Reliability of procedures used in the physical examination of non-specific low back pain: A systematic review

Stephen May[1], Chris Littlewood[2] and Annette Bishop[3]

[1]Sheffield Hallam University [2]Wakefield West PCT/ Mid Yorkshire Hospitals NHS Trust [3]Keele University UK

The purpose of this systematic review was to determine the quality of the research and to assess the reliability of different types of physical examination procedures used in the assessment of patients with non-specific low back pain. A search of electronic databases (MEDLINE, PEDro, AMED, EMBASE, Cochrane, and CINAHL) up to August 2005 identified 48 relevant studies which were analysed for quality and reliability. Pre-established criteria were used to judge the quality of the studies and satisfactory reliability, and conclusions emphasised high quality studies (≥ 60% methods score). The mean quality score of the studies was 52% (range 0 to 88%), indicating weak to moderate methodology. Based on the upper threshold used (kappa/ICC > 0.85) most procedures demonstrated either conflicting evidence or moderate to strong evidence of low reliability. When the lower threshold was used (kappa/ICC > 0.70) evidence about pain response to repeated movements changed from contradictory to moderate evidence for high reliability. Most procedures commonly used by clinicians in the examination of patients with back pain demonstrate low reliability. **[May S, Littlewood C and Bishop A (2006): Reliability of procedures used in the physical examination of non-specific low back pain: A systematic review. *Australian Journal of Physiotherapy* 52: 91–102]**

Key Words: Lumbar Spine, Physical Examination, Reliability, Systematic Review

Introduction

In patients with low back pain the terms non-specific or mechanical back pain are used to describe an entity whose pathoanatomical aetiology is unknown (AHCPR 1994, CSAG 1994, Deyo 2002). Although some recent studies suggest that specific structural pathology can be diagnosed with physical examination procedures (Laslett et al 2003, Young et al 2003), such reports are unusual and clinicians often base management decisions on findings from the physical examination. Types of physical examination items used in the assessment of patients with back pain include observation, palpation, and symptom response.

Observation may be used to help determine movement loss, the presence and direction of a lateral shift, the shape of the lumbar lordosis, muscle bulk, the state of the soft tissues, or the presence of asymmetry (McKenzie 1981, Petty and Moore 2001, Sammut and Searle-Barnes 1998). Chiropractors use tests to look for a 'manipulative lesion' at a particular joint (Hestboek and Leboeuf-Yde 2000). Physiotherapists palpate to determine the segmental level and nature of the disorder (Maitland 1986). Palpation is used to determine the presence of reduced or excessive motion at a segmental level, and the direction and grade of mobilisation that is to be used in treatment (Jull et al 1994). Osteopathy uses the term somatic dysfunction to refer to subtle changes that occur in and around the motion segment, and which can be detected by palpation (Sammut and Searle-Barnes 1998).

Some physical examination procedures emphasise symptom response to movement or manual pressure to determine management strategies (McKenzie 1981, Van Dillen et al 1998, Wilson et al 1999). Physical examination procedures can also be used to classify patients into different sub-groups that will determine the management strategy to be provided. There are numerous classification systems available for non-specific low back pain (McCarthy et al 2004).

Thus fundamentally different ways are used to undertake the physical examination of the lumbar spine. Tests are being applied that are determining future management. To be of optimum clinical value such tests need to have certain measurement properties of validity and reliability (Streiner and Norman 1996). Whilst double-blinded injection studies have identified the site of pain generation accurately (Schwarzer et al 1994, Drefuss et al 1996) reports establishing the validity of physical examination against such 'gold standards' are few (Laslett et al 2003, Young et al 2003). Reliability of test procedures is also important. It is imperative that physical examination findings are interpreted by clinicians with a high level of reliability for them to have clinical utility. Only with this evidence should their use to formulate management strategies be endorsed. If intertester reliability is poor then management decisions following the physical examination are based on unsound judgements.

Reliability of physical examination procedures has been explored over previous decades by different professional groups involved in the management of low back pain (Bergstrom and Courtis 1986, Gonnella et al 1982, McConnell et al 1980). Some of these studies have been reviewed in a non-systematic way (Johnston 1982, Russell 1983, Simmonds and Kumar 1993a, Panzer 1992, Huijbregts 2002, van Genderen et al 2003) and four systematic reviews have also been published (Hestboek and Leboeuf-Yde 2000, van der Wurff et al 2000, Seffinger et al 2004, van Trijffel et al 2005). One was concerned with physical examination procedures applicable to the sacroiliac joint (van der Wurff et al 2000), one reviewed chiropractic tests (Hestboek and Leboeuf-Yde 2000), one evaluated the reliability of palpation procedures throughout the spine (Seffinger et al 2004), and one evaluated the reliability of passive assessment of intervertebral motion (van Trijffel et al 2005). Straight leg

Table 1. Criteria for assessing reliability studies.

Criteria		Weighting
Model/patient population: total 25		
1	Adequate description of study population—some clinical description, or gender and age if volunteers	4
2	Representative of clinical practice	4
3	Subjects selected randomly or consecutively	7
4	Number of subjects: < 25, score – 0 > 25, score – 3 > 50, score – 6 > 75, score – 10	10
Test procedure: total 35		
5	Procedure clearly described and reproducible	5
6	Procedure executed in uniform manner	5
7	Adequate measures to reduce bias: e.g. examiner blinded to other examiner, results sealed, independent adjudicator.	10
8	Highest level of examiners: experienced with procedure, score – 10 experienced clinicians, score – 5 students/juniors, score – 2	10
9	Consensus procedure prior to testing with pilot study	5
Test results: total 40		
10	More than one pair of examiners tested	10
11	Multiple testing between examiners	5
12	Standardised measure of test outcome – dichotomous or clear outcome of test	5
13	Frequencies of outcome and agreement reported	10
14	Appropriate inferential statistics eg kappa/ICC; and measure of variance	10
Total Score		**100**

Derived from: van der Wurff et al (2000) and Bogduk (2001)

raise (SLR) is not appropriate to non-specific back pain, and its reliability (Dixon and Keating 2000) and diagnostic accuracy (Deville et al 2000) have been reviewed. A review of the reproducibility of functional measures of the low back, including range of movement measurements, has recently been published (Essendrop et al 2003).

No systematic review has attempted to investigate the reliability of different types of physical examination procedures commonly used for non-specific back pain. Previous reviews (Seffinger et al 2004, van Trijffel et al 2005) have included studies using asymptomatic subjects that do not truly reflect the clinical value of a procedure and may produce reliability statistics that do not reflect the clinical environment. Further, previous studies have commonly used the Landis and Koch (1977) scale for interpreting kappa statistics, which suggest that 0.4 and above is acceptable for clinical utility, whereas others consistently suggest that only higher values are satisfactory (McDowell and Newell 1987, Streiner and Norman 2003).

Thus the aim of the review was to evaluate the reliability of different types of examination procedures used in the assessment of non-specific back pain.

