Title:

The intra- and inter-observer reliability of a novel protocol for two-point discrimination in individuals with chronic low back pain.

Authors:

Ehrenbrusthoff K[1,2], Ryan CG[1], Grüneberg C[2], Wolf U[2,3], Krenz D[2], Atkinson G[1], Martin DJ[1].

Authors' Affiliations:

[1]Health and Social Care Institute, Teesside University, Middlesbrough, UK, TS1 3BA

[2]Hochschule fuer Gesundheit, Department of Applied Health Sciences, Gesundheitscampus 6-8, 44801 Bochum, Germany

[3]Hochschule Fulda, Department of Nursing and Health Sciences, Leipziger Strasse 123, 36037 Fulda, Germany

Corresponding Author:

Katja Ehrenbrusthoff, Department of Applied Health Sciences, Hochschule fuer Gesundheit Bochum, Gesundheitscampus 6-8, 44801 Bochum, Germany, E-Mail: katja.ehrenbrusthoff@hs-gesundheit.de

Short Title:

Low back 2-point discrimination reliability

ABSTRACT

Two-point discrimination is measured as an indicator of cortical reorganisation in musculoskeletal medicine. Nevertheless, data are lacking for the reliability of this measure in patients with non-specific chronic low back pain (NSCLBP). We aimed to quantify the intra- and inter-observer reliability of a novel protocol for measuring two-point discrimination in these patients. 35 participants (12 males, 23 females, mean age 52, SD 15years) with NSCLBP were recruited. Three clinicians made 14 consecutive measurements of two-point discrimination with callipers. One of these clinicians repeated the assessment protocol within 7 days. During each measurement, the calliper width was widened in 5-mm increments until participants could consistently identify two points. Intra- and inter-observer agreement was quantified using mean difference, within-subject SD and limits of agreement (LOA). After using the first measurement for familiarisation, the mean of measurements 2 to 5 within an assessment resulted in the optimum compromise between clinic time constraints and acceptable intra-observer reliability; the within-subjects SD being 7.5 mm (LOA: 20.8 mm). Inter-observer reliability was generally poorer; requiring the mean of measurements 2 to 9 within an assessment for a similar within-subjects SD of 8.6 mm (LOA: 23.7 mm). It was estimated that these within-subjects SDs were small enough for a clinically-important change to be detected with a feasible sample size in future studies. The intra-observer reliability of our assessment protocol is acceptable for detecting a clinically relevant difference in two-point discrimination for future research purposes. Nevertheless, individual patient measurement variability is relatively high, especially between different clinicians.

Key words: Tactile acuity, reliability, low back pain, measurement

INTRODUCTION

It has been reported that the somatosensory cortex is disrupted in patients with chronic pain - a phenomenon termed cortical reorganisation [1]. In patients with phantom limb pain and complex regional pain syndrome, the degree of cortical reorganisation has been shown to directly relate to their pain experience [2] and, as the pain intensity improves, the somatosensory representation normalises [3]. While the mechanisms have not been fully elucidated, normalisation of the somatosensory cortex is considered a viable target for the treatment for pain patients [4]. Interventions attempting to normalise cortical reorganisation, such as sensory discrimination training [5], have provided preliminary proof-of-concept for such interventions.

Cortical reorganisation has been shown in patients with non-specific chronic low back pain (NSCLBP) [6]. Furthermore, preliminary studies targeted at cortical reorganisation in patients with NSCLBP have shown promising results [5, 7]. As interventions that target cortical reorganisation for patients with NSCLBP become more common in clinical practice and research, there is a need to ensure that cortical reorganisation can be measured reliably and efficiently. The gold standard methods of measuring cortical reorganisation are functional Magnetic Resonance Imaging (fMRI) and Electromagnetic Encephalography (EEG) [8]. These methods are very expensive, require sophisticated technology, skilled technicians and can be time consuming. Therefore, there is a need to develop and appraise less expensive assessment methods, which also have acceptable clinical utility.

Two-point discrimination (TPD) is a simple clinical test of tactile acuity, which measures the minimum distance between two points on the skin that can be consciously detected [9]. Smaller distances indicate better acuity. Because TPD is correlated with cortical reorganisation, it is commonly-used as a proxy measure of cortical reorganisation [10]. TPD

was initially developed to assess finger and hand tactile acuity [11]. More recently, studies have used TPD to asses lower back tactile acuity as a proxy measure of lower back somatosensory reorganisation [12, 13]. To date, only one research group has investigated the reliability of lower back TPD and these researchers studied asymptomatic individuals [14]. Given that the test will be used clinically in patients with non-specific CLPB, there is a need to directly assess the reliability of this technique in that population.

It has been highlighted that current TPD techniques involve a considerable amount of subjectivity, where the clinician must make a judgement as to when sufficient consistency of distance has been attained [15]. This can be clinically challenging, time consuming and can introduce bias. There is a need to develop a TPD protocol, which reduces this source of variability by minimising clinician judgment. The overarching aim of this study was to develop and quantify the intra- and inter-observer reliability of a novel lower back TPD assessment protocol, which minimises subjective clinical judgement, in patients with NSCLBP. Specifically, the two objectives were to establish:

1. the minimum number of TPD measurements required within the assessment protocol to maximise intra-observer reliability whilst minimising the time required to complete the test;
2. the minimum number of TPD measurements required within the assessment protocol to maximise inter-observer reliability whilst minimising the time required to complete the test.

METHODS

Study overview

In this reliability study, 35 participants with NSCLBP underwent the same TPD test protocol at three different time points to assess tactile acuity of the lower back. To assess intra-observer reliability, assessor 1 measured TPD on day 1 and day 2. To assess inter-observer reliability, a second assessor measured each participant on day 1 and this was compared to the first assessor's measurement for day 1.

