Author's response to reviews

Title: Inter-examiner reproducibility of tests for lumbar motor control

Authors:

Flemming FE Enoch (fe@fysiq.dk)
Per PK Kjaer (pkjaer@health.sdu.dk)
Arne AE Elkjaer (arne.e@mail.dk)
Lars LR Remvig (lars.remvig@rh.regionh.dk)
Birgit BK Juul-Kristensen (bjuul-kristensen@health.sdu.dk)

Version: 3 Date: 27 January 2011

Author's response to reviews: see over
To the Editor:

Please find below our responses to all the questions and suggestions raised by you and the two reviewers. We hope that you find our answers sufficient and acceptable, covering all the aspects raised.

Yours sincerely,

Flemming Enoch

Response to the Editor’s comments:

Thank you very much for the thorough review and for the many relevant useful questions and proposals.

As a general comment, we wish to emphasise that our study is a study of test reproducibility – to test the ability of a test (or, as in this study, several tests) to give the same results when used by different examiners.

The examiners were experienced clinicians and trained in the test techniques – so as to reduce the variability in the data attributable to varying levels of experience and maximise measurement of the variability attributable to the inter-tester reproducibility of each test.

So, it is important to be aware that this is in essence not about reproducibility of a clinical diagnosis.

Internationally, there is no consensus about the terminology within this area, in particular when one talks about reliability versus reproducibility. Reviewer 2 gives reference to C. Terwee et al. 2007 [1], however, we wish to refer to another publication by the same group (LB. Mokkink et al. 2010)[2] in which they use a Delphi study to try to reach consensus on the definitions of the two words. Only by interference by the steering committee (which is against the principles of a Delphi study) did they reach consensus that was contrary to the dictionary definition after the second panel round. Worth noting is that it was an international panel, meaning that they didn’t have English as their first language.

According to Webster’s Encyclopedic Unabridged Dictionary [3]:

1. Reliability means “Trustworthy, which apply to persons, objects, etc that can be depended upon with confident certainty.”

2. Reproducibility – ability to reproduce - means “to produce again, as a play produced at an earlier time.”

3. Validity means “the state of being valid” and valid means “producing the desired result”

In our opinion this illustrates that reliability includes reproducibility and validity.

Reviewer no 1 (Luciana Macedo)
I would like to thank the authors for the opportunity to review this well designed and well written paper. I have only a few recommendations.

Major Compulsory Revisions

1. I believe that the authors have erroneously used the term diagnosis or diagnostic assessment. To my understanding diagnosis or diagnostic tests would deal directly with the identification of the nature/cause/tissue source of pain of low back pain. In this study the authors use the term diagnosis as a synonym to a functional assessment to identify deficits in motor control. I think the two are not the same.

   Answer: In our opinion this is a matter of semantics. One can make a patho-anatomic diagnosis such as osteoarthritis and in other circumstances one can make a diagnosis of a syndrome such as C1/C2 segmental dysfunction or loss of motor control. We have used the word in a broad sense.

2. Also, the authors mention that the tests measure motor control. However, I may disagree with this denomination as motor control relates to a neuromusculoskeletal progress. If the tests are not measuring motor control but position sense and control of the lumbo pelvic stability, the tests should called as such. Additionally, I think it would the study would be strengthened if the authors explain what specifically the tests are measuring and how that would affect the “motor control” of the spine. Overall it would be nice to know why one would choose to use these tests in clinical practice and how what is being assessed on the tests influence low back pain. The last paragraph of the introduction addresses in part this issue but I think this should be clearer.

   Answer: We have changed the term to ‘reposition and control of the lumbo pelvic complex’. The last part of the introduction has been rewritten to give a clearer understanding of how these tests can be used in clinical practice to get a better understanding about the patient movement pattern effect measure on LMC.

3. Why did the authors include in the sample asymptomatic patients? What is the clinical applicability of this? I don’t think this is relevant as the test would be commonly applied in patients with low back in which the reproducibility could be completely different due to the levels of pain and disability that the patients may present during testing conditions. I think that if the authors choose to maintain asymptomatic patients in their sample it would be interesting to present the results separately.
Answer: Inclusion of healthy controls/non-back pain patients was done for several reasons: 1) when performing inter-examiner reproducibility studies, it is of utmost importance that the examiners are as unbiased as possible. 2) Also, when analysing results with correlation coefficients, it is important to be sure that all possible measurement levels can be presented in the data set. 3) A priori we did not know whether healthy controls and patients with back pain did behave in the same way when tested with these tests. Consequently, we included approximately 35% (15 subjects) healthy controls, as well as patients with varying duration of low back pain (LBP) to cover the potential variation in test results. Eventually, as we already have mentioned in the discussion: "...the study was carried out on LBP and non-LBP subjects, for whom the test battery is intended, making the results relevant for screening purposes within this group."

However, we agree that it is interesting to analyse whether there is a difference in the reproducibility of the test when used with people who do and do not have back pain, even though the populations for each of these analyses are small. This is added as text (Figure 3) and described in the results section.

As suggested, we did a post hoc analysis for people with and without LBP separately. However, there were no statistically significant differences between the results from each group when looking at ICC 95% confidence intervals. When looking at the limits of agreement, there also were no marked differences. Furthermore, we have changed the symbols in the Bland-Altman plots in order to see more clearly that there is no systematic difference between people with and without LBP on the day of the examination.

**Minor Essential Revisions**

4. On the third phrase of the abstract the authors state that the reproducibility of LMC tests is unknown. This is not correct, as many studies have tested the reliability of other LMC tests not addresses in this study.

Answer: We apologise for this imprecise statement. The text is specified in the abstract (background) to “quantitative tests” which to our knowledge have not been reported sufficiently.

5. I think that the authors should consider including in their introduction what type of back pain the study is dealing with. Is it non-specific low back pain?

Answer: This information is now given on page 7 in the study sample section.

6. Second last line of page 2 (background). “The complex anatomy of the lumbo-pelvic region and the multidimensional…” Shouldn’t it be multidirectional?
Answer: We agree ‘multidirectional’ is more appropriate when talking about biomechanical issues. The text has been revised to accommodate this.

7. It seems on the introduction that the authors call LMC a specific/defined group of tests. However, the term lumbar motor control infers not only to the tests evaluated in this study but a number of other tests not addresses in this study. Additionally some of these tests (such as thickness of TrA) have had its reliability tested by many other studies. I think that the introduction would read better if the authors clearly identify the tests of interest before mentioning their reliability in the introduction.

Answer: We agree motor control is much broader than these 5 tests. We have inserted the following text in the background: ‘In this study LMC includes tests regarding the ability to control and reposition the lumbo-pelvic complex when challenged in different directions’. This does not include the assessment of TrA.

