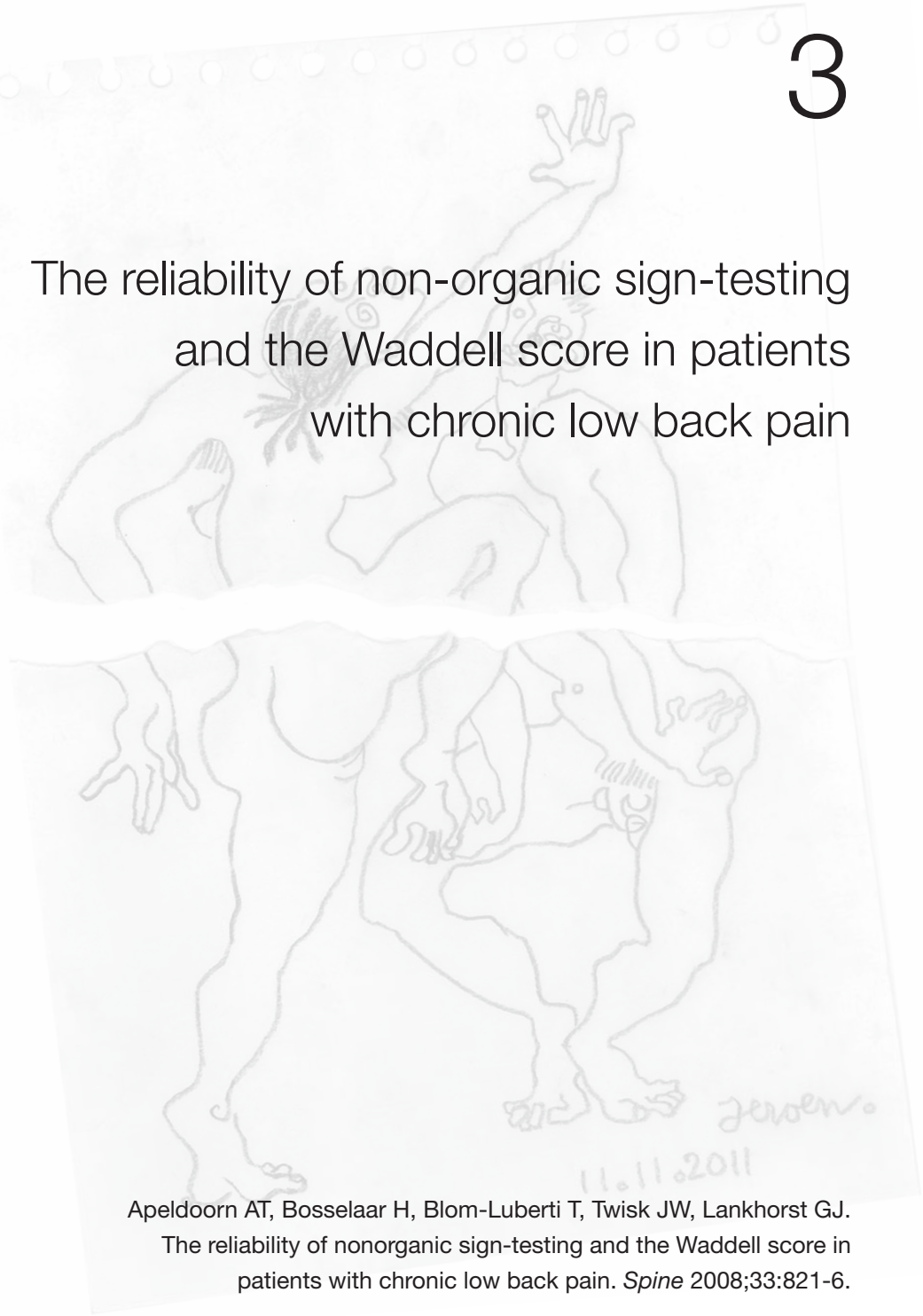
3

The reliability of non-organic sign-testing and the Waddell score in patients with chronic low back pain



Apeldoorn AT, Bosselaar H, Blom-Luberti T, Twisk JW, Lankhorst GJ. The reliability of nonorganic sign-testing and the Waddell score in patients with chronic low back pain. *Spine* 2008;33:821-6.