Recruitment

Patients were recruited consecutively from physiotherapy practices in Bochum, Germany between June 2013 and December 2014. Participants had to meet the following inclusion criteria: age ≥ 18 years; NSCLBP with or without leg pain (for those with leg pain, the back pain had to be dominant); duration of symptoms ≥ 6 months; sufficient cognitive and German language skills/ability to understand both oral and written instructions and to give informed consent; intact skin on the lower back. Participants were excluded if they had signs and symptoms indicating serious spinal pathologies (red flags). The study was approved by Teesside University's School of Health and Social Care Research Governance and Ethics Board and the Ethics committee of the German National Physiotherapists Society.

Two-Point Discrimination Assessment Procedure

The test procedure was developed from previous protocols for TPD threshold measurement [16, 17]. Participants were positioned in a comfortable lying prone position on a treatment bench with their back exposed. A pillow was positioned under the stomach to flatten the lumbar spine. Feet were supported by a half roll for participant comfort. Using the standardised palpation procedure according to Merz et al. [18], the tips of the transverse processes of L5 were located and marked with a washable pen. The measurement tool was a

2-point discrimination caliper (Nexgen Medical Systems, Florida, USA) with a 1mm precision. The calliper was applied, a sufficient amount of time to bring about the first blanching of the skin and was then removed, no more than 1-2 seconds. The calliper was placed horizontally to the spine and the transverse process of L5 was the center for the calliper. Testing was carried out in an ascending (or widening) manner, starting with a distance of 20mm between the two calipers and was increased in 5mm increments. This was based upon preliminary rehearsals of process which found 1mm increments too time consuming and distances between 0-20mm being constantly identified as one point. The time between each increment was no more than a few seconds. The assessment was carried out at one location of the lower back, either on the affected side or, with bilateral pain, on a prior randomly identified side. Participants were advised to say 'one', when they felt one point and 'two' when they felt two points. Catch trials were also included approximately every 5 measurements by either using only one point or the widest possible distance. The distance at which the participant first identified two points was noted. The callipers were then increased by 5mm. If the participant again noted two points, this was considered to indicate consistency of identification and the first distance of two-point identification was taken as the TPD result. If, however, the participant did not identify the stimulus as two separate points after the 5mm increment, the assessor continued to expand the distance until two consecutive correct ratings were provided by the participant. This test procedure, operationally defined within this paper as a measurement, was performed 14 times consecutively within each TPD assessment, with the first test considered a practice. The decision to consider the first test as a practice was based upon evidence of systematic effects, in that the first test was consistently poorer (a wider TPD score) than subsequent measurements. After discarding the first test, we did not find any clinically important upward or downward trend amongst the remaining measurements.

Order of testing

One assessor (KE) carried out the TPD protocol with each participant on two separate days. Day 1 and day 2 were never more than one week apart. Results between day 1 and day 2 for assessor 1 were compared to assess intra-observer reliability. The mean of sessions were consecutively calculated (i.e. mean of measurements 2 and 3, mean of measurements 2, 3 and 4 etc.) and compared either across day 1 and 2 for assessor 1. This was undertaken to identify the number of individual TPD measurements needed within each TPD assessment to establish a stable result within a clinically reasonable timeframe for a single assessor.

A second assessor (either DK or UW) carried out the TPD protocol on day 1 with each participant immediately after assessor 1 had completed testing. Results between assessor 1 and 2 were compared to assess inter-observer reliability. Again the mean of sessions were consecutively calculated (i.e. mean of tests 2 and 3, mean of tests 2, 3 and 4 etc.) and compared across assessors. This was undertaken to identify the number of individual TPD measurements needed within each TPD assessment to establish a stable result between raters within a clinically reasonable timeframe.

All 3 assessors were comparably experienced physiotherapists with more than 15 years working experience and postgraduate specialisation in Manual Therapy (IFOMPT degree). KE had over 50 hours experience completing the TPD over the previous year period. UW and DK received a 1-day training session prior to beginning the study, conducted by KE. All tests were performed in treatment rooms in a laboratory-based setting in the Hochschule fuer Gesundheit, Bochum, Germany.

Clinical characteristics

The following clinical measures were collected to provide a comprehensive clinical picture of the participants and were in line with international recommendations regarding core sets of outcome measures for back pain research [19, 20]: pain intensity (Brief Pain Inventory Short Form); back related physical Function (Roland Morris Disability Questionnaire); anxiety and depression (Hospital Anxiety and Depression Scale); health related quality of life (Euroquol 5D-3L). All questionnaires existed in a validated German version [21-23]. Demographics including age, gender, height, weight, Body Mass Index (BMI), duration of symptoms and working status were also documented.

Statistical analysis

For the adequate precision of sample estimates of error, Altman [24] advised the recruitment of at least 40 participants for an agreement-type study like ours. Fifty-two people expressed an interest in participating at the outset of the study. Seventeen people withdrew before the first measurements were obtained or did not meet our inclusion/exclusion criteria resulting in a final sample size of 35 participants (figure 1). Although this sample size is smaller than the 40 advised by Altman [24], we have reported 95% confidence intervals (95%CI) for the reliability statistics. These 95%CIs are useful for ascertaining if the precision of estimate affects the overall inferences that are made.

The greater the number of consecutive measurements averaged within a protocol period, the closer the average of these measurements approaches the "true value" [25]. Nevertheless, clinic time is obviously not exhaustive. Therefore, we examined intra- and inter-observer reliability for a range of consecutive measurements made within the protocol period. The mean (SD) systematic bias (and associated 95% confidence interval) between data collected in repeated protocols and between different assessors was first quantified using a paired t-test.