8. The authors mention that the study was a three-phase study. What was the phase three? The authors mention: three phase reproducibility and validity study protocol. What does validity study protocol mean? Are you talking about feasibility?

Answer: We used a three-phase study protocol for reliability in kappa studies as a background for designing our protocol. However, since this study deals with continuous data, we selected a two-phase protocol, excluding the agreement phase. The text is changed accordingly.

9. Approximately 50% of patients had LBP. This is not clear to readers. It would be better to present the exact percentage.

Answer: We agree. Now the exact figures are apparent from Table 2 and we have given exact percentages in the text of Table 2: Pain on day of examination: no pain 15=37.5%, 1-3: 13=32.5% >3 12=30%

10. Discussion 5th paragraph. “The study phase had an almost 50% prevalence of symptomatic subjects, which strengthens the statistical opportunity to demonstrate agreement for a given actual agreement.” This phrase is confusing.
and needs better explanation. “Finally, the study was carried out on LBP and non-LBP subjects, for whom the test battery is intended, making the results relevant for screening purpose within this group.” I don’t understand in which case scenario this type of test would be used clinically in asymptomatic patients.

Answer: These types of tests are mainly used in sports to screen athletes. It is believed that there is a correlation between poor lumbo-pelvic control and the risk of injury in the lumbo pelvic area. Further research will detect whether this screening is worthwhile doing. This is added to the discussion. We agree the phrase with 50% was confusing. It has been deleted.

**Discretionary Revisions**

11. I suggest that the authors change the title of the manuscript to “Inter-rater reliability of tests of lumbar motor control” this way the title would include information on what the study addresses. The same way I suggest changing the word reproducibility that is used throughout the study to reliability.

Answer: We appreciate this suggestion, however as mentioned above, we do not agree with the terminology reliability. We have changed the title to ‘Inter-examiner reproducibility.… ‘to underline that two testers are performing the study.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:** I declare that I have no competing interests

**Reviewer no 2. (Leonardo Costa)**

I found this paper interesting, specially for physiotherapists. I found some issues that should be clarified (see below).

**Minor revisions**

**Abstract:**

1. Reliability is not the same as reproducibility and validity, they are distinct concepts (see Terwee et al(1) to clarify this).
**Answer:** We do agree with this statement. However, there is no general consensus about the terminology. In a Delphi study 55% of the panellists were in favour of the term reproducibility instead of reliability; only due to an intervention of the steering committee (which is an unacceptable interference in a Delphi Study) did the panellists decide to recommend the term reliability [2].

In English “reliability” means something like trustworthy, which to us means that repeated test performances achieve the same result (reproducibility) and that the result of the test really expresses what the test is intended to express (validity). Consequently, we prefer to use the term “reliability” as an umbrella term for “reproducibility” and “validity”.

Before testing for validity you must be sure that the test you wish to test is reproducible, and this is why this study is about reproducibility.

The present terminology is defined and explained at the beginning of the Methods section.

2. **CCC? Do the authors actually mean ICC?**

No, we used concordance correlation coefficient, which we argued for and provided reference for. However, due to comments from the editor (and the fact that in this case CCC was equivalent to ICC) we have changed it to ICC and given new references in the statistical method section and left out the discussion on ICC versus CCC in the discussion.

**Essential revisions**

**Main text**

Methods: the sentence “When examiner A had tested a subject, the subject was examined by examiner B, and vice versa.” ...is a bit confusing for me. As stated above, there is some confusion with regards to reproducibility (which is an umbrella term for reliability and agreement), the authors also mentioned a test named “concordance correlation coefficient” which I am assuming that is the “Intraclass correlation coefficient”... if so, the authors need to name the type of ICC, because ICCs vary considerably depending on which type you use (see paper from Krebs for reference (2)). Given the extremely high “CCC” values obtained, I am assuming that the authors used the ICC type 3,1 which tends to overestimate the reliability estimates.

**Answer.** See above regarding examiners. We used CCC and compared it with a large one way Anova (type 1.1) using the STATA command “loneway”, which gave almost exactly the same results. As suggested by the editor, we have now changed to using ICC (type 2.1) and reported these results instead. More below.
4. I suggest the authors to demonstrate another agreement parameter other than Bland and Altman Plots (please keep the plots!), such as the SEM (Standard error of the measurement). SEM values gives us agreement estimates and they are expressed not in a dimensionless scale ranging from 0 to 1 (as ICCs do), but SEMs are expressed using the measurement units of the test itself (for example in centimetres or units of a questionnaire), and therefore SEMs are more informative than ICCs. SEMs are easy to calculate and very useful for clinicians. There are a lot of papers showing high ICCs which moderate to low SEMs(3). The time interval “same day” is too short to provide reasonable estimates of reproducibility. Ideally it should be from 2-7 days. This also explains partially the extremely high “CCCs” observed.

**Answer.** We appreciate the suggestion. However, we find that the Bland-Altman plots are giving a very precise image of the actual disagreements in the units used. To make this more comparable, we have changed the scales so all plots are on the same scale. We do not think that SEM is a better estimate of reproducibility nor do we think it is more easily understood by clinicians.

The aim was to test the reproducibility of this measurement method, thus eliminating day-to-day variation of the tests. If there is a larger time span between the tests it is much more likely that intra-individual changes in performance would be influencing the results.

B5. **Table 1. It would be nice if you provide the demographics from normal participants and patients separately.**

**Answer.** As suggested, we now have reported the information separately in Table 3

**Reference**

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:** I declare that I have no competing interests
**Editors Comment: Tests are normally evaluated in the target population so I am unsure why you included people without LBP. At a minimum you need to present the results separately for those with and without LBP; it is quite feasible that the reliability would be different in the two groups. You should also clarify how many of the LBP participants were currently seeking care for their LBP. Please also justify why you included non-clinical subjects.**

**Answer:** Inclusion of healthy controls/non-back patients was done for several reasons: 1) when performing inter-examiner reproducibility studies it is of utmost importance that the examiners are as unbiased as possible. 2) Also, when analysing results with correlation coefficients it is important to be sure that all possible measurement levels can be presented in the data set. 3) A priori we did not know whether healthy controls and patients with back pain did behave in the same way when tested with these tests. Consequently, we included approximately 35% healthy controls, as well as patients with varying duration of LBP to cover the potential variation in test results. Eventually, as we have already mentioned in the discussion: "...the study was carried out on LBP and non-LBP subjects, for whom the test battery is intended, making the results relevant for screening purposes within this group."

In the study sample section, the information about those subjects seeking care has now been inserted.

**Editors Comment: The IAMMM described how to perform the tests not the evaluation of reliability. Please revise the first sentence of the methods.**

**Answer:** The IAMMM protocol does not describe how to perform the tests, but describes the optimal way for performing a reproducibility study. Consequently there is no need for a revision of the first sentence in Methods. However we have deleted: ".. and validity study..." and changed the word "proposed" into "recommended".