Abstract

Study Design: An observational prospective cohort study.
Objectives: To determine the reliability of non-organic sign-testing in patients with chronic low back pain (CLBP), and to identify determinants of diagnostic disagreement.
Background: For the assessment of behavioral responses to examination, Waddell et al. published 'the Waddell score' in 1980. The Waddell score consists of eight non-organic signs, divided into five categories. The overall score is positive if at least three of the categories are scored positive. Although the Waddell score is widely used, little is known about its reliability.
Methods: Two observers examined 126 consecutive patients with CLBP referred for rehabilitation. Cohen's kappa was used to compute the inter-rater and intra-rater reliability of the sign maneuvers, categories and Waddell score. Cronbach's alpha was calculated for the five categories and eight signs in order to determine internal consistency. Chi-square tests were applied to determine the possible influence of clinical characteristics on inter-rater reliability.
Results: Inter-rater reliability varied from 0.33 to 0.74 for the sign maneuvers and categories, and was 0.48 and 0.49 for the overall Waddell score. Intra-rater reliability varied from 0.43 to 0.84 for the sign maneuvers and categories, and was 0.65 and 0.68 for the overall Waddell score. Internal consistency varied from 0.65 to 0.72 for the categories and from 0.71 to 0.78 for the signs. Determinants of diagnostic disagreement did not exceed levels of significance ($p < 0.05$).
Conclusions: For trained observers of a population of patients with CLBP in a rehabilitation setting, the inter-rater reliability of the Waddell score was moderate and the intra-rater reliability was good. No influence of clinical characteristics was found on inter-rater reliability. To optimize the homogeneity and variability of the Waddell score, we recommend summing up the individual signs instead of summing up the categories.

Introduction

Illness behavior has been suggested to comprise an important construct relevant in both the development and the maintenance of chronic pain.[1,2] Waddell et al. defined illness behavior as 'observable and potentially measurable actions and conduct which express and communicate the individual's own perception of disturbed health'.[3] Illness behavior is a normal part of human illness. In some patients, however, illness behavior becomes out of proportion to their physical problem, and therefore becomes counterproductive.[2] In 1980, Waddell et al. developed a screening tool to distinguish this medically incongruent behavior from clinical signs and symptoms that accurately portrayed the physical disease in patients with low back pain (LBP).[4] The assessment became known as 'the Waddell score'. According to Waddell, a positive score is an indication of emotional or psychological distress, which should alert the clinician to the need for more detailed psychological assessment.[5,6]

The Waddell score consists of eight standardized physical non-organic maneuvers, divided into five categories (Table 1). A category is scored positive if at least one test out of that category is scored positive. The (overall) Waddell score is positive if three or more categories are positive. Waddell stresses that non-organic contributions to a patient's LBP may coexist with organic pathology, so isolated signs should not be over-interpreted.[6]

Although there is increasing evidence that non-organic signs, also known as behavioral signs, are associated with poorer physical performance, more severe pain, psychological features and other measures of illness behavior, these signs have always been criticized. The most comprehensive criticism was published recently by Fishbain et al.[7] In their structured evidence-based review they concluded, among other things, that non-organic signs do not discriminate organic from non-organic problems, and that non-organic signs may represent an organic phenomenon. The available space does not permit an extended review of their publication, but we agree with Waddell, who criticized the Fishbain et al. study in his book 'The Back Pain Revolution', that their study is packed with methodological flaws, biased interpretations of the literature and unfounded conclusions, and does not do justice to the (initial) scientific basis of non-organic signs.[2]

Table 1 The five categories and eight non-organic tests of the Waddell score

Categories	Non-organic tests	Non-organic signs
I. Tenderness	*Superficial*	Widespread tenderness to light pinches over a wide area of lumbar skin.
	Non-anatomic	Deep tenderness felt over a wide area, not localized to one structure and often extended to the thoracic spine, sacrum or pelvis.
II. Simulation	*Axial loading*	Low back pain reported when light pressure is given on the patient's head while standing.
	Rotation	Low back pain reported when the shoulders and pelvis are passively rotated in the same plane as the patient stands with the feet together.
III. Distraction	*Distraction*	Inconsistent limitation of straight leg raising in supine and seated positions.
IV. Regional disturbance	*Weakness*	Partial cogwheel 'giving way' in many muscle groups.
	Sensory	Sensory disturbances include diminished sensation to light touch, pinprick, and sometimes other modalities fitting a 'stocking' rather than a dermatomal pattern.
V. Over-reaction	*Over-reaction*	Disproportionate verbalization, facial expression, muscle tension and tremor, collapsing or sweating during examination.