Method

A literature search of MEDLINE (January 1966 to August 2005), PEDro (August 2005), AMED (1985 to August 2005), EMBASE (1974 to August 2005), the Cochrane Library (2005, Issue 2), and CINAHL (1982 to August 2005) was conducted using the following search terms: lumbar, spine, lumbar spine, back pain, back ache, low back pain; reliability, reproducibility, inter examiner, inter-examiner, inter tester, inter-tester, inter observer, inter-observer, kappa, intra class correlation, intra-class correlation, ICC; assessment, physical examination, physical tests, manual examination, palpation, observation, classification, pain response, symptom response; centralization. This was supplemented by hand searching the references lists of the articles found from the electronic search.

Articles had to meet a number of criteria for inclusion. Results must be published as full reports before August 2005 (abstracts were not included.) The study involved procedures of the lumbar spine used for clinical decision making for non-specific back pain. The study did not involve tests for specific back pain, such as neurological testing for nerve root problems. The study involved human subjects, rather

Table 2. Levels of evidence for intervention studies.

Strong	Consistent findings from multiple high quality trials (n = 3 or more).
Moderate	Consistent findings among low quality trials and/or one high quality trial.
Limited	One low quality trial.
Conflicting	Inconsistent findings among multiple trials.
No evidence	No trials.

than a simulated spine testing apparatus. The study involved physical examination procedures, rather than instrumented examination procedures (an exception was made with timed endurance, as the 'low-tech' tool, a stopwatch, was considered to be commonly available.) The study involved adults (> 18 years) with low back pain. The study had to be an intra and/or inter-examiner reliability design. The study had to be reported in English.

The studies were identified by one of the authors (SM), and admitted to the study with agreement with another author. A total of 104 studies were identified. Articles were then excluded for the following reasons: 22 used instrumentation, a measuring device, or a questionnaire; 16 did not involve subjects with low back pain; 12 were discussion papers about reliability issues; five involved procedures that were not used in the clinical decision making process; and one involved non-adult subjects. Forty-eight articles were directly relevant to the review. AB and CL independently scored each publication according to a standardised set of 14 methodological criteria; differences of opinion were settled by negotiation and consensus with SM. At the first round there was agreement between CL and AB of 79.5% on 672 decision items, kappa value 0.56, which were resolved with consensus.

There are no established or commonly used criteria for judging the quality of reliability studies. The authors of the systematic review (van der Wurff et al 2000) into the reliability of clinical tests of the sacroiliac joint devised a criteria checklist consisting of three categories: study population, test procedure, and test results. Modifications to this set of criteria have been suggested for determining reliability in diagnostic tests (Bogduk 2001), which retained the main categories, but deleted, contracted, or replaced some specific criteria. Several of these changes seemed to be relevant to the goals of this study. However several of the modifications concerned knowledge of subject status, whether condition positive or control, which was not relevant to this study. Thus the final methodological criteria were derived from both (Table 1), and were pilot tested on three studies before being applied to all.

The criteria (van der Wurff et al 2000, Bogduk 2001) for methodological assessment from which our criteria were adopted were weighted, and a similar weighting was applied in this study. The maximum score was 100 points; a trial was considered to be of higher quality if it scored 60% or more.

There is general agreement that if the outcome is nominal data, kappa is the appropriate statistic; if ordinal data, weighted kappa; and if continuous data, an intra-class correlation coefficient (ICC) or the Bland-Altman test (Streiner and Norman 2003, Bruton et al 2000, Eliasziw et al 1994, Lantz 1997, Haas 1991, Rankin and Stokes 1998, Altman 1991). Percentage agreement is generally considered to be inappropriate as this can be heavily affected by chance agreement. However there is on-going debate about the most appropriate statistical evaluation of reliability studies (Streiner and Norman 2003).

Both kappa and ICC are presented as numerical values less than 1.00. There is not a consensus about what constitutes a clinically acceptable level of reliability, but from a statistical perspective there is no reason to interpret reliability coefficients differently (Streiner and Norman 2003). Kappa has been interpreted as follows: 0.00 to 0.20 poor or slight agreement; 0.21 to 0.40 fair; 0.41 to 0.60 moderate; 0.61 to 0.80 substantial or good; 0.81 to 1.00 very good or almost perfect (Landis and Koch 1977). Several other authors have suggested that 0.4 or 0.6 may represent acceptable reliability (Haas 1991, Seffinger et al 2004, Landis and Koch 1977, Altman 1991); however this is not in agreement with other statistical authorities who demand higher values, especially when individuals rather than groups are being assessed (McDowell and Newell 1987, Streiner and Norman 2003). Streiner and Norman (2003, p146) suggest that a kappa value of '0.75 is a fairly minimal value for a useful instrument', and McDowell and Newell that 'values above 0.85 may be considered satisfactory' (1987, p32).

Regarding the ICC the closer to 1.00 the greater the reliability, but it has been argued that this numerical value cannot, by itself, give a true picture of reliability (Rankin and Stokes 1998). It has been recommended that any useful measure should have a correlation coefficient of at least 0.6 or 0.7 (Chinn 1991, Nunnally 1978). Other authors have recommended that when the scores of individuals, as opposed to groups, are being considered then coefficients of 0.85 or 0.90 are appropriate (McDowell and Newell 1996, Ware et al 1981, Weiner and Stewart 1984). ICC has been interpreted as follows: less than 0.40 poor reliability; 0.40 to 0.75 fair to good reliability; greater than 0.90 excellent reliability (Fleiss 1986).

It was decided, therefore, that the pre-determined criteria for satisfactory reliability would be 0.85 for both kappa and ICC. Given that such cut-off points are necessarily somewhat arbitrary it was determined that a sensitivity analysis would be conducted by lowering the cut-off point to 0.70.

When justifying such arbitrary criterion it would be helpful to understand how much error is tolerable in clinical practice with regard to these procedures; however, this is currently unknown. The physical examination procedures here are inherently safe but reliability error may compromise the diagnostic process which could compromise outcomes.

For intervention studies, levels of evidence adapted from van Tulder et al (2003) are detailed in Table 2.

Some studies reported on more than one physical examination procedure and thus occurred more than once in the results table. Where the results were presented as single kappa values or ICC these were presented in the results table. A number of studies listed multiple results on a particular procedure—in these instances the range of results were presented. Studies that used only percentages were listed but results were not reported. As well as the kappa or ICC the variance was presented if available. Conclusions concerning reliability were made with reference to the quality of the studies.

Results

Of the 48 studies included (listed in alphabetical order by first author in Table 3) 47 involved analysis of intertester reliability, nine involved intratester reliability as well, and one involved intratester reliability only. Multiple professional groups were evaluated: physiotherapists most commonly (32 studies); chiropractors (8), physicians (3) and osteopaths (1) less commonly, and diverse professional groups in 4 studies. Studies evaluated the reliability of palpation (24 studies), symptom response (23), observation (15), classification systems (12), or timed muscle endurance (3); with some studies evaluating more than one aspect.