In English “reliability” means something like trustworthy, which to us means that repeated test performances achieve the same result (reproducibility) and that the result of the test really expresses what the test is intended to express (validity). Consequently, we prefer to use the term “reliability” as an umbrella term for “reproducibility” and “validity”.

Also to specify our perception of "reliability", we have introduced a short section in the first part of the methods section, where we define Reliability, Reproducibility and Validity.
3. Editors Comment: Please replace the CCC analyses with ICC (2,1). I am unconvinced of the advantage of the CCC and the ICC is more widely understood by our readers. Please define the LOA measure you calculated. Please add the MDC as requested by the reviewer. Please report the statistical package used for the analyses.

Answers: The questions are already mentioned in the discussion and in the statistical methods section:
"In the current study, we have used CCC [37] as a measure of correlation between the two examiners, because of the concerns raised about the intra-class correlation coefficient (ICC) most often used in studies of correlation [41]. CCCs are more robust to the order of examination than ICCs. Post hoc ICCs were also calculated, to enable comparison of the current results with other studies. The differences between the two methods (CCC and ICC) of estimating correlation coefficients revealed a difference of less than 0.002, making the results directly comparable."
"...the STATA statistical package was used (Stata Corp., 2000, Stata Statistical Software: Release 10.1, College Station, TX)."
In the previous calculation we have used the large one-way ANOVA command (type 1.1) in STATA to compare CCC to ICC. We have addressed the issue of type of ANOVA and therefore, now have reported the results following a two-way ANOVA analysis. Interestingly, there are no statistically significant differences although the ICCs are slightly higher. The statistical methods have been changed accordingly.

We do not understand MDC. If you are referring to SEM, see above. We do not find that this information is necessary. In general, it only provides a standard error for the measurements.

Limit of agreement has been defined more clearly in the statistical methods and in the legend for Figure 1.

4. Editors comment: Can you clarify what was the order of tests and the order of examiners. The manuscript is silent on this issue.

Answer: It is mentioned in Methods that: "When examiner A had tested a subject, the subject was examined by examiner B, and vice versa."
However, to clarify the procedure the text has been changed into: “The subjects were examined independently and in random order by the two examiners on the same day. When examiner A had tested a subject, the subject was examined by examiner B, and vice versa. Half the subjects started with examiner A and half with examiner B."
Both examiners performed the tests in the same order on each subject now written as 1-5."

5. **Editors comment:** I found the discussion uninspiring and superficial. It left me ambivalent about the study and unsure of how it contributed to the field. You may wish to adopt the structured discussion headings advocated by BMJ to help you provide a stronger discussion.

- ? statement of principal findings
- ? strengths and weaknesses of the study
- ? strengths and weaknesses in relation to other studies, discussing important differences in results
- ? meaning of the study: possible explanations and implications for clinicians and policymakers
- ? unanswered questions and future research

**Answer:** The discussion is now divided into the above mentioned points, and text revisions have been done accordingly here and there.

**Editors comment:** Please replace sentence three of the abstract which is incorrect and contradicts the discussion.

**Answer:** We apologize for this imprecise statement. The text is specified in the abstract (background) to “quantitative test” which to our knowledge has not been reported sufficiently.

7. **Editors comment:** Please include as a limitation of the study that you only included two highly experienced raters and so it is unclear to what extent these results would generalise to other testers.

**Answer:** We have included the following text:

In the abstract “Whether reproducibility of these tests is as good in daily clinical practice, when used by untrained examiners, also needs to be examined” “This will increase the intention to test the test per se and not to test the combination of test and examiner.”

In the discussion section we have included the following text:

‘The weakness of the study is that we do not know the reproducibility of the current tests carried out by inexperienced clinicians, which of course might be different from the reproducibility of experienced clinicians and trained examiners, as also shown in previous studies [4, 5]. However, in case inexperienced examiners have a low inter-examiner test reproducibility our study shows that is should be possible through education and training to obtain high enough skills to perform the tests in a reproducible way.'
8. **Editors Comment:** Informed consent must also be documented. Manuscripts may be rejected if the editorial office considers that the research has not been carried out within an ethical framework, e.g. if the severity of the experimental procedure is not justified by the value of the knowledge gained.

**Answer:** A last sentence has been added to the study design section: “...which includes the principles of the Declaration of Helsinki, hereunder oral as well as written informed consent.”

9. Eyeballing the scores in Figure 1 I am surprised that the CCC values were so high. Eg it is hard to imagine a CCC value of 0.89 for the data for reposition error shown in the figure. Can you double check the analyses. Please provide the scores for examiner A and B for each of the 5 tests so that I can also double check the analyses.

**Answer:** Nevertheless, it is a fact that, CCC and ICC can disguise fairly large disagreement. However, we have made the graphs uniform by using the same scale for the test, e.g. the y-axis is in cm. Looking closely at these, it is obvious that what appears as large differences as a matter of fact are really small. We do not find it appropriate to send you the data for a re-check. We have sought statistical advice and the analyses have been checked by a statistician for errors and none were found. We will however, be happy to send you log-files from the analyses on request.

11. Please explain what you did with the results of tests that were replicated eg leg lowering was repeated five times and it is unclear if you took the mean or the best score. Either approach has been used in other studies so we need this clarified.

**Answer:** We took the mean of the 5 repetitions (3 repetitions for the repositioning). The analyses of reproducibility are based on the mean values. This has been clarified in the section of the text “Test for lumbar motion control”

14. In table 2 add a third column so that you can provide descriptive data separately for those with and without LBP. Please include the mean (SD) scores for each measure.’

**Answer:** The columns have now been divided into two columns including data separately for subjects with and without LBP. This is clarified in the study sample section.
15. Please present the reliability statistics in Table 3 to two decimal places. Provide the results for LBP participants separately to those who are pain-free.

Answer. This has been done (see Table 3).

Please provide a sharper figure for Fig 1. It is difficult to read. Please use the same name for each test in the Fig and table. Ex A and Ex B is unclear, please say test 1 and test 2.

Answer. This error has been corrected. Figure 1 will be uploaded in a format that is re-scalable. We have kept the wording Examiner A and Examiner B to avoid its being misunderstood as first and second test.