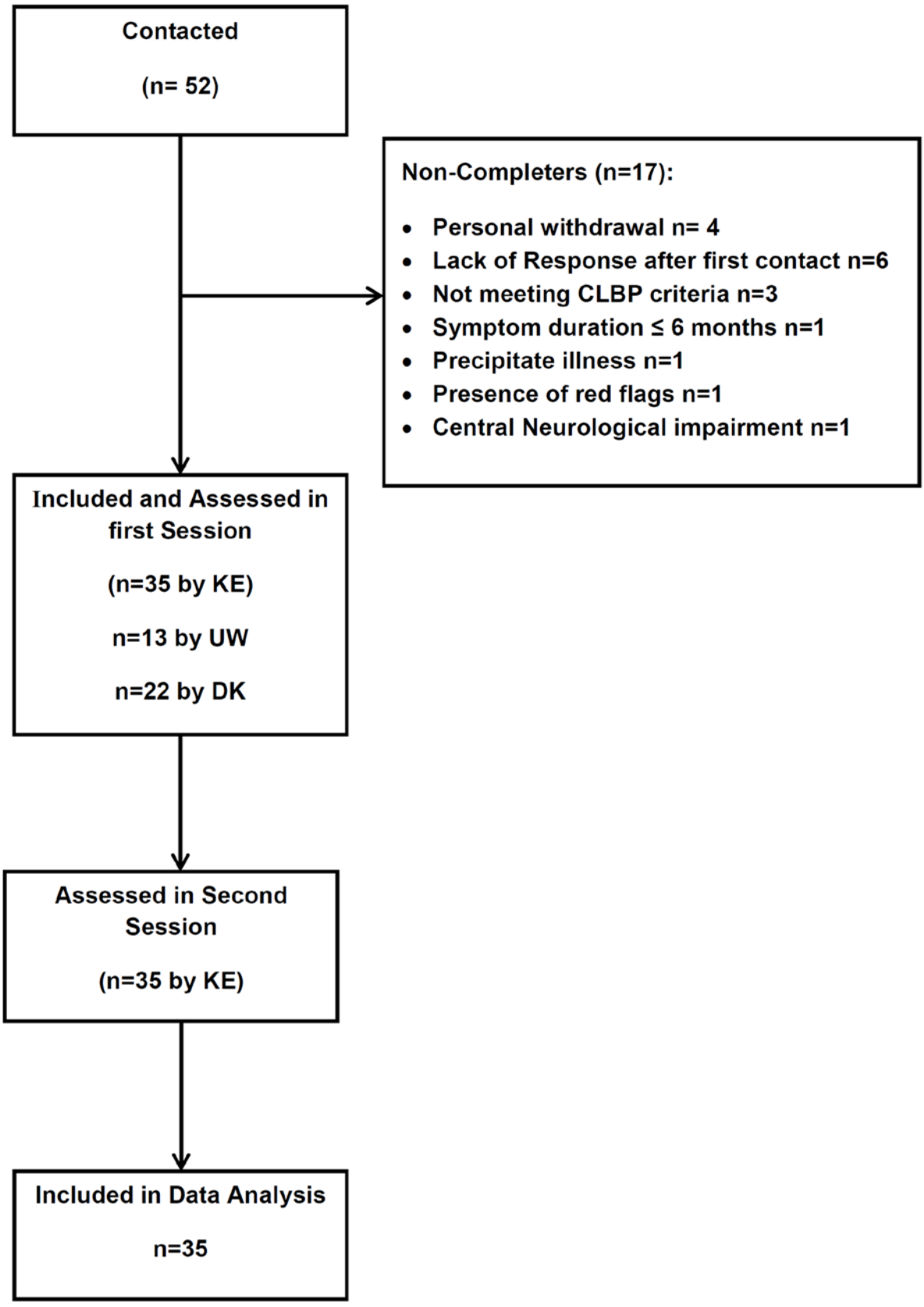
Random error within and between assessors' measurements was quantified with the within-subjects SD (standard error of measurement), coefficient of variation, limits of agreement, and a random-error only (model 3.1) intra class correlation coefficients (ICC). Correlations which collapse different components of bias, as well as random error between and within assessors have been criticized in the literature for obfuscating separate sources of variability [25-27]. The within-subjects SD was then used as an input in statistical power calculations to estimate whether the random measurement error was small enough to detect a clinically relevant change in TPD with a feasible sample size [25, 28].

RESULTS

Participant Characteristics

Fifty-two people enquired about the study of which 35 met the inclusion criteria and consented to participate; and all those who consented to participate completed the study (figure 1). Of the 35, 20 were employed, 10 retired, 3 had retired early due to back pain, and 2 were on sick leave. The overall average for the intra-observer TPD data was 50.5mm (SD 19.2mm). The average levels of pain severity at time 1 and 2 were 3.6 and 3.5 respectively, defined as mild-to-moderate severity [29].The average back related physical function was 7.5, similarly defined as a mild-to-moderate disability [30]. For all participants, the area of pain included the L5 level. The participant characteristics are detailed in table 1.

Figure 1: Flow Chart of Participant Recruitment Process, Assessment for Eligibility, Testing and Data Analysis



Intra-observer reliability

The mean difference between test days 1 and 2 in participants' levels of reported pain severity was 0.12 arbitrary units (95% CI: -0.25 to 0.48, p=0.52). The difference in pain interference scores between day 1 and 2 was 0.53 arbitrary units (95% CI: 0.27 to 0.80, p<0.01).