The (overall) Waddell score is positive if three or more categories are positive. A category is positive if at least one non-organic test in that category is positive.

Nevertheless, we agree with Fishbain et al. that literature on the reliability of non-organic signs is scarce and results are conflicting.[7] Waddell et al. investigated the intra-rater and inter-rater reliability of non-organic signs and the Waddell score in a population of 50 patients with chronic low back pain (CLBP).[4] Intra-rater and inter-rater agreement ranged from 76%-90% for signs, and was 86% for the Waddell score. The inter-rater reliability has been investigated in only a few other studies.[8-12] However, for the most part, the data are difficult to interpret because of the small size of the patient groups ($n < 27$)[8-10] and the deviating scoring methods.[11] Only McCombe et al. have investigated the inter-rater reliability with a sufficient number of patients.[12] In their study, a group of 50 patients was examined by two orthopedic surgeons, and another group of 33 patients was examined by a physiotherapist and an orthopedic surgeon. They found low, and even negative, kappa values (-0.16 to 0.48) for the inter-rater reliability of the non-organic categories. However, the prevalence of non-organic categories found by the examiners was low. Furthermore, they only investigated the inter-rater reliability of the categories, but did not present data about the inter-rater reliability of individual non-organic signs or about the (overall) Waddell score. To the authors' best knowledge, apart from the Waddell et al. study[4], there are no other studies that have reported data on the intra-rater reliability of non-organic signs.

Finally, measures of internal consistency (homogeneity) have been reported. Waddell et al. and Lehmann et al. analyzed the data for a total of 680 patients, and concluded that homogeneity of the eight non-organic signs and five categories was sufficient.[4,5,13,14] It must be noted that most of their data were analyzed with correlation and factor analysis, which is less suitable for use with dichotomous items.[15]

In summary, existing data on the intra-rater and inter-rater reliability of non-organic sign-testing are insufficient, and have yielded contradictory results. Non-organic signs probably form a homogeneous group.

The main objective of this study was to assess the homogeneity, intra-rater and inter-rater reliability of non-organic sign maneuvers, categories and the Waddell score in a sufficiently large population. To guarantee maximal prevalence of positive non-organic signs, patients with a long history of CLBP were selected. The secondary objective was to identify determinants of diagnostic disagreement between examiners.

Materials and Methods

Subjects

From January 2002 until May 2004, all patients with CLBP who were recruited by rehabilitation physicians for an observation period in the outpatient rehabilitation center of the Medical Centre Alkmaar (MCA) were invited by the first author (AA) to participate in the study. The patients had been referred to the rehabilitation clinic by general practitioners and medical specialists for the management of CLBP and disability, because they had not responded to conventional and/or surgical treatment. Inclusion criteria were the following: CLBP (low back pain for more than three months) as the primary problem with or without radiation, age over 18, and adequate command of the Dutch language. Patients who had previously been treated for LBP in a multidisciplinary rehabilitation setting were excluded, but patients who had received mono-disciplinary treatment, for example from a physical therapist or a psychologist, were included. The Regional Medical Ethics Committee approved the study design and all patients gave written informed consent.

Procedure

Patient observation period

Each patient underwent an observation period of three weeks, during which the patient was screened by a physiotherapist, an occupational therapist, a social worker and a psychologist. Furthermore, each patient completed a battery of questionnaires. After observation, the findings were discussed in a multidisciplinary team meeting to formulate a treatment strategy.

Training of the observers

Prior to the study, the two examiners (AA and HB), both physiotherapists with more than 10 years of experience in rehabilitation, scored for 12 patients the eight signs according to Waddell et al,[6] and gave feedback to each other to establish accuracy. Vague test descriptions were standardized by consensus. For example, in the axial loading test, the pressure on the patient's head was established between two and a half and five kilograms.