Reference


Inter-examiner Reproducibility of tests for lumbar motor control

Flemming Enoch\textsuperscript{a}, Per Kjaer\textsuperscript{b}, Arne Elkjaer\textsuperscript{c}, Lars Remvig\textsuperscript{a}, Birgit Juul-Kristensen\textsuperscript{d}

\textsuperscript{a}Department of Rheumatology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
\textsuperscript{b}Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark
\textsuperscript{c}Fredensborg Health Centre, Fredensborg, Denmark
\textsuperscript{d}Research Unit for Musculoskeletal Function and Physiotherapy, University of Southern Denmark, Odense, Denmark

**Corresponding author:**

✉ Flemming Enoch: flemming@fysiocenter.dk

Specialist in musculoskeletal physiotherapy MScR

Department of Rheumatology
Rigshospitalet,
Blegdamsvej 9,
2100 Copenhagen Ø
Tel: +4526258271  Fax: +4535452008
Abstract

Background: Many studies show a relation between reduced lumbar motor control (LMC) and low back pain (LBP). However, test circumstances vary and during test performance, subjects may change position. In other words, the reliability – i.e. reproducibility and validity – of tests for LMC should be based on quantitative data. This has not been considered before are unknown. The aim was to analyse the test the reproducibility of five different quantitative tests for LMC commonly used in daily clinical practice.

Methods: The five tests for LMC were: joint position sense repositioning (JPSRPS), sitting forward lean (SFL), sitting knee extension (SKE), and bent knee fall out (BKFO), all measured in cm, and leg lowering (LL), measured in mm Hg. A total of 40 subjects (14 males, 26 females) 25 with and 15 without LBP, with a mean age of 46.5 years (SD 14.8), were examined independently and in random order by two examiners on the same day. LBP subjects were recruited from patients from three different private physiotherapy clinics, who had had a connection to the gym in the clinic or to back-school classes. Non-LBP subjects were recruited from the clinic’s staff acquaintances, as well as from patients without a back pain problem. Reproducibility was determined by way of calculation of concordance correlation coefficients (CCC) intra class correlation (ICC) and Bland and Altman plots to establish limits of agreement.

Results: Four of the All five tests for LMC had reproducibility with the following CCC’s ICCs: 0.90 for RPS, 0.96 for SFL, 0.965 for SKE, 0.944 for BKFO, and 0.98 for LL. The JPS test had good reproducibility with a CCC of 0.89. Bland and Altman plots showed that most of the differences between examiners A and B were less than 0.2 cm.

Conclusion: These five tests for LMC displayed good (JPS) to excellent (SFL, SKE, BKFO, LL) reproducibility. The tests are easy to use and time efficient in daily clinical practice. However, the diagnostic accuracy of these tests needs to be addressed in larger cohorts of subjects, establishing values for the normal population. Also cut-points between subjects with and without LBP must be determined, taking into account age, level of activity, degree of impairment and participation in sports. Further, establishment of the reproducibility of these tests in daily clinical practice must be performed.
Whether reproducibility of these tests is as good in daily clinical practice when used by untrained examiners also needs to be examined.

Background

Pain in the lumbar region is a common problem, corresponding to a point prevalence of approximately 15-27% of all adults [1, 2]. It is estimated that 60 to 80% of the Danish population will experience low back pain (LBP) sometime during their lifetime [3]. The vast majority of these LBP episodes will settle within two to three months, however more than 70% of those with non-treated LBP will have a recurrence within a year [4, 5]. It may be, that problems for the individual patient are cumulative with each episode of LBP [6, 7]. It is disturbing concerning that about 10% of the people having an episode of LBP will develop a chronic pain condition and related disability [8]. Half a year after the first episode of LBP, more than 60% still have pain, and 16% will still be on sick leave [8-10].

There is almost no consensus among different professional groups with regard to examination and treatment methods for patients with low back pain [11]. The lack of a specific diagnosis for the majority of chronic LBP patients has led to the development of many alternative diagnostic assessment processes.

Of increasing interest in recent years has been the assessment of static and dynamic motor control of the lumbo-pelvic complex in LBP, called lumbar motor control (LMC). Various methods of LMC evaluation are currently applied clinically for diagnostic purposes, as part of the physiotherapy examination [5, 12-19]. In this study, the evaluation of LMC included tests regarding the ability to control and reposition the lumbo-pelvic complex, when challenged in different directions. The importance of joint stabilisation in its neutral zone has been demonstrated [20, 21], and inter-segmental instability and altered recruitment of the stabilising muscles have been proposed as possible contributing factors to the development of LBP [12, 13, 16, 22-24].

An optimum static and dynamic stability of the lumbo-pelvic complex, as an expression of the LMC, is considered important in order to maintain the functional and structural integrity of the lumbar region. Deficits in dynamic stability can compromise segmental spinal stability and may lead to tissue damage,
and the development of chronic LBP [24-26]. In particular, the dynamic stability of the lumbo-pelvic complex can be biomechanically challenged by both trunk and limb movements. Appropriate muscle coordination is considered important for the function of the lumbar spine as an effective ‘force-bridge’ between the trunk, the lower and the upper extremities, as well as for force development within the lumbar region itself. The complex anatomy of the lumbo-pelvic region and the multidirectional functional demands placed on it, constitute a challenge for those responsible for determining a specific structural diagnosis. A diagnosis based on movement control impairment is considered by many authors to be a relevant way to subgroup low back pain patients [13, 17, 18].

In a number of studies, several tests for LMC and movement control impairment have been evaluated for their reliability and reproducibility of several tests for LMC, and movement impairment has been evaluated in a number of studies [18, 19, 27-30]. In essence, all these studies report test reproducibility, which the reported reproducibility ranges from poor to almost perfect, apparently depending on the qualifications of the examiners, the focus of the test (symptoms or alignment/movement), and the number of possible subcategories. Clinically, it is difficult without any technical equipment to visually estimate how much the lumbar region is moving during tests for LMC. Previous studies have judged LMC tests dichotomously, as “can–cannot/yes-no”. However, a lot of information is hidden between these two end-points. Besides, there has been no clear consensus for when the test is passed/not passed, or at what level the relevant dichotomous cut-point of each test should be. Consequently, there is a need for more precise test descriptions, in addition to tests with more quantitative and reproducible methods for measuring LMC.

A test battery consisting of five tests, described in several articles and textbooks [12-14, 19, 31-34], is often used in daily clinical practice. The tests have evolved and been modified over the past ten years, including a method for continuous quantification. This has been done in order to achieve clear standards for quantifying LMC as the tests challenge LMC in three directions: flexion, extension and rotation. Information from these tests contributes to making a directionally specific diagnosis, which Based on the results from these tests, it should make it possible to design a retraining program and provide more specific advice on appropriate physical activity, including measurement of the effect on
LMC. However, in order to achieve clear standards for quantifying LMC, these tests have evolved and been modified over the past ten years, including a method for continuous quantification. However, the reproducibility of the tests still needs to be determined. Therefore, the aim of the current study was to test the inter-examiner reproducibility of these tests for LMC in a mixed population of subjects with and without LBP.

Methods

Definitions

A reliable test is a test that ‘can be depended upon with confident certainty’ (that is, it is trustworthy)\[35]\, meaning that it can be reproducible as well as valid. A reproducible test is a test where one can achieve the same result from two or more different measurements. A valid test is a test in which the result really expresses the characteristics the test is expected to express.