Table 1: Participant characteristics and clinical measures

	Mean (SD[a])	Range
Participant characteristics		
Age (years)	52 (15)	22 - 79
Sex	12♂ 23♀	NR[b]
Height (m)	1.72 (0.1)	1.58 - 2.00
Weight (kg)	76 (17)	47 - 105
Body Mass Index (kg/m^2)	25 (5)	19 - 35
Symptom Duration (years)	11 (11)	0.5 - 40
Clinical measures		
BPI[c] T1[d]:		
Pain Severity	3.6 (2.0)	0.0 – 7.3
Pain Interference	2.6 (2.0)	0.0 – 8.4
BPI T2[e]:		
Pain Severity	3.5 (1.9)	0.0 – 6.8
Pain Interference	2.1 (1.8)	0.0 – 6.9
RMDQ[f]	7.5 (4.6)	1 – 18
HADS[g]:		
Anxiety	5.2 (3.4)	0.0 – 15.0
Depression	4.5 (3.1)	0.0 – 14.0
EuroQol:		
Thermometer (%)	65 (21)	30 – 100
Index Value	0.81 (0.20)	0.18 – 1.00

Legend: [a] SD = Standard Deviation, [b] NR = Not Reported, [c] BPI= Brief Pain Inventory, [d] T1 = Session 1,[e] T2 = Session 2, [f] RMDQ = Roland Morris Disability Questionnaire, [g] HADS = Hospital Anxiety and Depression Scale

The intra-observer reliability statistics for rater 1, and across all fourteen consecutive measurements within the assessment, are shown in table 2. It was judged that taking the

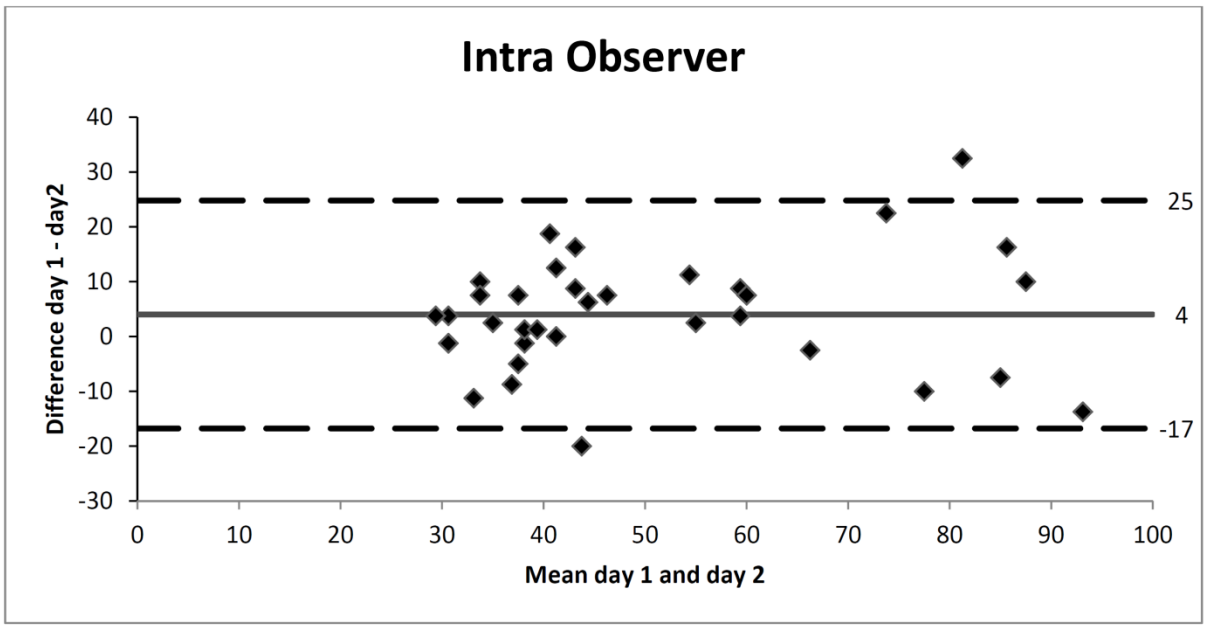
average of consecutive measurements 2 to 5 resulted in the optimum trade-off between measurement stability and the clinic time needed to complete testing. In clinical practice, the shorter the time required to complete the better. The Bland and Altman plot for the individual differences between days is shown in figure 2. The reliability appraisal was based on the reasoning that for the mean of 2 to 5 measurements the systematic bias was less than 5mm (the resolution of the measurement procedure) and the random error began to plateau with further measurements resulting in minimal reductions in error in relation to the measurement resolution. The 2-5 consecutive measurements took approximately 5 min to obtain with each subsequent measure adding approximately 1 min for those participants with the poorest TPD ability.

Table 2: Intra-observer reliability

Number of tests	**2**	**2 to 3**	**2 to 4**	**2 to 5**	**2 to 6**	**2 to 7**	**2 to 8**	**2 to 9**	**2 to 10**	**2 to 11**	**2 to 12**	**2 to 13**	**2 to 14**
Mean session difference	3.4	3.6	5.5	4.0	3.9	3.4	2.7	1.9	1.3	1.3	0.6	0.4	0.3
SD[a] of session differences	17.7	12.8	11.5	10.6	11.4	11.3	11.0	10.8	10.8	10.4	9.8	10.5	10.1
Within-subjects SD (SEM[b])	12.5	9.0	8.1	7.5	8.1	8.0	7.8	7.6	7.6	7.4	7.0	7.4	7.2
Coefficient of variation (%)	24.0	17.7	16.1	14.9	16.1	16.1	15.6	15.4	15.6	15.1	14.3	15.3	14.7
Limits of agreement	34.8	25.1	22.5	20.8	22.4	22.2	21.5	21.1	21.1	20.4	19.3	20.6	19.8
ICC[c]	0.65	0.80	0.81	0.85	0.82	0.82	0.83	0.83	0.83	0.84	0.85	0.83	0.85

Legend: Intra-observer reliability data for day 1 and 2 for rater 1 for 14 test repetitions with test 1 excluded as it was considered a practice test. The values are based upon the cumulative test scores; [a]SD = Standard Deviation, [b]SEM = Standard error of measurement, [c] Intra class correlation coefficient

Figure 2: The Limits of Agreement for intra-observer reliability



Legend 2: For intra-observer reliability, test-retest differences are plotted against the pooled means for two sessions for measurement repetitions 2-5. Mean session differences (systematic bias) are displayed by solid lines and limits of agreement by dashed lines.