Inter-rater and intra-rater reliability

At the moment when the patient started the observation period, the two physiotherapists administered the non-organic tests directly after each other, without being informed

about the patient or about each other's findings. The physiotherapist who was not involved in the observation of the patient made the first examination. To determine the inter-rater and intra-rater reliability, both physiotherapists performed sign-retesting, according to the same procedure, preferably at the end of the observation period, with a minimum of a few days and a maximum of three weeks. To prevent a patient's awareness of the evaluation of non-organic signs, the tests were preceded by three active tests measuring lumbar mobility. Scores were given for signs, categories and the Waddell score. After the examination, the results were given to an independent research assistant. For the sake of clarity, all retesting took place before the start of any intervention. During the study, the test results were not open to inspection, and the examiners were blinded for each other's findings.

Questionnaires

Demographic characteristics (age, gender, nationality) were recorded by means of a questionnaire.

All patients recorded the severity of their current pain on a visual analogue scale (VAS), ranging from 0 ('no pain') to 100 ('unbearable pain').[16] The VAS has been proven to be user-friendly, valid and reliable.[17-19]

Fear of movement was measured with the Dutch version of the Tampa Scale of Kinesiophobia (TSK), which consists of 17 questions (Miller RP, et al. Unpublished report, 1991).[20-21] Each item is scored on a 4-point Likert scale, with answer categories ranging from 'strongly disagree' to 'strongly agree'. Items 4, 8, 12 and 16 are reverse-scored. The total score ranges from 17 (no fear-avoidance beliefs) to 68 (strong fear-avoidance beliefs). The reliability and validity of the Dutch version proved to be adequate.[21-23]

Disability was measured with the Dutch version of the Roland Disability Questionnaire (RDQ), which consists of 24 yes/no questions pertaining to difficulties in performing various daily activities in the past 24 hours.[24] The total score ranges from 0 'no disability' to 24 'difficulty with all applicable items'. In this study, the sum-scores of the RDQ ranged from 0-100 by multiplying each score by 100/24. The Dutch translation has been proved to be a valid instrument.[25]

Analysis

The prevalence of positive signs, categories and Waddell scores was calculated. Percentages of agreement and Cohen's kappa, including its 95% confidence interval (CI), were calculated in order to quantify the inter-rater and intra-rater reliability of

sign-testing, categories and the Waddell score. For Cohen's kappa the following classification was used: < 0.20 poor, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 good, 0.81-1.00 very good agreement.[26] To obtain insight into the internal consistency of the Waddell score, Cronbach's alpha was calculated for both the original Waddell score (five categories) and a Waddell score consisting of signs only. The desired value of Cronbach's alpha depends on the purpose of the test.[26] The following classification for tests of moderate importance was used: < 0.70 insufficient, 0.70-0.80 sufficient, > 0.80 good internal consistency.[27]
Univariate analysis was applied to determine whether demographic characteristics (gender, age and nationality), pain, disability and fear of movement were determinants of diagnostic disagreements between the two examiners on the (overall) Waddell score. The statistical analysis was performed in SPSS (version 14.0) and VasserStats.

Results

128 Patients were eligible for the study, two of whom were unwilling to participate. The total number of patients participating in the study was thus 126. Table 2 presents the main characteristics of the patients, and Table 3 shows the prevalence of positive signs, categories and the Waddell score. The prevalence of positive signs and categories varied from 14% to 49%, and was 36% for the Waddell score.

Inter-rater reliability

Almost all participating patients were tested at the start and at the end of the observation period by examiner A and B, so inter-rater reliability could be calculated twice. Examiner B re-examined 110 patients who were first tested by examiner A after a mean period of 13.4 days. Examiner A performed the retesting procedure after examiner B for 108 patients with a mean time-interval of 13.7 days. Table 3 presents the percentages of agreement and the kappa values of the inter-rater reliability for sign-testing, categories and the Waddell score for both test procedures. For signs and categories, agreement expressed in kappa values varied from 0.33 (superficial) to 0.74 (rotation). The kappa values for the Waddell score were 0.48 and 0.49.

Table 2 Main characteristics of participating patients

Female (%)	57.9
Dutch nationality (%)	92.1
No low back operations (%)	81.6
Mean age in years (SD)	44.4 (11.4)
Mean current pain[1](SD)	55.6 (23.9)
Mean disability[2](SD)	51.3 (17.7)
Mean fear of movement[3](SD)	37.9 (7.2)
Median duration of low back pain in months (IQR)	84.0 (23-162)

SD, standard deviation; IQR, inter-quartile range
[1] visual analogue scale (0-100), [2] Roland Disability Questionnaire (0-100),
[3] Tampa Scale of Kinesiophobia (17-68).