Characteristics of the included studies are displayed in Tables 4 and 5. The mean quality score was 52% (Table 3); common weaknesses involved subjects not being representative of clinical practice, subjects selected other than randomly or consecutively, subjects few in number, procedures not clearly standardised, and failure to provide adequate measures to reduce possible bias. However 40% of studies scored more than 60% and were considered higher quality (totals in bold, Table 3).

Most studies used kappa or ICC, a few failed to provide adequate statistical analysis, reporting percentage values only, or percentage agreement greater than that expected by chance. These studies were listed, but excluded from the strength of evidence conclusions. Overall levels of intertester reliability are listed in Table 4.

Palpation There was conflicting evidence indicating reliability of identifying the spinal level in two high quality studies. There was moderate evidence about the low reliability of passive accessory movements from two high quality studies and three other studies. There was moderate evidence about the low reliability of establishing comparable level and passive physiological movements from two high quality studies and one other. There was conflicting evidence about the reliability of evaluating muscle tension or spasm from three high quality studies, one study reported two results (four results in total); three results demonstrated lack of reliability, and one a range of kappa values that crossed the 0.85 threshold. There was moderate evidence about the low reliability in determining the existence of a fixation or 'manipulative' lesion in eight low quality studies. There was conflicting evidence about the reliability of 'instability' tests in one high quality study, with the prone instability test demonstrating reasonable reliability.

Symptom response There was conflicting evidence regarding the reliability of pain response to repeated movements in two high quality studies and four other studies. There was strong evidence indicating low reliability of pain on movement from three high quality studies and one other. There was strong evidence for the low reliability of pain on palpation and trigger points from six high quality studies and seven others; however, in one high quality study and four others the upper range of kappa values or the upper 95% confidence interval was greater than 0.85.

Observation There was strong evidence indicating low reliability of detecting a lateral shift by observation from five high quality studies and two others. There was conflicting evidence about making judgements on lordosis by observation from two high quality studies, with two others demonstrating lack of reliability. There was moderate evidence about the low reliability in evaluating abnormal movement, posture, or coupling patterns from one high quality study and four others. There was moderate evidence indicating high reliability of timed muscle endurance from three low quality studies.

Classification systems There was conflicting evidence regarding reliability of the McKenzie classification system from three high quality studies; two reported reliability greater than 0.85, one did not. There was conflicting evidence about the reliability of the movement impairment system of classification, at least regarding symptom behaviour, from two high quality studies. There was moderate evidence indicating low reliability of the treatment-based classification system from three low quality studies. There was limited evidence indicating low reliability of the Canadian Back Institute classification system from one low quality study. There was moderate evidence indicating low reliability of the diagnostic classification system from one high quality study; however in three or nine of 15 decisions kappa values or 95% confidence intervals respectively exceeded 0.85.

When the lower threshold of acceptable reliability was applied only one clear change occurred in the conclusions. Evidence regarding pain response to repeated movements changed from contradictory to moderate evidence for reliability in two high quality and two other studies. The weight of evidence shifted in two other instances without creating definite changes. Moderate evidence indicating low reliability in evaluating abnormal movement became less clear with one high quality study demonstrating low reliability and one other study demonstrating reliability. The conclusion about pain on palpation, strong evidence for low reliability of pain on palpation and trigger points, became less clear—in three high quality studies and six others the upper range of kappa values or the upper 95% confidence interval was greater than 0.70.

Intratester reliability was evaluated less often and conclusions from high quality studies were very limited (Table 5). There was moderate evidence indicating high reliability of timed muscle endurance from three low quality studies. There was moderate evidence indicating low reliability of observation from one high quality and one other study. For palpation one high quality study included kappa values that nearly crossed the 0.85 threshold, one other study did, but three other studies did not. For pain response one high quality study included kappa values and 95% confidence intervals that crossed the 0.85 threshold.

Discussion

The general quality of the research into reliability of procedures is moderate at best. In 50% or more of the studies

Table 3. Reliability testing—methods score.

	Reference*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total**
1	Binkley	4	4	0	0	5	0	0	10	5	10	5	5	0	10	58
2	Boline 1988	0	0	0	3	0	0	10	5	5	0	0	5	10	0	38
3	Boline 1993	0	0	0	3	0	0	0	5	0	0	5	5	10	0	28
4	Clare 2003	4	0	0	3	5	5	10	10	0	10	5	5	0	10	**67**
5	Clare 2005	4	4	0	3	0	0	10	10	0	10	5	5	10	10	**71**
6	Cook	4	4	0	0	0	0	0	5	0	10	5	5	0	0	33
7	Delitto	0	0	0	6	5	0	0	10	5	10	5	5	0	0	46
8	Donahue	4	4	7	3	5	5	0	2	5	10	5	5	10	0	**65**
9	Downey 1999	0	0	0	6	5	0	10	10	0	10	5	5	0	10	**61**
10	Downey 2003	0	0	0	6	5	0	10	10	0	10	5	5	0	10	**61**
11	Fedorak	4	0	0	3	5	5	0	10	0	10	5	5	0	10	57
12	French	4	4	0	0	0	0	0	10	0	10	5	5	10	0	48
13	Fritz 2000a	0	0	0	0	5	5	0	5	5	10	5	5	0	10	50
14	Fritz 2000b	4	4	0	3	0	0	0	10	0	10	0	5	10	0	46
15	Fritz 2003	4	4	0	0	5	0	0	10	0	0	0	5	0	10	38
16	Haswell	4	4	7	3	5	5	10	5	5	10	5	5	10	10	**88**
17	Hawk	4	0	0	0	0	0	10	10	5	10	5	5	10	0	59
18	Heiss	4	4	0	3	0	0	10	5	0	10	5	0	10	0	51
19	Hicks	4	4	7	6	5	5	0	10	5	10	0	5	10	10	**81**
20	Horneij 2002a	0	0	0	6	5	5	0	10	5	10	5	5	10	10	**61**
21	Horneij 2002b	4	0	0	0	5	5	0	5	0	0	0	5	0	10	34
22	Hsieh	4	0	0	6	5	5	10	5	5	10	5	5	10	0	**70**
23	Inscoe	4	0	0	0	5	0	0	10	0	0	0	5	10	0	34
24	Keating	0	0	0	3	0	0	0	5	0	0	5	5	10	0	28
25	Keller	4	0	0	6	5	5	0	5	0	0	0	5	10	0	40
26	Kilby	0	4	0	3	0	0	10	5	0	0	0	5	10	0	37
27	Kilpikoski	4	4	7	3	0	0	10	10	0	0	0	5	10	0	53
28	Latimer	4	0	0	6	5	5	0	5	5	10	0	5	0	10	55
29	Ljungquist: Inter	4	0	0	0	5	5	0	10	0	0	0	5	0	10	39
	Ljungquist: Intra	4	0	0	3	5	5	0	10	0	0	5	5	0	10	47
30	Maher	4	4	0	10	5	0	10	10	0	10	0	5	0	10	**68**
31	McCombe	0	0	7	10	5	0	0	10	0	10	0	5	0	10	57
32	McConnell	0	0	0	0	0	0	0	5	0	10	5	5	0	0	25
33	Mootz	0	0	0	6	5	0	0	10	5	0	0	5	10	0	41
34	Nelson	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	Nice	4	4	0	3	5	5	0	10	5	10	0	5	10	10	**71**
36	Njoo	4	4	0	10	5	0	0	5	5	10	5	5	10	10	**73**
37	Petersen	4	4	0	10	5	0	10	5	0	10	5	5	10	10	**78**
38	Razmjou	4	4	7	3	5	0	10	10	5	0	0	5	10	0	**63**
39	Riddle	4	4	0	10	0	0	10	5	5	10	0	5	10	0	**63**
40	Rhudy	0	0	0	0	0	0	0	10	0	0	0	5	0	0	15
41	Sebastian	4	4	0	3	0	0	0	10	0	0	5	5	10	0	41
42	Seymour	0	0	0	0	0	0	10	5	5	10	5	5	10	0	50
43	Spratt	4	4	0	3	0	0	0	0	0	0	5	5	10	0	31
44	Strender	4	4	0	6	5	5	0	10	5	10	0	5	10	10	**74**
45	Van Dillen	4	4	0	10	0	5	0	5	5	10	5	5	10	0	**63**
46	Waddell	0	0	0	3	5	0	0	10	5	0	0	5	0	0	28
47	White	4	4	0	3	5	0	10	5	5	10	5	5	10	10	**76**
48	Wilson	4	4	0	10	5	0	10	10	0	10	5	5	10	0	**73**
Possible points		**4**	**4**	**7**	**10**	**5**	**5**	**10**	**10**	**5**	**10**	**5**	**5**	**10**	**10**	**100**
Mean		**2.7**	**2**	**.9**	**3.9**	**3**	1.6	**3.7**	**7.5**	**2.1**	**6.3**	**2.9**	**4.8**	**6.1**	**4.5**	**51.7**