Inter-observer reliability

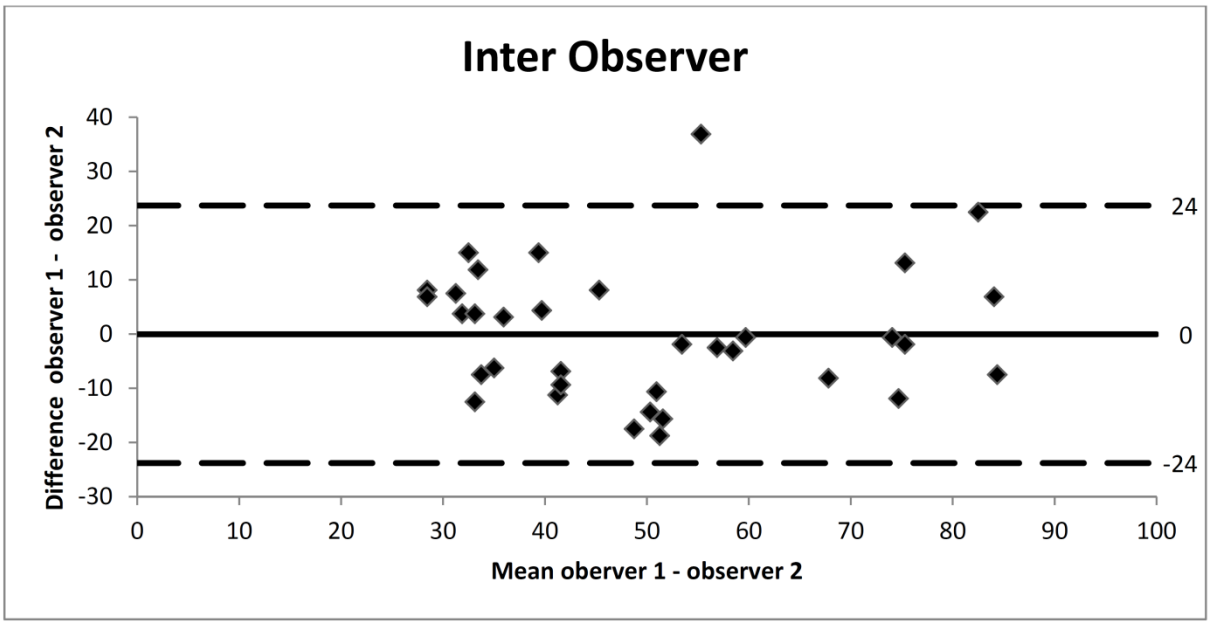
The inter-observer reliability statistics between rater 1 and 2 across all fourteen measurements are shown in table 3. The data for the two second raters were pooled as the inter-observer reliability between rater 1 and both raters 2 and 3 were similar. Using the same reliability criteria as above, it was judged that averaging the 2nd to 9th consecutive measurements resulted in the optimum trade-off between measurement stability and the time needed to complete testing. The Bland and Altman plot for these data is shown in figure 3. The systematic bias was less than one unit of resolution (5mm) regardless of the number of consecutive measurements.

Table 3: Inter-observer reliability

Number of tests	**2**	**2 to 3**	**2 to 4**	**2 to 5**	**2 to 6**	**2 to 7**	**2 to 8**	**2 to 9**	**2 to 10**	**2 to 11**	**2 to 12**	**2 to 13**	**2 to 14**
Mean session difference	3.8	2.2	2.0	1.3	1.4	0.9	0.6	-0.1	-0.8	-1.0	-1.8	-2.1	-2.2
SD of session differences	22.4	18.8	16.0	15.4	14.6	13.3	13.2	12.8	13.1	13.2	13.3	13.0	13.0
Within-subjects SD[a] (SEM[b])	15.8	12.8	10.9	10.5	10.3	9.3	9.0	8.6	8.7	8.7	8.7	8.6	8.3
Coefficient of variation (%)	29.9	23.9	20.4	19.5	19.5	17.8	17.3	16.7	17.1	17.2	17.3	17.0	16.5
Limits of agreement	43.8	35.4	30.3	29.0	28.6	25.9	24.9	23.7	24.0	24.2	24.2	23.7	23.0
ICC[c]	0.30	0.53	0.66	0.70	0.70	0.74	0.75	0.76	0.75	0.74	0.74	0.75	0.77

Legend: Inter-observer reliability data for observer 1 and observer 2 for 14 consecutive measurements with test 1 excluded as it was considered a practice test. The values are based upon the cumulative measurement scores. The values for rater 2 are the combined values for DK who measured 22 of the participants and UW who measured 13 of the participants. Data for both 2nd raters were sufficiently similar with respect to reliability with rater 1 to allow pooling of the data; [a]SD = Standard Deviation, [b]SEM = Standard error of measurement, [c]ICC= Intra class correlation coefficient

Figure 3: The limits of agreement for inter-observer reliability



Legend 3: For inter-observer reliability, the differences from rater 1 and the pooled differences for rater 2 are plotted against the pooled means for measurement repetition 2-9. Mean session differences (systematic bias) are displayed by solid lines and limits of agreement by dashed lines.