Table 3 **The prevalence of positive signs, categories and Waddell score and its inter-rater reliability**

Type of test	Prevalence (%) (n=126)	Examiner A-B (n=110)[1]		Examiner B-A (n=108)[2]	
		Agreement (%)	Kappa (95%CI)	Agreement (%)	Kappa (95%CI)
I. Tenderness	42	75	0.49 (0.32, 0.66)	76	0.49 (0.32, 0.66)
Superficial	19	78	0.33 (0.11, 0.55)	85	0.49 (0.28, 0.71)
Non-anatomic	36	73	0.40 (0.22, 0.59)	74	0.40 (0.21, 0.58)
II. Simulation	49	85	0.69 (0.56, 0.83)	85	0.70 (0.57, 0.84)
Axial loading	23	83	0.55 (0.37, 0.73)	82	0.53 (0.34, 0.71)
Rotation	41	87	0.74 (0.61, 0.87)	82	0.65 (0.50, 0.79)
III. Distraction	14	87	0.46 (0.22, 0.70)	83	0.37 (0.14, 0.61)
IV. Regional disturbance	27	83	0.55 (0.37, 0.73)	81	0.51 (0.32, 0.69)
Weakness	18	86	0.53 (0.33, 0.74)	84	0.49 (0.28, 0.70)
Sensory	14	91	0.61 (0.40, 0.83)	94	0.68 (0.46, 0.90)
V. Over-reaction	49	70	0.42 (0.24, 0.59)	68	0.37 (0.19, 0.55)
Waddell score	36	76	0.48 (0.30, 0.65)	78	0.49 (0.31, 0.67)

[1]Examiner A examined the patient first and examiner B performed retesting of the signs after a time-interval of 7-21 days (mean; 13.4 days). [2]Examiner B examined the patient first and examiner A re-examined the patient after a time-interval of 4-21 days (mean; 13.7 days).
CI, confidence interval

Intra-rater reliability

Of 126 patients eligible for the study examiner A performed the non-organic tests twice on 110 patients, and examiner B on 114 patients, with a mean interval of respectively 13.4 and 13.8 days. Table 4 presents the results for both examiners. The lowest kappa value was found for the sign superficial tenderness (0.43), and the highest values were found for the signs regional disturbance and sensory (0.84). The kappa values for the Waddell score were 0.65 and 0.68. In the inter-rater reliability and the intra-rater reliability were no statistical differences between patients who were retested and not retested, with regard to age, gender, nationality, low back operations, current pain, disability, depression, kinesiophobia and duration of low back pain (exact values available from the first author).

Table 4 **Intra-rater reliability of the signs, categories and Waddell score for examiners A and B**

Type of test	Examiner A (n=110)[1]		Examiner B (n=114)[2]	
	Agreement (%)	Kappa (95%CI)	Agreement (%)	Kappa (95%CI)
I. Tenderness	77	0.54 (0.38, 0.69)	77	0.50 (0.33, 0.66)
Superficial	82	0.47 (0.27, 0.67)	85	0.43 (0.20, 0.65)
Non-anatomic	76	0.50 (0.34, 0.67)	82	0.55 (0.38, 0.72)
II. Simulation	81	0.62 (0.47, 0.76)	86	0.72 (0.59, 0.85)
Axial loading	85	0.63 (0.46, 0.79)	82	0.50 (0.31, 0.68)
Rotation	79	0.58 (0.43, 0.73)	85	0.69 (0.56, 0.83)
III. Distraction	87	0.55 (0.35, 0.76)	89	0.49 (0.25, 0.73)
IV. Regional disturbance	93	0.80 (0.67, 0.93)	94	0.84 (0.72, 0.95)
Weakness	91	0.69 (0.52, 0.87)	92	0.72 (0.55, 0.89)
Sensory	96	0.84 (0.68, 0.99)	96	0.84 (0.68, 0.99)
V. Over-reaction	84	0.66 (0.52, 0.81)	82	0.61 (0.46, 0.76)
Waddell score	84	0.65 (0.50, 0.80)	87	0.68 (0.53, 0.84)

[1]Time-interval; 4-21 days (mean; 13.4 days). [2]Time-interval; 7-21 days (mean; 13.8 days).
CI, confidence interval

Internal consistency

For the original Waddell score with five categories the internal consistency expressed by Cronbach's alpha for examiner A and B respectively was 0.72 and 0.65, and for a Waddell score consisting of signs only it was respectively 0.78 and 0.71. If one sign or category was omitted, Cronbach's alpha decreased in all cases, which was indicative of a positive contribution of all signs and categories to the (overall) Waddell score.