*first author, date if more than 1 paper by author **totals in **bold** are higher quality studies (≥ 60).

Table 4. Intertester reliability studies—items of physical examination.

Item	Reference	Statistic[1]	Variance[2]
Palpation			
Identifying spinal level	Binkley et al 1995	ICC: 0.69	0.53 to 0.82
	Downey et al 1999	**0.92**	**0.86 to 0.98**[f]
	Downey et al 2003	0.09	–0.05 to 0.23[f]
Passive accessory movements	Binkley et al 1995	ICC: 0.25	0.0 to 0.39
	Hicks et al 2003	–0.02 to 0.26	–0.26 to 0.53
	Inscoe et al 1995	SP: 0.18	
	Keating et al 1990	–0.18 to 0.31	
	Maher & Adams 1994	ICC: 0.03 to 0.37	0.18 to 0.53
Comparable level	Downey et al 2003	0.37	0.19 to 0.55[f]
Passive physiological movements	Keating et al 1990	–0.13 to 0.47	
	Strender et al 1997[g]	0.54 and 0.75	0.26 to **0.94**[f]
	Strender et al 1997[h]	–0.08	–0.50 to 0.34[f]
Muscle tension	Boline et al 1988	0.10 to 0.31	
	Horneij et al 2002a[d]	0.03 to 0.68	–0.22 to **0.92**
	Horneij et al 2002a[e]	0.56 to **0.94**	0.07 to **1.00**
	Keating et al 1990	–0.14 to 0.32	
Muscle spasm	Hsieh et al 2000	0.06	
	Strender et al 1997[g]	0.50 to 0.73	0.30 to **0.90**[f]
	Strender et al 1997[h]	0.43	0.0 to **0.86**[f]
'Manipulative lesion'/fixation	Boline et al 1988	–0.05 to 0.31	
	Boline et al 1993	–0.30 to 0.56	
	French et al 2000	0.27	
	Hawk et al 1999	–0.42 to 0.44	
	Mootz et al 1989	–0.17 to 0.17	
	Sebastian & Chovvath 2004	0.69	
Misalignment	Keating et al 1990	–0.28 to 0.22	
Subluxation	Rhudy et al 1988	11/15 < 0.20	
Instability tests			
—Posterior shear	Hicks et al 2003	0.35	0.20 to 0.51
—Prone instability	Hicks et al 2003	**0.87**	0.80 to **0.94**
Pain response			
Pain response to repeated movements	Spratt et al 1990		
Centralisation	Fritz et al 2000a	0.79	0.78 to 0.81
	Kilby et al 1990	0.51	
	Kilpikoski et al 2002	0.7	
Directional preference	Kilpikoski et al 2002	**0.9**	
Relevance of lateral shift/ component	Donahue et al 1996	0.74	
	Kilpikoski et al 2002	0.7/0.4	
	Razmjou et al 2000	**0.85/0.95**	
	Seymour et al 2002	0.56	
Pain on movement	Haswell et al 2004	0.17 to 0.60	–0.08 to 0.79
	Hicks et al 2003	0.61 to 0.69	0.44 to 0.84
	McCombe et al 1989	0.10 to 0.58	–0.13 to **0.90**[f]
	Spratt et al 1990		
	Strender et al 1997[g]	0.51 to 0.76	0.31 to **0.96**[f]
	Strender et al 1997[h]	0.06 to 0.71	–0.1 to **1.00**[f]
Pain on palpation	Boline et al 1988	–0.03 to 0.49	
	Fritz & Piva 2003	0.35	–0.33 to **1.00**
	Hicks et al 2003	0.25 to 0.55	0.11 to 0.67
	Horneij et al 2002a[d]	0.36 and 0.41	0.07 to 0.70
	Horneij et al 2002a[e]	0.49	0.15 to 0.83
	Maher & Adams 1994	ICC 0.67 to 0.73	0.55 to 0.81

