabil.* 2000;81:57–61.
- 45 Werneke M, Hart DL, Cook D. A descriptive study of the centralization phenomenon. A prospective analysis. *Spine.* 1999;24:676–83.