Determinants of diagnostic disagreement

Univariate analysis showed that gender, age, nationality, pain, disability and fear of movement were not significantly associated with inter-rater disagreement on the Waddell score (Table 5).

Table 5 Determinants of inter-observer disagreement on the Waddell score

	Examiner A-B (n=110)[1]				Examiner B-A (n=108)[2]			
	Agreement	Disagreement			Agreement	Disagreement		
Variables	**Mean**	**Mean**	**p**	**OR**	**Mean**	**Mean**	**p**	**OR**
Gender			0.46	1.42			0.19	0.55
Nationality (Dutch-foreign)			0.93	1.08			0.40	0.41
Age (years)	45.4	43.5	0.45		44.1	47.4	0.19	
Disability (RDQ)	50.9	51.7	0.88		49.5	57.5	0.13	
Current pain (VAS)	55.3	57.1	0.72		55.9	57.4	0.74	
Fear of movement (TSK)	37.5	39.6	0.21		38.4	36.6	0.24	

Dichotomous variables were analysed with the χ^2 test and the dependent variable was defined to be 1 if the examiners disagreed on the Waddell score and 0 if they agreed. Continuous variables were analysed with the t-test.

[1]Examiner A made the first examination and examiner B the second with a time-interval of 7-21 days (mean; 13.4 days). [2]Examiner B made the first examination and examiner A the second with a time-interval of 4-21 days (mean; 13.7 days).

RDQ, Roland Disability Questionnaire (0-100); VAS, visual analogue scale (0-100); TSK, Tampa Scale of Kinesiophobia (17-68); OR, odds ratio.

Discussion

The present study investigated the inter-rater and intra-rater reliability and internal consistency of non-organic sign-testing according to Waddell et al.[4] The inter-rater reliability of the signs varied from fair to good, and the intra-rater reliability varied from moderate to very good. The results indicated that the inter-rater reliability of the overall Waddell score is moderate and the intra-rater reliability is good. Furthermore, the internal consistency of a Waddell score consisting of signs only was sufficient and had higher Cronbach's alpha values, compared to a Waddell score consisting of the original five categories. To our knowledge, the present study was the first to analyze possible influences of clinical characteristics on the inter-rater reliability of the Waddell score. However, no influences were found.

The agreement percentages for inter-rater reliability found in this study are in concordance with the findings of Waddell and his colleagues in 1980.[4] In their study, two examiners independently re-examined 50 patients with a time-interval of 1-6 days. The kappa values of inter-rater reliability varied from 0.55 and 0.71 for the sign maneuvers and were reported in 1982.[28] These somewhat higher kappa values, compared to our findings, are probably a result of prevalence-dependence of the kappa values[29] i.e. low kappa values will be found when the prevalence of a finding is very high or very low. There was a 50% prevalence of a positive score in the Waddell et al. study, and therefore the kappa values were probably higher than in the present study, in which the prevalence was 36%. This prevalence-dependence of kappa values can also partially explain the unsatisfactory kappa values for inter-rater reliability found by McCombe et al,[12] who reported a low prevalence of positive categories and excessively large 95% CIs for kappa values.

In the present study the intra-rater reliability was found to be superior to the inter-rater reliability. This is according to our expectations, because in contrast to intra-rater reliability, inter-rater reliability contains the error of variation between observers. Our results of intra-rater reliability expressed in agreement percentages are better than the original findings of Waddell et al.[4] The lower agreement percentages found in their study could have been caused by the longer time-intervals they used. Waddell et al. carried out testing sessions at an average interval of 23 days, whereas in the present study the average intervals were 13-14 days, with a maximum of 21 days.