Discussion

The overarching aim of this study was to develop and quantify the intra- and inter-observer reliability of a novel lower back TPD assessment protocol, which minimises subjective clinical judgement, in patients with NSCLBP. The study had two objectives - to establish: 1. the minimum number of TPD measurements required within the assessment protocol to maximise intra-observer reliability whilst minimising the time required to complete the test; and 2. the minimum number of TPD measurements required within the assessment protocol to maximise inter-observer reliability whilst minimising the time required to complete the test?

Five measurements (with the first used as a practice trial only and not used to calculate the mean) was identified as the minimum number of TPD measurements required within the TPD

assessment protocol to maximise intra-observer reliability whilst minimising the time required to complete the test (approximately 5 minutes). Nine measurements (with the first used as a practice trial only and not used to calculate the mean) was identified as the minimum number of TPD measurements required within the TPD assessment protocol to maximise inter-observer reliability whilst minimising the time required to complete the test (approximately 9 minutes).

Only one previous study has investigated the intra- and inter-observer reliability of the TPD for the lower back [14]. In this study, 28 clinicians assessed the TPD of the lower back in 28 healthy young adults. The mean TPD reported was 55.5mm (SD 12.7mm) with an intra-observer ICC of 0.81 and inter-observer ICC 0.86; and an intra-observer limits of agreement of (mean difference [lower limit to upper limit]) 0.6mm [-14.1 to 15.4] and an inter-observer reliability limits of agreement of (1.9mm [-19.0 to 22.8]). In our study, the mean TPD reported was 50.5mm (SD 19.2mm) with an intra-observer ICC of 0.85 and inter-observer ICC 0.76; and an intra-observer limits of agreement of (mean difference [lower limit to upper limit]) 4.0mm [-16.8 to 24.8] and inter-observer reliability limits of agreement of (-0.1mm [-23.8 to 23.7]). Broadly, the level of systematic error and the ICCs were similar between studies though our study had slightly wider limits of agreement. This is likely due to the inherent greater variability that would be expected in participants with low back pain compared with healthy participants.

In a meta-analysis, Catley et al [31] indicated that a minimal clinically important difference in TPD between NSCLBP patients and healthy controls is 11.7 mm (26%). Using these values as a basis for a power estimation and our intra-observer reliability within-subjects SD of 7.5 mm, it can be estimated that 11 participants would be required for a future single arm pre-post study (two-tailed $P<0.05$, statistical power = 90%). We also estimate that 44 participants (22

in each study arm) would be required for a future two-arm randomised controlled trial, which is a reasonably achievable sample size within the musculoskeletal rehabilitation research context. Hence, our TPD assessment protocol can be seen as possessing acceptable intra-observer reliability as a measure of tactile acuity for research purposes.

Whether this measure is sufficiently reliable to detect change on an individual patient basis within a clinical setting is less clear. The SEM (or typical error) identified in our study was 7.5mm, which is below the MCID of 11.7mm in the literature. Nevertheless, with 95% limits of agreement of ±20.8 mm, it can be estimated that an individual back pain patient could change, in a worst scenario, by as much as 20.8 mm due to normal variation with this measure. This questions its usefulness in clinical practice, based on the assumption that 11.7 mm is a clinically relevant change. This minimum clinically important difference (MCID) was estimated from a systematic review comparing individuals with back pain to healthy controls [15] rather than from a formal estimation of what is a clinically relevant improvement in patients with back pain. Further work needs to be undertaken to establish the MCID for back pain patients before firm conclusions can be made about the reliability of our TPD assessment protocol. Reliability decisions are inherently contingent on the magnitude of the MCID [25].

It has previously been identified that the need for assessor judgement to quantify the exact TPD in previous protocols has introduced considerable capacity for bias [14]. This study is the first to present a detailed protocol which eliminates that judgement. While it is not possible from our data to examine the effect this judgement elimination has had on reliability, it is reasonable to suggest that it will help to reduce bias. It would be interesting to compare our TPD protocol with previous protocols that did require assessor judgement to investigate what effect, if any, assessor judgment has on TPD reliability in the lower back region.

Clinical Implications

The implications of the higher inter-rater reliability compared with the intra-rater reliability are that fewer measurements are needed when 1 clinician is treating and monitoring the same patient, than when test results from more than one clinician are being used. Our findings indicate that the mean of 5 TPD measurements (with the first considered a practice and not used to calculate the mean) provided the optimum balance of intra-observer reliability aspects and practical duration of testing with one clinician. For two clinicians, we identified 9 measurements (with the first considered a practice and not used to calculate the mean) as the optimum balance of inter-observer reliability aspects and duration of testing. These figures should be interpreted cautiously in light of the limitations outlined below.

Strengths and limitations

One limitation in all measurement studies is sample size and, although 52 participants were contacted initially to take part, our final sample size for analysis was slightly lower than established recommendations of 40 participants [24]. This recommendation was based on the adequate precision of estimate of a sample SD. Precision of statistical estimates is indicated by the 95%CIs. It can be seen in table 4 that these are sufficiently narrow not to alter our conclusions that the assessment protocol is useful for research but unclear in terms of reliability of clinical decisions made on individual patients. Our participant sample size is also somewhat larger than those in other related studies [14, 32, 33].