Item	Reference	Statistic[1]	Variance[2]
	McCombe et al 1989	0.11 to 0.50	–0.07 to 0.80[f]
	Spratt et al 1990		
	Strender et al 1997[a]	0.22 and 0.27	–0.11 to 0.55[f]
	Strender et al 1997[b]	0.40 to 0.56	7.07 to 0.79[f]
	Strender et al 1997[c]	0.24 to 0.38	–0.17 to 0.66[f]
	Waddell et al 1982	0.63 and **1.00**	
Osseous pain	Keating et al 1990	0.19 to 0.66	
	Boline et al 1993	0.48 to **0.90**	
Soft tissue pain	Boline et al 1993	0.40 to 0.78	
	Horneij et al 2002a[d]	0.28 to 0.63	–0.02 to **0.92**
	Horneij et al 2002a[e]	0.49 to **0.88**	0.07 to **1.00**
	Keating et al 1990	0.13 to 0.59	
Trigger point symptoms	Hsieh et al 2000	0.008 to 0.33	
	Nice et al 1992	0.29 to 0.38	–0.20 to 0.67[f]
	Njoo & Does 1994	–0.02 to 0.73	–0.99 to 0.96
Observation			
Timed muscle endurance			
—Abdominals	Horneij et al 2002b	ICC: **1.00**	**1.00** to **1.00**
	Ljungquist et al 1999		
—Back extensors	Horneij et al 2002b	ICC: **1.00**	**1.00** to **1.00**
	Latimer et al 1999	ICC: **0.85**	0.76 to **0.90**
Lateral shift/direction	Clare et al 2003	0.26 to 0.38	0.25 to 0.39
	Donahue et al 1996	0.00	
	Kilpikoski et al 2002	0.2/0.4	
	Razmjou et al 2000	0.52	
	Riddle & Rothstein 1993	0.26	
	Strender et al 1997	0.13 and 0.39	–0.10 to 0.67[f]
	Waddell et al 1982	0.53	
Lordosis	Fedorak et al 2003	0.16	0.00 to 0.48
	Razmjou et al 2000	**1.0**	
	Strender et al 1997	0.32	–0.26 to **0.90[f]**
	Waddell et al 1982	0.71	
Visual observation of abnormality	Boline et al 1993	0.34 to 0.84	
	Keating et al 1990	0.29	
Disturbance of normal lumbopelvic rhythm	Waddell et al 1982	0.82	
	Hicks et al 2003	0.00 to 0.25	–0.15 to 0.60
Coupling patterns	Cook et al 2004	0.02 to 0.17	
Classifications			
McKenzie syndromes/sub-syndromes	Kilpikoski et al 2002	0.6/0.7	
	Razmjou et al 2000	0.7/**0.96**	0.45 to **0.96**/**0.88** to **1.00**
	Riddle & Rothstein 1993	0.26	
	Clare et al 2005	**1.0**/**0.89**	0.35 to **1.00**/0.66 to **1.00**
Movement system impairment categories	Van Dillen et al 1998[k]	0.0 to 0.78	
	Van Dillen et al 1998[l]	**0.87** to **1.0**	
	White & Thomas 2002	0.02 to 0.62	–0.39 to **0.87**
Treatment-based classification system	Delitto et al 1992[m]	PCC: –0.1 to 0.53	
	Fritz & George 2000	0.49 and 0.56	
	Heiss et al 2004	0.14 and 0.15	
Canadian Back Institute system	Wilson et al 1999	0.61	
Diagnostic classification system	Petersen et al 2004	0.26 to **1.00**	–0.19 to **1.00**

[1]statistic is kappa or weighted kappa unless ICC (Intra-class Correlation Coefficient), PCC (Pearson Correlation Coefficient) or SP (Scott's Pi); where no statistic only % frequency was stated or mean difference and limits of agreement (Ljungquist et al. 1999). Bolded figures > 0.85. Horneij et al 2002: [d] = first stage; [e] = following further standardisation of unreliable tests. Strender et al 1997: [a] = paravertebral tenderness; [b] = intersegmental tenderness; [c] = pain or decreased elasticity; [g] = testing by physiotherapist; [h] = testing by physician. Data for individual items could not be extracted from: McConnell et al 1980, Nelson et al 1979. Van Dillen et al 1998: [k] = alignment and movement; [l] = symptom behaviour; Delitto et al 1992: [m] = mobilisation/flexion/extension category. [2] 95% confidence interval where given, or [f] = calculated from original data.

Table 5. Intratester reliability studies—items of physical examination.

Item	Reference	Statistic[1]	Variance[2]
Palpation			
Passive accessory movements	Inscoe et al 1995	SP: 0.42 to 0.61	
Muscle evaluation–tension	Horneij et al 2002a	0.18 to 0.84	–0.25 to **1.00**
'Manipulative lesion'/fixation	French et al 2000	0.47	
	Hawk et al 1999	–0.17 to **0.85**	
	Mootz et al 1989	–0.09 to 0.48	
Pain response			
Pain on palpation	Horneij et al 2002a	0.56 / 0.78	0.18 to **1.00**
Soft tissue pain	Horneij et al 2002a	0.61 to **0.87**	0.21 to **1.00**
Observation			
Timed muscle endurance			
—Abdominals	Horneij et al 2002b	ICC: **0.95**	**0.85** to **0.98**
	Ljungquist et al 1999		
—Back extensors	Horneij et al 2002b	ICC: **0.91/0.92**	0.79 to **0.97**
	Keller et al 2001	ICC: **0.93**	
	Ljungquist et al 1999		
Lateral shift	Clare et al 2003	ICC: 0.48 to 0.59	0.43 to 0.63
Lordosis	Fedorak et al 2003	0.50	0.02 to **0.98**

[1]statistic is kappa unless ICC (Intraclass Correlation Coefficient), or SP (Scott's Pi); where no statistic mean difference and limits of agreement (Ljungquist et al 1999). Bolded figures ≥ 0.85. [2]95% confidence interval where given.

limitations included: subjects not being representative of clinical practice, subjects selected other than randomly or consecutively, procedures not clearly standardised, failure to provide adequate measures to reduce bias, failure to use a pilot study to establish consensus, and failure to report appropriate inferential statistic *and* a measure of variance. Between 35% and 49% of studies failed to: give an adequate description of the procedure, include more than one pair of examiners, perform multiple testing between examiners, or report frequencies and agreement rate.

The common types of physical examination procedures that were evaluated were classified as palpation, symptom response, timed muscle endurance, observation, or classification. The pre-established criteria for reliability were based on a value equal to or greater than a reliability coefficient of 0.85, with special emphasis being placed on studies with methods scores equal to or greater than 60%. Using these criteria most procedures commonly used in the examination of the lumbar spine were found to have moderate or strong evidence for low reliability between clinicians or to have conflicting evidence. There was moderate evidence about the high reliability of timed muscle endurance from 3 low quality studies; a procedure in which subjective decision making is minimal. When the lower threshold for acceptable reliability of 0.7 was applied evidence concerning pain response to repeated movements changed from contradictory to moderate evidence for high reliability in two high quality and two other studies. The sensitivity analysis was somewhat arbitrary, but appeared to represent a reasonable level of reliability coefficient without the compromise inherent in lower values.

The present review involved a systematic and transparent analysis of reliability studies. It is the first systematic review to focus on physical examination procedures used by all professions in the examination of the lumbar spine, and the first to make an analysis of different types of physical examination procedures, using the quality of studies to make conclusions about the levels of evidence. Furthermore only studies that evaluated reliability in patients with low back pain were used, and a high threshold for satisfactory reliability was used. There are not standardised and widely accepted tools for scoring the quality of such studies. However the methods scores used were derived from previous work (van der Wurff et al 2000, Bogduk 2001), and the quality assessment tool used in this review is similar to that used in another recent publication (Seffinger et al 2004).