Study design

The study was a test-retest reproducibility study with two examiners, who followed a three-phase reproducibility and validity study protocol, recommended by the International Academy of Manual/Musculoskeletal Medicine (IAMMM) \[30]\. Since this study included continuous data, the protocol was adjusted accordingly to a two-phase study, and excluded the overall agreement phase.

In phase one, the five tests (see Table 1), were described in detail by the two examiners A and B (FE and AE). They were both teachers in the Danish Manual Therapy Society, and they had both had 20 years of clinical experience, including experience in using these tests for LMC. Both examiners tested 10 subjects with LBP in an open study, in order to become familiar with the test procedures and the method for interpreting test results, thereby reducing examiner bias.

In phase two, the two examiners applied the five tests for LMC on all subjects (40 subjects, approximately 75\% of whom had LBP on the day of the examination, see Table 2) in two separate rooms. Each examiner provided the subject with the necessary instruction for the tests, and all subjects were appropriately unrobed to allow visualisation of the lumbosacral spine. The subjects were
examined independently and in random order by two examiners on the same day, and after examiner A had tested a subject, the subject was examined by examiner B, and vice versa. Half the subjects started with examiner A and half with examiner B. Both examiners performed the tests in the same order on each subject, specified in the current manuscript (Test 1-5) in the section ‘Tests for lumbar motor control’.

Healthy controls were included in order to maximise variability in the subjects’ test performance, partly to reduce examiner bias and partly to cover the spectrum of all possible measurement levels available for the tests.

The Regional Committee on Biomedical and Research Ethics approved the study (H-A-2008-082), which includes the principles of the Declaration of Helsinki. All participants gave, hereunder their consent after receiving oral as well as written information about the study and consent.

Study sample
The LBP subjects were recruited from patients seeking care from three different private physiotherapy clinics who had had a connection to the clinic gym or to back school classes, from three different private physiotherapy clinics. The non-LBP subjects were recruited from the clinical staff’s acquaintances, as well as from patients without back pain problems.

The inclusion criteria were men and women, aged 18-85 years, with (25 subjects) or without (15 subjects) non-specific LBP problems (see Table 2), while the exclusion criteria were neurological or rheumatologic disorders, acute pain in the hip and leg, diabetes and cancer, and inability to speak and understand Danish. The Numeric Pain Rating Scale (NPRS), previously shown to be valid [36], was used to describe the severity of the LBP.

Tests for lumbar motor control (LMC)
Five different tests for LMC were used, including one for joint position sense repositioning (RPS) and four for dynamic stability, including sitting forward lean (SFL), sitting knee extension (SKE), bent knee fall out (BKFO) and leg lowering (LL) (Table 1). Generally, the subjects performed a maximum
of 10 repetitions of each test. The subjects were allowed to practise the JPSRPS test twice, and the remaining four dynamic stability tests a maximum of five times, before the test examination started. Thereafter, three repetitions of the RPS test and five repetitions of the other tests were performed, and the mean value of these was calculated. Within this range, the amount of instruction and tactile feedback before the test evaluation started varied among subjects, depending on the subject’s ability to understand and perform the tests. The tests are summarised in Table 1.

**Joint position Sense 1) Repositioning (JPSRPS)** was performed by measuring how accurately the subject during sitting could re-position the low back (LB) into the former lumbar position, after having actively moved around, in flexion and extension. The subject was sitting with feet supported, and the examiner guided the subject's LB into neutral position. The examiner ensured that the LB was in neutral position, i.e. midway between the posterior and anterior tilt. A 5 cm tape-measure with mm markings was placed on the LB with the 0 cm marking on Sacral segment 1 (S1). A laser pointer (Class 3A Laser product, Wen Zhou Xinke, China), placed on a stable base and adjusted to be level, was positioned to have the mark line directly on 0 cm. The subject was instructed to remember this position, and then to move the pelvis twice from the maximum anterior to the maximum posterior tilt and then return to the neutral position. With the laser line on the tape-measure, the deviation from the 0 point was measured in cm, and this could be read within ±0.25 mm accuracy. The test was performed three times.

**2) Sitting Forward Lean (SFL)** was designed to measure the amount of LB movement that was necessary for a sitting forward leaning movement of the upper body. The range of motion (ROM) of the LB was measured by a 15 cm ruler. The subject was sitting upright with the knees and the hips at 90°, and with the hands resting on the thighs. The examiner placed the subject’s LB in neutral position and marked the SI point and a point 10 cm cranially, using a pen on the skin. The subject was instructed to hold that position of the two points relative to each other, during the subsequent movements. To guide the range of movement, the examiner firmly grasped the subject's pelvis and moved the pelvis anteriorly, until a maximum of 120° hip flexion was reached, measured by a
plurimeter V gravity inclinometer (Access Health, Melbourne, Australia), placed on the LB. The examiner placed the LB in neutral position, i.e. midway between the posterior and anterior tilt. The subject was then instructed to remain in neutral position of the LB, while moving the trunk and pelvis forward until the hips reached 120° flexion or within the available ROM. Initially, the first test performance movement was guided with tactile feedback, by the examiner’s 1st and 2nd finger on the S1 and 10 cm mark on the tape-measure. Once the subject was well instructed, the examination of the test started, and the subject performed five repetitions without feedback. At the forward lean position of each repetition, the distance between S1 and 10 cm mark was measured in cm with a ruler, to within one decimal point.

3) Sitting knee extension (SKE) was designed to determine the magnitude of LB movements that occurred during a sitting knee extension, using a tape-measure (in cm). Using the same setup as with JPSRPS, the couch was raised until the subject’s feet were off the floor. In order to define ROM at the knee during the test, the examiner manually fixed the subject's pelvis in neutral with one hand, and extended the knees as much as possible, however only to a maximum of minus 10° extension. This was controlled using the plurimeter, placed at the tibia just distal to the tibial tuberosity. A 5 cm tape-measure was placed on the LB with 0 cm at S1, and with the laser pointing at 0 cm. The examiner ensured that the LB was in neutral position, i.e. midway between the posterior and anterior tilt. The subject was instructed to remain in a neutral position of the LB, while moving the knee to minus 10° extension or within the available ROM. Initially, the movement was guided with feedback by the examiner’s 1st and 2nd finger placed on the S1 and 10 cm mark, previously marked on the skin cranially to the S1. Once the subject was well instructed, the test started, and the subject performed five repetitions without feedback. At the end of each knee extension, the LB movement was measured as the distance from 0 cm to the laser pointer mark.