In 1980, Waddell et al. grouped the eight signs into five categories, but did not offer data to support this classification.[4] In general, this reduction in data is not advisable, because information about the patient will be lost before the analysis is performed.

Moreover, as long as the test items are not perfectly correlated, the internal consistency of a test will improve with an increase in the number of items included in a test.[30] From this statistical point of view it was not surprising that in the present study the internal consistency expressed by Cronbach's alpha was higher for a Waddell score consisting of signs only (0.71 and 0.78) than for the original Waddell score consisting of five categories (0.65 and 0.72). As far as we are aware, our results can only be compared with the results of the Lehmann et al. study (n = 56), which reported a Cronbach's alpha of 0.77 for a Waddell score consisting of four non-organic signs.[13] In line with our findings, and on the basis of theoretical considerations, we and some other authors recommend that the signs should be summed up instead of the categories.[13,14,31-34]

Recently, Waddell dropped the sign (and category) over-reaction from the Waddell score, because it was found to be unreliable and prone to observer bias.[2] In the present study, the inter-rater agreement percentages for over-reaction are low compared to the other signs, but the corresponding kappa values and the intra-rater reliability of this non-organic sign are comparable with those of other non-organic signs. Therefore, according to our results, there is no indication that over-reaction should be omitted from the Waddell score. Moreover, in our study, internal consistency would decrease if over-reaction was omitted (data not shown).

Due to occasional absence of the examiners, not all patients were retested, and this could be a threat to internal validity. However, the periods of absence for both examiners were randomly divided, and probably did not result in selection bias. Moreover, for both examiners, there were no statistical differences with regard to demographical and biographical data between patients who were retested and not retested. Another point of discussion is the appropriate time-interval between examinations. Short and long intervals both have their disadvantages, and the appropriate interval varies, depending on the task. The mean time-interval between examinations to determine the inter-rater and intra-rater reliability was 13-14 days. In our opinion, this is long enough for patients and observers to forget the first responses, and sufficiently short to assume that the underlying process has not changed.

An issue that should be addressed concerning the Waddell score is the inconsistency of terminology in the literature. It is not always clear whether the word 'signs' means non-organic signs or categories. Occasionally, even Waddell himself uses signs instead of categories ('…most patients had either 0-1 behavioral signs or showed a constellation of three or more').[2] This confusing nomenclature is not conducive to the establishment of a body of knowledge about the Waddell score.

In summary, it may be concluded that for trained observers of a population of patients

with CLBP, the inter-rater reliability of non-organic sign-testing according to Waddell et al.[4] was moderate and the intra-rater reliability was good. Clinical characteristics were not associated with disagreement. Moreover, the present study provided no support for the categories recommended by Waddell et al. or for omitting one particular non-organic sign. Therefore, we recommend summing up the non-organic signs instead of summing up the categories. However, when generalizing the results to other settings, they should be interpreted with caution.

References

1. Turk DC, Flor H. Pain>pain behaviors: the utility and limitations of the pain behaviour construct. *Pain* 1987;31:277-95.
2. Waddell G. *The Back Pain Revolution.* 2nd ed. Edinburgh: Churchill Livingstone, 2004.
3. Waddell G, Pilowsky I, Bond MR. Clinical assessment and interpretation of abnormal illness behaviour in low back pain. *Pain* 1989;39:41-53.
4. Waddell G, McCulloch JA, Kummel E, et al. Nonorganic physical signs in low-back pain. *Spine* 1980;5:117-25.
5. Waddell G, Main CJ, Morris EW, et al. Chronic low-back pain, psychologic distress, and illness behavior. *Spine* 1984;9:209-13.
6. Waddell G. Nonorganic signs [editorial]. *Spine* 2004;13:1393.
7. Fishbain DA, Cole B, Cutler RB, et al. A structured evidence-based review on the meaning of nonorganic physical signs: Waddell signs. *Pain Med* 2003;4:141-81.
8. Hadjistavropoulos HD, Craig KD. Acute and chronic low back pain: cognitive, affective and behavioral dimensions. *J Consult Clin Psychol* 1994;62:341-9.
9. Reesor KA, Craig KD. Medically incongruent chronic back pain: physical limitations, suffering and ineffective coping. *Pain* 1988;32:35-45.
10. Sobel JB, Sollenberger P, Robinson R, et al. Cervical nonorganic signs: a new clinical tool to assess abnormal illness behaviour in neck pain patients: a pilot study. *Arch Phys Med Rehabil* 2000;81: 170-5.
11. Korbon GA, DeGood DE, Schroeder ME, et al. The development of a somatic amplification rating scale for low-back pain. *Spine* 1987;12:787-91.
12. McCombe PF, Fairbank JCT, Cockersole BC, et al. Reproducibility of physical signs in low-back pain. *Spine* 1989;14:908-18.
13. Lehmann TR, Russell DW, Spratt KF. The impact of patients with nonorganic physical findings on a controlled trial of transcutaneous electrical nerve stimulation and electroacupuncture. *Spine* 1983;8:625-34.
14. Waddell G, Richardson J. Observation of overt pain behaviour by physicians during routine clinical examination of patients with low back pain. *J Psychosom Res* 1992;36:77-87.
15. Comrey AL. Common methodological problems in factor analysis. *J Consult Clin Psychol* 1978;46:648-59.
16. Jensen MP, Karoly P. Self-report scales and procedures for assessing pain in adults. In: Turk DC, Melzack R, eds. *Handbook of Pain Assessment.* New York: The Guilford Press, 1992:135-51.
17. Carlsson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain* 1983;16:87-101.
18. Revill SI, Robinson JO, Rosen M, et al. The reliability of a linear analogue for evaluating pain.