A consistent finding from work in this field is the generally low reliability of palpation-based assessment, and the moderate reliability of some examination procedures based on symptom response (Hestboek and Leboeuf-Yde 2000, van der Wurff et al 2000, Seffinger et al 2004, van Trijffel et al 2005). This review found high reliability for timed muscle endurance and highlighted the potential value of symptom response with repeated movements.

Reliability is not an absolute property but is dependent on a number of variables such as the population in which the procedure is being tested, the prevalence of the attribute, the bias index, and the threshold of how much reliability is 'good enough' (Sim and Wright 2005, Streiner and Norman 2003). As it is not possible or appropriate to define 'reliability' or 'unreliability' in absolute kappa or ICC values (Haas 1991) the cut-off points used in this review were necessarily somewhat arbitrary, but were higher than those applied in previous reviews and in line with contemporary interpretations of reliability coefficients (McDowell and Newell 1987, Streiner and Norman 2003).

Certain caveats should be made about how the results of this review affect clinical judgements. Direct comparison

Table 6. Improving reliability studies.

1.	Subjects should be patients with the condition of interest, representative of clinical practice and recruited randomly or consecutively.
2.	Procedures should have clear operational definitions with uniform and reproducible descriptions.
3.	Several measures should be used to reduce bias, for example the examiner is blinded to the other examiner; results are sealed; there is an independent adjudicator.
4.	Large numbers of patients (at least 50), and multiple pairs of testers.
5.	A standardised measure for test outcome.
6.	Frequencies of outcome and agreement must be reported so that prevalence and bias can be considered.
7.	The appropriate inferential statistic should be reported with a measure of variance.
8.	The study should be reported following the STARD statement (Bossuyt et al 2003).

of reliability studies is inappropriate given differences in statistical analysis, experimental design, prevalence, bias, and dissimilar scales (Haas 1991, Sim and Wright 2005). Kappa values are affected by prevalence of sub-groups in the sample and by 'observer expectation bias' (Lantz 1997), and are not simply a measure of the reliability of a procedure. With a large prevalence index the kappa value is lower than when the prevalence index is low; when there is a large bias kappa is higher than when bias is low (Sim and Wright 2005). Because reliability is not an absolute property with clearly defined numerical values, qualitative judgements between examination procedures are inappropriate. Interpretation of reliability studies is also affected by the purpose and the amount of error acceptable in a decision making process. Whilst perfect agreement is 'blatantly absurd' and a certain amount of error is unfortunate but unavoidable (Lantz 1997), if clinical management decisions are to be made from these procedures then using measures with moderate or high levels of reliability is clearly preferred to using measures with low levels of reliability. Low levels of reliability may compromise the diagnostic process which could compromise clinical outcomes.

Reliability has been defined as a ratio of subject variance to subject and error variance, and therefore the way to improve reliability is to reduce error variance (Streiner and Norman 2003). This may be done by standardising tests, providing clear operational definitions for their use and interpretation, training users, and making use of procedures that demonstrate greater reliability. Traditionally and intuitively it has been argued that reliability is a necessary precursor to validity, and that clinical procedures must be reliable to be used effectively in management decisions.

There are important research implications given that the quality of the literature in this field has generally been found to be moderate at best. It has been suggested that internationally acceptable methods need to be established for the way reliability research should be performed and reported (van Genderen et al 2003). Some of the limitations of previous reliability studies have been detailed above against the methodological criteria we used to judge quality. Key suggestions for an improvement in the conduct and reporting of reliability studies are provided in Table 6.

Conclusions

This systematic review identified 48 studies that evaluated the reliability of physical examination procedures used in the assessment of the lumbar spine for non-specific low back pain. The methodological quality was only moderate, and conclusions emphasised the findings from high quality studies, defined as ≥ 60% methods score. Many commonly used examination procedures were found either to lack reliability or to have conflicting evidence about their reliability. Timed muscle endurance tests and evaluation of symptom response during repeated movements may be exceptions.

Correspondence Stephen May, Faculty of Health and Wellbeing, Sheffield Hallam University, Sheffield, UK. Email: s.may@shu.ac.uk

References

AHCPR (Agency for Health Care Policy and Research) (1994): Acute Low Back Problems in Adults. Bigos S, Bowyer O, Braen GR et al (Eds): Department of Health and Human Services, USA.

Altman DG (1991): Practical Statistics for Medical Research. London: Chapman & Hall.

Bergstrom E and Courtis G (1986): An inter- and intra-examiner reliability study of motion palpation of the lumbar spine in lateral flexion in the seated position. *European Journal of Chiropractic* 34: 121–141.

Binkley J, Stratford PW and Gill C (1995): Interrater reliability of lumbar accessory motion mobility testing. *Physical Therapy* 75: 786–795.

Bogduk N (2001): Unpublished instruments developed in the course of the Australian National Musculoskeletal Medicine Initiative. (Personal communication.)

Boline PD, Keating JC, Brist J and Denver G (1988): Interexaminer reliability of palpatory evaluations of the lumbar spine. *American Journal of Chiropractic Medicine* 1: 5–11.

Boline PD, Haas M, Meyer JJ, Kassak K, Nelson C and Keating JC (1993): Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality: Part II. *Journal of Manipulative and Physiological Therapeutics* 16: 363–374.

Bossuyt PM, Reitsma JB, Bruns DE et al (2003): The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry* 49: 7–18.

Bruton A, Conway JH and Holgate ST (2000): Reliability: What is it, and how is it measured? *Physiotherapy* 86: 94–99.

Chinn S (1991): Repeatability and method comparison. *Thorax* 46: 454–456.

Clare HA, Adams R, Maher CG (2003): Reliability of detection of lumbar lateral shift. *Journal of Manipulative and Physiological Therapeutics* 26: 476–480.

Clare HA, Adams R, Maher CG (2005): Reliability of McKenzie classification of patients with cervical or lumbar pain. *Journal of Manipulative and Physiological Therapeutics* 28: 122–127.

Cook C, Stickley L, Akram N, Benavides Y, Renz C and Ramey K (2004): Inter-rater reliability of coupling pattern observations of the pathological lumbar spine: a pilot study. *Journal of Manual and Manipulative Therapy* 12: 192–198.

CSAG (Clinical Standards Advisory Group) (1994): Back Pain. London: HMSO.

Delitto A, Erhard RE and Bowling RW (1995): A treatment-based classification approach to low back syndrome: Identifying and staging patients for conservative treatment. *Physical Therapy* 75: 470–489.

Delitto A, Shulmann AD, Rose SJ et al (1992): Reliability of a clinical examination to classify patients with low back syndromes. *Physical Therapy Practice* 1: 1–9.

Deville WLJM, van der Windt DAWM, Dzaferagic A, Bezemer PD and Bouter LM (2000): The test of Lasegue. Systematic review of the accuracy in diagnosing herniated discs. *Spine* 25: 1140–1147.

Deyo RA (2002): Diagnostic evaluation of LBP. Reaching a specific diagnosis is often impossible. *Archives of Internal Medicine* 162: 1444–1448.