4) Bent Knee Fall Out (BKFO) was designed to evaluate the range of LB movement that takes place during a supine lying external rotation of the hip, using a tape-measure (in cm). The subject was supine lying with right hip flexed, the knee flexed at 120°, with the feet resting on the surface of the couch,
and the arms lying relaxed beside the body. The examiner ensured that the LB was in neutral position, i.e. midway between the posterior and anterior tilt. A 5 cm tape-measure was placed laterally to the anterior superior iliac spine (ASIS opposite to the bent leg) with 0 cm placed on the ASIS and pointing laterally towards the laser. The laser line was adjusted to the 0 cm point to determine the amount of hip movement on the tape-measure. The examiner manually fixed the subject's pelvis and moved the hip of the bent leg into as much abduction/external rotation as possible, however, only to a maximum of 45°, measured by the plurimeter, placed at the medial side of the knee. The subject was instructed to abduct the knee to the determined point and return to the starting position, and in the beginning, the subject received feedback via the examiner’s finger on the ASIS in order to detect the movement. Once the subject was well instructed, the test started, and the subject performed five repetitions without feedback. At the extreme of hip abduction in each repetition, the LB movement was measured as the distance from the 0 cm on the tape-measure to the laser pointer mark.

5) **Leg Lowering (LL)** was designed to quantify the extent of LB movement accompanying a supine lying unilateral leg lowering using a pressure biofeedback unit (PBU) (Chattanooga Ltd Hixson, USA). The PBU instrument was developed to monitor LMC by recording pressure changes in mm Hg during the different repetitions. The PBU has been shown to be reliable and capable of detecting even small changes in pressure during movement [37]. The subject was placed in supine position with the hips at 90° flexion. The knees were in maximum relaxed flexion. The examiner ensured that the LB was in neutral position, i.e. midway between the posterior and anterior tilt, and the ASIS were at a horizontal level. The arms were relaxed and beside the body. A BPU was placed under the LB, and inflated to 40 mm Hg. First, the subject was asked to actively push the LB downwards, increasing the BPU pressure to 45 mm Hg. Then the subject was instructed to lower the feet to just above the surface of the couch. In the early attempts, the subject was allowed to have visual feedback from the BPU. Once the subject was well instructed, five repetitions were performed without feedback. At each repetition, the pressure in mm Hg was recorded, when the feet were as close as possible to the couch.
Statistical analyses
To evaluate the inter-examiner reproducibility of test performance, intraclass correlation coefficients (ICC) type 2.1 [38, 39] the concordance correlation coefficient (CCC)-[40, 41]-and Bland and Altman’s [42] limits of agreement (LOA) were used (Figure 1). A correlation coefficient above 0.90 is considered ‘excellent’ reproducibility, greater than 0.75 is considered ‘good’ reproducibility, and less than 0.75 indicates ‘poor’ reproducibility [43]. The ICC was calculated for all subjects as a group, and separately for those with and without LBP on the day of the examination.

LOA is based on the difference between results from examiners A and B. The average of the differences in measures from examiner A and B is reported (Table 3), together with the standard deviation and the range within 95% of the differences (95% LOA). Data are presented for the groups with and without LBP on the day of the examination, separately, in addition to the whole group. Bland and Altman plots were constructed by plotting the differences between A and B measures (y-axis) against the mean of A and B (x-axis) for each of the tests, as shown in Figure 1. The green line (y=0) is perfect average difference and the purple line the observed average difference. The distance between these lines indicates the bias towards one of the observers’ measures. The distance from the purple line (average difference) to each dot represents the difference between the examiners rating for the observed mean value for the two examiners on the x-axis. The red lines in the figures indicate 95% LOA as described above. The closer the dots are to the line, the less disagreement in measures. People with and without LBP on the day of the examination are marked as orange and blue symbols, respectively. The LOA provides a mean difference between measurements of the same parameter, which indicates whether or not a systematic bias is present (i.e. deviations from 0). This statistical method also provides upper and lower limits, indicating where most of the measurements exist [95% confidence interval (CI)]. For statistical analyses, the STATA statistical package was used (Stata Corp., 2000, Stata Statistical Software: Release 11.1, College Station, TX). The command “icc23” (two way ANOVA) was used to calculate ICC type 2.1 with 95% confidence intervals (CI), and the command “concord” was used to calculate LOA as well as Bland and Altman plots.
Results
In total, 40 subjects were recruited for this reproducibility study, of whom 14 were men and 26 were women, having an age range from 20 to 82 years. The mean age of the subjects was 46.5 years (SD 14.8) (Table 2), and 15 (37.5%) of them did not have pain on the test day. Pain intensity, measured by the NPRS on the test day, ranged from 0 to 8. In total, nine subjects (22.5%) had never had backache for more than three days. In contrast, 19 subjects (47.5%) had had more than 10 episodes of LBP that lasted more than three days.

All of the tests (SFL, SKE, BKFO, LL, and RPS) had excellent inter-examiner reproducibility (ICC >0.930) for the whole group, except for while one test (JPS) which had good reproducibility with a ICC of 0.890 (Table 3). The Bland and Altman plots showed that the majority of the differences were less than 0.2 cm for the whole group (Figure 1). From the limits of agreement 95% of the measurements’ variation is within the range of -0.44cm to 0.35cm for tests 1-4, while for test 5 (LL) the range is from -3 to 3mmHg, representing the absolute measurement differences in relation to the mean of the measurements. In the Bland and Altman plots, most of the measurements are located within a smaller range (Figure 1).

When analysing groups with and without LBP on the day of the examination separately, the ICC values were about the same level as for the whole group.

Discussion
The principal findings were a good to excellent inter-examiner reproducibility of the five tests for LMC, with the ICC ranging from 0.890 to 0.98 for the whole group. The differences between the two examiners were less than 0.2 cm. To our knowledge, this is the first study reporting excellent reproducibility for tests of LMC, using a quantitative method. Previously, the tests have evolved and been refined over a period of time involving several clinicians, via a trial and error method in several different clinical settings. The emphasis has been on reaching a feasible and accurate test battery, which clinicians can easily use and agree upon. The results of this study show promise for reaching this goal, at least when the test battery is applied and
all five studies on reproducibility of LMC tests have been tested in a dichotomous setup with qualitative data. Three other studies have shown substantial and almost perfect inter-tester reproducibility in qualitative rating of similar LMC tests [18, 19, 27]. In two of the studies, “reproducibility of specific classification systems for motor control impairments was tested” in which kappa was 0.96 for experienced and 0.61 for inexperienced clinicians [18], respectively and 0.75 for experienced clinicians [27]. However, both these studies analysed the diagnostic reproducibility based on a whole battery of tests, and thus they are not comparable with each individual test in the current study. The third study, which tested reproducibility of individual tests for LMC, rated dichotomously, have shown kappa values of 0.72 for SKE, and 0.38 for BKFO [19], both of which were not as reproducible as in the current study (0.95 and 0.94). Further, only six out of the ten tests for LMC were classified as having substantial reliability with kappa > 0.60 [19]. Of these ten tests, only SKE and BKFO are comparable with the current tests (same position, same test procedure, rating deficits in same direction, although dichotomously). A similar study of LMC tests showed substantial kappa values for hip extension with 0.72 and 0.76 (left and right) for 80% of the cases, but neither test was included in the current study [29].