Anaesthesia 1976;31:1191-8.

19. Sriwatanakul K, Kelvie W, Lasagna L, et al. Studies with different types of visual analog scales for measurement of pain. *Clin Pharmacol Ther* 1983;34:234-9.
20. Kori SH, Miller RP, Todd DD. Kinisophobia: a new view of chronic pain behaviour. *Pain Management* 1990;3:35-43.
21. Vlaeyen JWS, Kole-Snijders AMJ, Boeren RGB, et al. Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. *Pain* 1995;62:363-72.
22. Goubert L, Crombez G, Van Damme S, et al. Confirmatory factor analysis of the Tampa scale for kinesiophobia. Invariant two-factor model across low back pain patients and fibromyalgia patients. *Clin J Pain* 2004;20:103-10.
23. Vlaeyen JWS, Kole-Snijders AMJ, Rotteveel AM, et al. The role of fear of movement/(re)injury in pain disability. *J Occup Rehabil* 1995;5:235-52
24. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low back pain. *Spine* 1983;8:141-4.
25. Beurskens AJHM, De Vet HCW, Köke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain* 1996;65:71-6.
26. Altman DG. *Practical statistics for medical research.* London: Chapman & Hall, 1997.
27. Evers AVAM, Van Vliet-Mulder JC, Groot CJ. *Documentatie van tests en testresearch in Nederland (Documentation of tests and testresearch in the Netherlands).* 7th ed. Assen: Van Gorcum, 2000.
28. Waddell G, Main CJ, Morris EW, et al. Normality and reliability in the clinical assessment of backache. *Br Med J* 1982;284:1519-23.
29. Bouter LM, Van Dongen MCJM. *Epidemiologisch onderzoek, opzet en interpretatie (Epidemiologic research: principles and methods)* 3rd ed. Houten/Diegem: Bohn Stafleu Van Loghum, 1995.
30. Streiner DL, Norman GR. Health Measurement Scales. A practical guide to their development and use. 3rd ed. New York: Oxford University Press Inc, 2003.
31. Prkachin KM, Schultz I, Berkowitz J, et al. Assessing pain behaviour of low-back pain patients in real time system; concurrent validity and examiner sensitivity. *Beh Res Ther* 2002;40:595-607.
32. Gaines WG, Hegmann KT. Effectiveness of Waddell's nonorganic signs in predicting a delayed return to regular work in patients experiencing acute occupational low back pain. *Spine* 1999;24:396-401.
33. Gunzburg R, Keller TS, Szpalski M, et al. Clinical and psychofunctional measures of conservative decompression surgery for lumbar spinal stenosis: a prospective cohort study. *Eur Spine* 2003;12:197-204.
34. Werneke MW, Harris DE, Lichter RL. Clinical effectiveness of behavioral signs for screening chronic low-back pain patients in a work-oriented physical rehabilitation program. *Spine* 1993;18:2412-8.