Dixon JK and Keating JL (2000): Variability in straight leg raise measurements. *Physiotherapy* 86: 361–370.

Donahue MS, Riddle DL and Sullivan MS (1996): Intertester reliability of a modified version of McKenzie's lateral shift assessment obtained on patients with low back pain. *Physical Therapy* 76: 706–726.

Downey BJ, Taylor NF and Niere KR (1999): Manipulative physiotherapists can reliably palpate nominated lumbar spinal levels. *Manual Therapy* 44: 151–156.

Downey B, Taylor N and Niere K (2003): Can manipulative physiotherapists agree on which level to treat based on palpation? *Physiotherapy* 89: 74–81.

Dreyfuss P, Michaelsen M, Pauza K, McLarty J and Bogduk N (1996): The value of medical history and physical examination in diagnosing sacroiliac joint pain. *Spine* 21: 2594-2602.

Eliasziw M, Young SL, Woodbury MG and Fryday-Field K (1994): Statistical methodology for the concurrent assessment of interrater and Intrarater reliability: Using goniometric measurements as an example. *Physical Therapy* 74: 777–788.

Essendrop M, Maul I, Laubli T, Riihimaki H and Schibye B (2003): Measures of low back function: A review of reproducibility studies. *Physical Therapy in Sport* 4: 137–151.

Fedorak C, Ashworth N, Marshall J and Paull H (2003): Reliability of the visual assessment of cervical and lumbar lordosis: How good are we? *Spine* 28: 1857–1859.

Fleiss J (1986): The Design and Analysis of Clinical Experiments. New York: Wiley.

French SD, Green S and Forbes A (2000): Reliability of chiropractic methods commonly used to detect manipulable lesions in patients with chronic low-back pain. *Journal of Manipulative and Physiological Therapeutics* 23: 231–238.

Fritz JM, Delitto A, Vignovic M and Busse RG (2000a): Interrater reliability of judgements of the centralization phenomenon and status change during movement testing in patients with low back pain. *Archives of Physical Medicine and Rehabilitation* 81: 57–61.

Fritz JM, George S (2000b): The use of a classification approach to identify subgroups of patients with acute low back pain. Interrater reliability and short-term treatment outcomes. *Spine* 25: 106–114.

Fritz JM and Piva SR (2003): Physical impairment index: reliability, validity, and responsiveness in patients with acute low back pain. *Spine* 28: 1189–1194.

Gonnella C, Paris SV and Kutner M (1982): Reliability in evaluating passive intervertebral motion. *Physical Therapy* 62: 436–444.

Haswell K, Williams M and Hing W (2004): Interexaminer reliability of symptom-provoking active sidebend, rotation and combined movement assessments of patients with low back pain. *Journal of Manual and Manipulative Therapy* 12: 11–20.

Hawk C, Phongphuna C, Bleecker J, Swank L, Lopez D and Rubley T (1999): Preliminary study of the reliability of assessment procedures for indications for chiropractic adjustments of the lumbar spine. *Journal of Manipulative and Physiological Therapeutics* 22: 382–389.

Heiss DG, Fitch DS, Fritz JM, Sanchez WJ, Roberts KE and Buford JA (2004): The interrater reliability among physical therapists newly trained in a classification system for acute low back pain. *Journal of Orthopaedic and Sports Physical Therapy* 34: 430–439.

Hestboek L and Leboeuf-Yde C (2000): Are chiropractic test for the lumbo-pelvic spine reliable and valid? A systematic critical literature review. *Journal of Manipulative and Physiological Therapeutics*. 23: 258–275.

Hicks GE, Fritz JM, Delitto A and Mishock J (2003): Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Archives of Physical Medicine and Rehabilitation* 84: 1858–1864.

Horneij E, Hemborg B, Johnsson B and Ekdahl C (2002a): Clinical tests on impairment level related to low back pain: A study of test reliability. *Journal of Rehabilitation Medicine* 34: 176–182.

Horneij E, Holmstrom E, Hemborg B, Isberg PE and Ekdahl C (2002b): Inter-rater reliability and between-days repeatability of eight physical performance tests. *Advances in Physiotherapy* 4: 146–160.

Hsieh CYJ, Hong CZ, Adams AH et al (2000): Interexaminer reliability of the palpation of trigger points in the trunk and lower limb muscles. *Archives of Physical Medicine and Rehabilitation* 81: 258–264.

Huijbregts PA (2002): Spinal motion palpation: A review of reliability studies. *Journal of Manipulative and Physiological Therapeutics* 10: 24–39.

Inscoe EL, Witt PL, Gross MT and Mitchell RU (1995): Reliability in evaluating passive intervertebral motion of the lumber spine. *Journal of Manual and Manipulative Therapy* 3: 135–143.

Johnston W (1982): Interexaminer reliability studies: Spanning a gap in medical research—Louisa Burns Memorial Lecture. *Journal of the American Osteopathic Association* 81: 819–829.

Jull G, Treleaven J and Versace G (1994): Manual examination: Is pain provocation a major cue for spinal dysfunction? *Australian Journal of Physiotherapy* 40: 159–165.

Keating JC, Bergmann TF, Jacobs GE, Finer BA and Larson K (1990): Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. *Journal of Manipulative and Physiological Therapeutics* 13: 463–470.

Keller AK, Hellesnes J and Brox JI (2001): Reliability of the isokinetic trunk extensor test, Biering-Sorensen test, and Astrand bicycle test. *Spine* 2001:771-777.

Kilby J, Stigant M and Roberts A (1990): The reliability of back pain assessment by physiotherapists, using a 'McKenzie algorithm'. *Physiotherapy* 76: 579–583.

Kilpikoski S, Airaksinen O, Kankaanpaa M, Leminen P, Videman T and Alen M (2002): Inter-tester reliability of low back pain assessment using the McKenzie method. *Spine* 27: E207–E214.

Landis JR and Koch GG (1977): The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.

Lantz CA (1997): Application and evaluation of the kappa statistic in the design and interpretation of chiropractic clinical research. *Journal of Manipulative and Physiological Therapeutics* 20: 521–528.

Laslett M, Young SB, Aprill CN and McDonald B (2003): Diagnosing painful sacroiliac joints: A validity study of the McKenzie evaluation and sacroiliac provocation tests. *Australian Journal of Physiotherapy* 49: 89–97.

Latimer J, Maher CG, Refshauge K and Colacco I (1999): The reliability and validity of the Biering-Sorensen test in asymptomatic subjects and subjects reporting current or previous non-specific low back pain. *Spine* 24: 2085–2090.

Ljungqist T, Harms-Ringdahl K, Nygren A and Jensen I (1999): Intra- and inter-tester reliability of an 11-test package for assessing dysfunction due to back or neck pain. *Physiotherapy Research International* 4: 214–232.

Maher C and Adams R (1994): Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Physical Therapy* 74: 801–811.

Maitland GD (1986): Vertebral Manipulation (5th Edn). London: Butterworths.