Inter-examiner reproducibility of 53 different tests showed kappa to be ≥0.75 for tests related to symptoms, but when related to alignment and movement, kappa was only ≥0.41 [28]. Two of these tests (SKE, BKFO) were similar to the current tests with kappa of 0.58 and 0.52; however, test results in the study by Van Dillen et al. [28] were rated dichotomously (yes/no), and solely rated by visual observations. Since it is well known that ‘judgments based on visual and tactile information are often difficult to make reliable’ [28], use of a visual rating method may have been one of the reasons for the poor kappa values.

However, in kappa studies it is essential to secure a high Overall Agreement, and a 50/50% prevalence of positive and negative findings in order to measure the true reproducibility of the test (34). We are not provided with this information in the above-mentioned papers and consequently the true kappa value may be higher than presented by the authors. Likewise, when using dichotomous scales and without using the ‘50/50% contrast group’ design concept, low kappa values may be due to test
responses being close to the threshold values provided for the study, thereby also increasing the risk for a poor kappa.

The strengths of this study are that despite the differences in study design, tests, examiner expertise, and selection criteria for the study population, the results are in line with data from previous studies and show better reproducibility. This may be due to our use of the standardised protocol by IAMMM [30], including a standardised training procedure for the examiners, a protocol with well defined procedures and operational definitions that provided quantifiable values. According to IAMMM, these are the prerequisites for excellent reproducibility of clinical tests [30].

Also, in this protocol, the two examiners went through a training phase in order to minimise bias during performance of the tests and to increase their overall agreement on test performance and interpretation. This precaution will increase the intention of the study to test the test independent of the examiner, and not to test the combination of test and examiner, illustrating that the tests per se are good to excellent. Finally, the study was carried out on LBP and non-LBP subjects, for whom the test battery is intended, making the results relevant for screening purposes within this group. The examiners had many years of clinical experience and use of the tests.

The weakness of the study is that we do not know the reproducibility of the current tests must be determined in studies carried out by inexperienced clinicians, which of course might be different from the reproducibility of experienced and trained examiners, as previous studies have shown experienced examiners to have a higher test reproducibility than inexperienced clinicians [18, 19]. However, the current tests were developed to include only quantifiable variables, and do not include other more subjective factors, such as breathing, co-contraction, rigidity and perceived effort, which are commonly included in daily clinical practice. This absence is likely to have increased the observed reproducibility, but reproducibility also needs to be tested in a more normal clinical environment.

Further, in case inexperienced examiners have a low inter-examiner test reproducibility, our study has shown that it is should be possible through education and training to obtain high enough skills to perform the tests in a reproducible way.

In the current study, we have used CCC [40] as a measure of correlation between the two examiners because of the concerns raised about the intra-class correlation coefficient (ICC) most often used in studies of correlation [44]. CCC is more robust to the order of examinations compared with ICC. Post
hoc ICCs were also calculated, to enable comparison of the current results with other studies. The differences between the two methods (CCC and ICC) of estimating correlation coefficients revealed a difference of less than 0.002, making the results directly comparable.

The use of correlation coefficients for reliability can easily disguise large differences in measurements. Therefore, also the Bland and Altman plots [42] were used, from which the variation in each measure from each examiner is demonstrated. This provides the reader with the true variation, as a supplement to the ICC.

The current tests were developed to include only quantifiable variables, and do not include other more subjective factors, such as breathing, co-contraction, rigidity and perceived effort, which are commonly included in daily clinical practice. This absence is likely to have increased the observed reproducibility, but reproducibility also needs to be tested in a more normal clinical environment.

Testing the inter-examiner reproducibility is a more valid expression of reproducibility than intra-examiner reproducibility, as intra-examiner reproducibility is usually better.

Several aspects need to be considered and analysed in the future: Since an excellent reproducibility of clinical tests for LMC was obtained, the validity as well as the relevant cut-point (distance moved from the 0 point) for abnormality for each of the tests must be determined by testing the human variation in the normal population. Further, the validity, sensitivity and specificity must be tested, i.e. the discriminative ability of the tests to discriminate between subjects with and without LPB in a larger study sample. A recent pilot study showed the predictive validity of a poor performance on two selected LMC tests in relation to an increased risk of lower limb/lumbar spine injuries in professional dancers [12]. Another pilot study (without a control group) including 38 LBP patients showed that after treatment focusing specifically on increasing LMC, the pain decreased, and, physical function and LMC improved [33]. The addition of tests for LMC with excellent reproducibility may also enhance future validity studies, such as those previously described in other positions [12] and other movement directions [29]. Using a whole test battery may make it possible to determine the optimum number and combination of tests with that have the highest diagnostic accuracy (i.e. sensitivity and specificity). The future of LMC tests is challenging, and further studies of these interactions are required.
Conclusion
The current five tests for LMC had a good (JPS) to excellent (SFL, SKE, BKFO, and LL, and RPS) reproducibility. The tests are easy to use and not time-consuming in daily clinical practice. However, reproducibility is only the first step on the path to establishing the diagnostic value of these tests. Therefore, subsequent studies need to include larger cohorts of subjects, including establishment of values for the normal population, and cut-points between subjects with and without LBP, while taking into account age, levels of activity, degree of impairment and participation in sports. Further, establishment of the reproducibility of these tests in normal clinical practice must be performed.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
FL was involved in the planning and acquisition of the data, the making of the videos, the data analysis and the writing of the paper. AE was involved in the planning and acquisition of the data, LR and BJK were involved in the planning, methodological considerations, analysis of the data, and revision of the paper. PK was involved in the data analysis, calculation of the statistics and revision of the paper. All authors read and approved the final manuscript.

Acknowledgements
The authors wish to thank The Danish Society for Rheumatism for supporting this project.
Table 1. Clinical tests for Lumbar Motor Control (LMC). LB=low back, ASIS=Anterior Superior Iliac Spine, BPU=Biopressure Unit.