Table 4: Intra- and Inter-observer reliability for the recommended number of test repetitions

	Intra-observer	Inter-observer	
Number of tests	**2 to 5**	**2 to 5**	**2 to 9**
Mean session difference (95% CI)	4.0 (0.49-7.51)	1.3 (-3.84-6.34)	-0.1(-4.07-3.96)
SD[a] of session differences	10.6 (8.6-13.9)	15.4 (12.5-20.2)	12.8 (10.4-16.8)
Within-subjects SD (SEM[b])	7.5 (6.1-9.8)	10.5 (8.8-14.2)	8.6 (7.4-11.9)
Coefficient of variation (%)	14.9 (12.1-19.4)	19.5 (16.4-26.5)	16.7 (14.4-23.2)
Limits of agreement	20.8 (16.9-27.2)	29.0 (24.5-39.6)	23.7 (20.4-32.9)
ICC[c] (95% CI[d])	0.85 (0.72-0.92)	0.70 (0.48-0.84)	0.76 (0.57-0.87)

Legend: Intra-observer reliability data for day 1 and 2 for rater 1 for 5 measurements with measurement 1 excluded as it was considered a practice measurement. Inter-observer reliability data for rater 1 and rater 2 (pooled) for 5 and 9 measurements with measurement 1 excluded respectively. The values are based upon the cumulative test scores. The values for the 95% Confidence Intervals were calculated according to the recommended method by Zar [34]; [a]SD = Standard Deviation, [b]SEM = Standard error of measurement, [c]ICC= Intra class correlation coefficient, [d]CI = Confidence Interval

One practical limitation of our assessment protocol is that it is carried out in prone lying. While this in line with other protocols, there is the possibility that for some people with back pain this may be uncomfortable, which could affect the practicality of testing. Future studies could be carried out to investigate the measurement properties of this assessment protocol in sitting.

Because of our limited number of assessors and the fact that they were experienced in the use of the protocol, the findings of the present study should not be directly generalized to any other assessor using the same protocol. That will require further work with a larger number of

assessors. Also, all of our assessors were experienced and so we cannot comment on the influence of experience. That said, a previous study has suggested that this factor is of limited importance [14].

In addition, a further limitation could constitute the use of catch trials. We used catch trials in keeping with previous protocols [12, 17, 35] in order to account for potential effects of participants guessing. However, patients may be more likely to report the application of a stimuli after a catch trial as 'two points' simply because it feels different (e.g. *duller or coarser*) from "one point" but not because of any perceived difference in spatial acuity [36]. Thus the use of catch trials could artificially lower a patient's TPD. This may explain why the TPD measurements reported in our study are lower than those reported by others [15]. Further work could compare the current assessment protocol to another not using catch trials or employing them differently to weigh up the effects of these catch trials.

The primary strength of this study was the comprehensive range of statistical estimates used for quantifying the systematic and random changes that can occur in a reliability study [25]. We concentrated on absolute indicators of reliability such as the within-subjects SD in order to arrive at our conclusions, especially in terms of extrapolating how the degree of reliability might impact on future research. Standard procedures for quantifying systematic bias tend to include the reporting of ICCs. However, in samples with large heterogeneity, a high ICC might obfuscate substantial and clinically relevant random error [37]. It also does not make sense to use a type of ICC which combines systematic and random errors into a single value, because the solutions to reducing systematic and random errors can be very different [26].

Another strength of the present study constituted the TPD focus on the population of patients with NSCLBP. As mentioned earlier, the TPD reliability data in a young and healthy population displayed a large variability [14]. With the expectation of an even greater

heterogeneity in a population with back pain, here was a need to specifically look at reliability of the test in this population.

A final strength of this study was that the participants were clinically stable between assessment times 1 and 2. While there was a statistically significant improvement in pain interference between time 1 and 2 this difference was lower than 1 unit on the BPI, which is not considered to be clinically relevant. [38]. Such stability between measures is an important methodological issue in reliability studies [39].

Conclusion

This study presented data on the development and reliability assessment of a novel TPD assessment protocol for the lower back. The protocol attempted to overcome previously noted limitations of TPD measurement protocols that require assessor judgement. This study found that five measurements (with the first used as a practice trial only and not used to calculate the average) was the minimum number of TPD measurements required within the assessment TPD protocol to maximise intra-observer reliability whilst minimising the time required to complete the assessment. Additionally, nine measurements (with the first used as a practice trial only and not used to calculate the average) should be used as the minimum number of TPD measurements within the assessment protocol to maximise inter-observer reliability whilst minimising the time required to complete the assessment. The protocol described demonstrates a level of intra-observer reliability sufficient for research purposes. However, it is unclear, as yet, whether the level of reliability is sufficient for individual patient measurements in clinical practice.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

1 Moseley GL. Pain, brain imaging and physiotherapy—Opportunity is knocking. Manual therapy 2008;13(6):475-7.

2 Flor H. Cortical reorganisation and chronic pain: implications for rehabilitation. Journal of Rehabilitation Medicine 2003;35(0):66-72.

3 Flor H, Denke C, Schaefer M, Gruesser S. Effect of sensory discrimination training on cortical reorganisation and phantom limb pain. The Lancet 2001;357(9270):1763-4.

4 Moseley L, Flor H. Targeting cortical representations in the treatment of chronic pain: a review. Neurorehabil Neural Repair 2012;26(Suppl 6):646 - 52.

5 Wand BM, O'Connell NE, Di Pietro F, Bulsara M. Managing Chronic Nonspecific Low Back Pain With a Sensorimotor Retraining Approach: An Exploratory Multiple-Baseline Study of 3 Participants. Physical Therapy 2011.

6 Flor H, Braun C, Elbert T, Birbaumer N. Extensive reorganization of primary somatosensory cortex in chronic back pain patients. Neuroscience Letters 1997;224(1):5-8.