McCarthy CJ, Arnall FA, Strimpkos N, Freemont A and Oldham JA (2004): The biopsychosocial classification of non-specific low back pain: A systematic review. *Physical Therapy Reviews* 9: 17–30.

McCombe PF, Fairbank JCT, Cockersole BC and Pynsent PB (1989): Reproducibility of physical signs in low-back pain. *Spine* 14: 908–918.

McConnell DG, Beal MC, Dinnar U et al (1980). Low agreement of findings in neuromusculoskeletal examinations by a group of osteopathic physicians using their own procedures. *Journal of the American Osteopathic Association* 79: 441–450.

McDowell I and Newell C (1987): Measuring Health: A Guide to Rating Scales and Questionnaires (2nd Edn). New York: Oxford University Press.

McKenzie RA (1981): The Lumbar Spine. Mechanical Diagnosis and Therapy. Waikanae: Spinal Publications.

Mootz RD, Keating JC, Kontz HP, Milus TB and Jacobs GE (1989): Intra and interobserver reliability of passive motion mobilisation of the lumbar spine. *Journal of Manipulative and Physiological Therapeutics* 12: 440–445.

Najm WI, Seffinger MA, Mishra SI et al (2003): Content validity of manual spinal palpatory exams–A systematic review. *BMC Complementary and Alternative Medicine* 3: 1.

Nelson MA, Allen P, Clamp SE and de Dombal FT (1979): Reliability and reproducibility of clinical findings in low-back pain. *Spine* 4: 97–101.

Nice DA, Riddle DL, Lamb RL, Mayhew TP and Rucker K (1992): Intertester reliability of judgements of the presence of trigger points in patients with low back pain. *Archives of Physical Medicine and Rehabilitation* 73: 893–898.

Njoo KH and van der Does E (1994): The occurrence and inter-rater reliability of myofascial trigger points in the quadratus lumborum and gluteus medius: a prospective study in non-specific low back pain patients and controls in general practice. *Pain* 58: 317–323.

Nunnally JC (1978): Psychometric Theory. New York: McGraw-Hill.

Panzer DM (1992): The reliability of lumbar motion palpation. *Journal of Manipulative and Physiological Therapeutics* 15: 518–524.

Petty NJ and Moore AP (2001): Neuromusculoskeletal Examination and Assessment (2nd Edn). Edinburgh: Churchill Livingstone.

Petersen T, Olsen S, Laslett M et al (2004): Inter-tester reliability of a new diagnostic classification system for patients with non-specific low back pain. *Australian Journal of Physiotherapy* 50: 85–91.

Rankin G and Stokes M (1998): Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. *Clinical Rehabilitation* 12: 187–199.

Razmjou H, Kramer JF and Yamada R (2000): Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low-back pain. *Journal of Orthopaedic Sports Physical Therapy* 30: 368–389.

Rhudy TR, Sandefur MR and Burk JM (1988): Interexaminer/ intertechnique reliability in spinal subluxation assessment: a multifactorial approach. *American Journal of Chiropractic Medicine* 1: 111–114.

Riddle DL and Rothstein JM (1993): Intertester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine* 18: 1333–1344.

Russell R (1983): Diagnostic palpation of the spine: A review of procedures and assessment of their reliability. *Journal of Manipulative and Physiological Therapeutics* 6: 181–183.

Sahrmann SA (2001): Diagnosis and Treatment of Movement Impairment Syndromes. Saint Louis: Mosby.

Sammut E and Searle-Barnes P (1998): Osteopathic Diagnosis, Cheltenham: Thornes.

Schwarzer A, Aprill C, Derby R, Fortin JD, Kue G and Bogduk N (1994): The relative contribution of the disc and the zygapophyseal joints in chronic low back pain. *Spine* 19: 801–806.

Sebastian D and Chovvath R (2004): Reliability of palpation assessment in non-neutral dysfunctions of the lumbar spine. *Orthopaedic Physical Therapy Practice* 16: 23–26.

Seffinger MA, Najm WI, Mishra SI et al (2004): Reliability of spinal palpation for diagnosis of back and neck pain. A systematic review of the literature. *Spine* 29: E413–E425.

Seymour R, Walsh T, Blankenberg C, Pickens A and Rush H (2002): Reliability of detecting a relevant lateral shift in patients with lumbar derangement: a pilot study. *Journal of Manual and Manipulative Therapy* 10: 129–135.

Sim J and Wright CC (2005): The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85: 257–268.

Simmonds MJ and Kumar S (1993): Health care ergonomics. Part I: The fundamental skill of palpation—a review and critique. *International Journal of Industrial Ergonomics* 11: 135–143.

Spratt KF, Lehmann TR, Weinstein JN and Sayre HA (1990): A new approach to the low-back physical examination. Behavioral assessment of mechanical signs. *Spine* 15: 96–102.

Streiner DL and Norman GR (1996): PDQ Epidemiology (2nd Edn). St Louis: Mosby.

Streiner DL and Norman GR (2003): Health Measurement Scales (3rd Edn). Oxford: Oxford University Press.

Strender LE, Sjoblom A, Sundell K, Ludwig R and Taube A (1997): Interexaminer reliability in physical examination of patients with low back pain. *Spine* 22: 814–820.

Van der Wurff P, Hagmeijer RHM and Meyne W (2000): Clinical tests of the sacroiliac joint. A systematic methodological review. Part 1: Reliability. *Manual Therapy* 5: 30–36.

Van Dillen LR, Sahrmann SA, Norton BJ et al (1998): Reliability of physical examination items used for classification of patients with low back pain. *Physical Therapy* 78: 979–988.

Van Genderen FR, de Bie RA, Helders PJM and van Meeteren NLU (2003): Reliability research: Towards a more clinically relevant approach. *Physical Therapy Reviews* 8: 169–176.

Van Trijffel E, Anderegg Q, Bossuyt PMM and Lucas C (2005). Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Manual Therapy* 10: 256–269.

Van Tulder M, Furlan A, Bombardier C et al (2003): Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. The Cochrane Library, Issue 4. Oxford: Update Software.

Waddell G, Main CJ, Morris EW et al (1982): Normality and reliability in the clinical assessment of backache. *BMJ* 284: 1519–1523.

Ware JE, Brook RH, Davies AR and Lohr KN (1981): Choosing measures of health status for individuals in general populations. *American Journal of Public Health* 71: 620–625.

Weiner EA and Stewart BJ (1984): Assessing Individuals. Boston: Little Brown.

White LJ and Thomas JS (2002): The rater reliability of assessments of symptom provocation in patients with low back pain. *Journal of Back and Musculoskeletal Rehabilitation* 16: 83–90.

Wilson L, Hall H, McIntosh G and Melles T (1999): Intertester reliability of a low back pain classification system. *Spine* 24: 248–254.

Young S, Aprill C and Laslett M (2003): Correlation of clinical examination characteristics with three sources of low back pain. *Spine Journal* 19: 460–465.