7 Wand BM, Abbaszadeh S, Smith AJ, Catley MJ, Moseley GL. Acupuncture applied as a sensory discrimination training tool decreases movement-related pain in patients with chronic low back pain more than acupuncture alone: a randomised cross-over experiment. British Journal of Sports Medicine 2013;47(17):1085-9.

8 Diers M, Koeppe C, Diesch E, et al. Central processing of acute muscle pain in chronic low back pain patients: an EEG mapping study. Journal of Clinical Neurophysiology 2007;24(1):76-83.

9 Nolan MF. Quantitative measure of cutaneous sensation two-point discrimination values for the face and trunk. Physical Therapy 1985;65(2):181-5.

10 Moseley GL, Flor H. Targeting Cortical Representations in the Treatment of Chronic Pain A Review. Neurorehabilitation and neural repair 2012;26(6):646-52.

11 Moberg E. Two Point Discrimination Test. Scandinavian Journal of Hand Surgical Rehabilitation 1990;22:8.

12 Moseley GL. I can't find it! Distorted body image and tactile dysfunction in patients with chronic back pain. Pain 2008;140:239-43.

13 Loumajoki H, Moseley G. Tactile acuity and lumbopelvic motor control in patients with back pain and healthy controls. Br J Sports Med 2011;45:437 - 40.

14 Catley MJ, Tabor A, Wand BM, Moseley GL. Assessing tactile acuity in rheumatology and musculoskeletal medicine—how reliable are two-point discrimination tests at the neck, hand, back and foot? Rheumatology 2013.

15 Catley MJ, O'Connell NE, Berryman C, Ayhan FF, Moseley GL. Is tactile acuity altered in people with chronic pain? A systematic review and meta-analysis. The Journal of Pain 2014;15(10):985-1000.

16 Dinse HR, Kleibel N, Kalisch T, Ragert P, Wilimzig C, Tegenthoff M. Tactile coactivation resets age-related decline of human tactile discrimination. Annals of Neurology 2006;60(1):88-94.

17 Wand B, Di Pietro F, George P, O'Connell NE. Tactile thresholds are preserved yet complex sensory function is impaired over the lumbar spine of chronic non-specific low back pain patients. A preliminary investigation. Physiotherapy 2010;96:317-23.

18 Merz O, Wolf U, Robert M, Gesing V, Rominger M. Validity of palpation techniques for the identification of the spinous process L5. Manual therapy 2013.

19 Deyo RA, Battie M, Beurskens AJHM, et al. Outcome Measures for Low Back Pain Research: A Proposal for Standardized Use. Spine 1998;23(18):2003-13.

20 Bombardier C. Outcome Assessments in the Evaluation of Treatment of Spinal Disorders: Summary and General Recommendations. Spine 2000;25(24):3100-3.

21 Radbruch L, Loick G, Kiencke P, et al. Validation of the German Version of the Brief Pain Inventory. Journal of pain and symptom management 1999;18(3):180-7.

22 Wiesinger GF, Nuhr M, Quittan M, Ebenbichler G, Woelfl G, Fialka-Moser V. Cross-Cultural Adaptation of the Roland-Morris Questionnaire for German-Speaking Patients With Low Back Pain. Spine 1999;24(11).

23 Oesch P, Hilfiker R, Keller S, et al. *Assessments in der Rehabilitation Band 2: Bewegungsapparat*. Bern: Verlag Hans Huber; 2011.

24 Altman DG. *Practical statistics for medical research*. New York;London;: Chapman and Hall; 1991.

25 Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports medicine 1998;26(4):217-38.

26 Atkinson G, Nevill A. Comment on the use of concordance correlation to assess the agreement between two variables. JSTOR; 1997.

27 Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. The Journal of Strength & Conditioning Research 2005;19(1):231-40.

28 Batterham AM, Atkinson G. How big does my sample need to be? A primer on the murky world of sample size estimation. Physical Therapy in Sport;6(3):153-63.

29 Palos GR, Mendoza TR, Mobley GM, Cantor SB, Cleeland CS. Asking the Community About Cutpoints Used to Describe Mild, Moderate, and Severe Pain. The Journal of Pain 2006;7(1):49-56.

30 Roland M, Fairbank J. The Roland Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine 2000;25(24).

31 Catley MJ, Tabor A, Miegel RG, Wand BM, Spence C, Moseley GL. Show me the skin! Does seeing the back enhance tactile acuity at the back? Manual Therapy 2014;19(5):461-6.

32 Luomajoki H, Lean C, Di Pietro F, Ariaksinen O. Reliability of movement control tests in the lumbar spine. BMC Musculoskeletal Disorders 2007;8(90).

33 Tidstrand J, Horneij E. Inter-rater reliability of three standardized functional tests in patients with low back pain. BMC Musculoskeletal Disorders 2009;10(1):58.

34 Zar J. Biostatistical analysis (p. 663). NewJersey: Prentice Hall 1999.

35 Peters ML, Schmidt AJ. A comparison of two-point discrimination threshold of tactual, non-painful stimuli between chronic low back pain patients and controls. Pain 1991;44(1):57-60.

36 Johnson KO, Van Boven RW, Hsiao SS. The perception of two points is not the spatial resolution threshold. Touch, temperature, and pain in health and disease: Mechanisms and assessments 1994:389-404.

37 Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Computers in biology and medicine 1990;20(5):337-40.

38 Dworkin RH, Turk DC, Wyrwich KW, et al. Interpreting the Clinical Importance of Treatment Outcomes in Chronic Pain Clinical Trials: IMMPACT Recommendations. The Journal of Pain 2008;9(2):105-21.

39 Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist manual. Amsterdam: VU University Medical Centre 2009.