<table>
<thead>
<tr>
<th>Test name</th>
<th>Subject position, performance and measurement equipment</th>
<th>Modified by previous test (reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Joint position sense (JPSRPS)</td>
<td>Sitting, feet unsupported, LB in neutral, 5 cm tape-measure, taped at 0 cm at S1, and marked by laser pointer. LB movement from max anterior-max posterior tilt. Subject reposition of LB (neutral), distance measured between 0 cm (S1) and laser pointer.</td>
<td>[13, 14]</td>
</tr>
<tr>
<td>2) Sitting Forward lean (SFL)</td>
<td>Sitting, feet supported, LB in neutral, mark with 15 cm ruler at S1 and 10 cm above. 5 repetitions of hip flexion to max 120°, distance between marks (0 cm and 10 cm) measured (cm).</td>
<td>[13, 14] [28]</td>
</tr>
<tr>
<td>3) Sitting knee extension (SKE)</td>
<td>Sitting, feet unsupported, LB in neutral, 5 cm tape-measure, taped at 0 cm at S1, and marked by laser pointer. 5 repetitions in knee extension up to at least -10°, distance measured between 0 cm (S1) and laser pointer.</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>4) Bent knee fall out (BKFO)</td>
<td>Supine lying, one knee flexed 120°, LB and pelvis in neutral. 5 cm tape-measure placed between right and left ASIS, with 0 cm and laser pointer placed lateral to right ASIS. 5 repetitions of abduction/external hip rotation up to max. 45°, distance measured between laser pointer and 0-point (cm).</td>
<td></td>
</tr>
</tbody>
</table>
5) Leg lowering (LL)

Supine lying, hips flexed 90°, knees maximally flexed, LB in neutral.

BPU placed under LB, inflated to 40 mmHg. LB downward press to increase BPU to 45 mmHg. 5 repetitions of leg lowering. Increase measured in BPU (mmHg).

[31]
[13, 14]
Table 2. Demographics obtained by questionnaires and NPRS=Numeric Pain Rating Scale.

<table>
<thead>
<tr>
<th>Variable</th>
<th>People with LBP on day of examination (n=25(62.5%))</th>
<th>People without LBP on day of examination (n=15(37.5%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years (mean (SD))</td>
<td>47 (12)</td>
<td>45 (19)</td>
</tr>
<tr>
<td>Gender (n(%))</td>
<td>11/14</td>
<td>3/12</td>
</tr>
<tr>
<td>(Male/female)</td>
<td>(44/56%)</td>
<td>(20/80%)</td>
</tr>
<tr>
<td>Previous episodes (n(%))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>9 (60%)</td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>2 (8%)</td>
<td>1 (7%)</td>
</tr>
<tr>
<td>Less than 5</td>
<td>1 (4%)</td>
<td>2 (13%)</td>
</tr>
<tr>
<td>5 or more</td>
<td>4 (16%)</td>
<td>2 (13%)</td>
</tr>
<tr>
<td>more than 10</td>
<td>18 (72%)</td>
<td>1 (7%)</td>
</tr>
<tr>
<td>Pain on day of examination* (n(%))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no pain</td>
<td>15 (100%)</td>
<td></td>
</tr>
<tr>
<td>1-3</td>
<td>13 (52%)</td>
<td></td>
</tr>
<tr>
<td>&gt;3</td>
<td>12 (48%)</td>
<td></td>
</tr>
<tr>
<td>Back History in months (n(%))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>2 (8%)</td>
<td>12 (80%)</td>
</tr>
<tr>
<td>4-7</td>
<td>2 (8%)</td>
<td>2 (13%)</td>
</tr>
<tr>
<td>8-12</td>
<td>13 (52%)</td>
<td>1 (7%)</td>
</tr>
<tr>
<td>13-16</td>
<td>8 (32%)</td>
<td></td>
</tr>
</tbody>
</table>

* Numeric Pain Rating Scale (NPRS) 0-10
Table 3. Intraclass correlation coefficient (ICC type 2.1), with 95 percent confidence intervals (95% CI) for the five lumbar motor control tests. Average refers to the mean difference between the two examiners A and B in cm (for Leg Lowering in mmHg) with standard deviation (SD). The 95% Limits Of Agreement (LOA) express the interval in which 95% of the differences in measures from examiner A and B lie. Values are given for all people and for the subsamples with and without low back pain (LBP) on day of examination.

<table>
<thead>
<tr>
<th>Test</th>
<th>ICC</th>
<th>[95% CI]</th>
<th>Difference</th>
<th>95% LOA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Repositioning (all)</td>
<td>0.90</td>
<td>[0.81; 0.94]</td>
<td>0.02 (0.12)</td>
<td>[-0.22; 0.26]</td>
</tr>
<tr>
<td>-&gt; LBP = 0</td>
<td>0.85</td>
<td>[0.60; 0.95]</td>
<td>0.06 (0.13)</td>
<td>[-0.19; 0.31]</td>
</tr>
<tr>
<td>-&gt; LBP = 1</td>
<td>0.92</td>
<td>[0.82; 0.96]</td>
<td>0.00 (0.12)</td>
<td>[-0.23; 0.22]</td>
</tr>
<tr>
<td>2) Sitting Forward Lean (all)</td>
<td>0.96</td>
<td>[0.92; 0.98]</td>
<td>-0.06 (0.18)</td>
<td>[-0.41; 0.29]</td>
</tr>
<tr>
<td>-&gt; LBP = 0</td>
<td>0.94</td>
<td>[0.82; 0.98]</td>
<td>-0.08 (0.18)</td>
<td>[-0.43; 0.28]</td>
</tr>
<tr>
<td>-&gt; LBP = 1</td>
<td>0.96</td>
<td>[0.92; 0.98]</td>
<td>-0.05 (0.18)</td>
<td>[-0.40; 0.30]</td>
</tr>
<tr>
<td>3) Sitting Knee Extension (all)</td>
<td>0.95</td>
<td>[0.90; 0.97]</td>
<td>-0.01 (0.10)</td>
<td>[-0.20; 0.18]</td>
</tr>
<tr>
<td>-&gt; LBP = 0</td>
<td>0.94</td>
<td>[0.84; 0.98]</td>
<td>0.00 (0.08)</td>
<td>[-0.17; 0.16]</td>
</tr>
<tr>
<td>-&gt; LBP = 1</td>
<td>0.95</td>
<td>[0.88; 0.98]</td>
<td>-0.02 (0.11)</td>
<td>[-0.22; 0.19]</td>
</tr>
<tr>
<td>4) Bend Kne Fall Out (all)</td>
<td>0.94</td>
<td>[0.88; 0.97]</td>
<td>-0.03 (0.19)</td>
<td>[-0.40; 0.34]</td>
</tr>
<tr>
<td>-&gt; LBP = 0</td>
<td>0.97</td>
<td>[0.92; 0.99]</td>
<td>0.01 (0.15)</td>
<td>[-0.29; 0.31]</td>
</tr>
<tr>
<td>-&gt; LBP = 1</td>
<td>0.89</td>
<td>[0.77; 0.95]</td>
<td>-0.06 (0.21)</td>
<td>[-0.46; 0.35]</td>
</tr>
<tr>
<td>5) Leg Lowering (all)</td>
<td>0.98</td>
<td>[0.96; 0.99]</td>
<td>-0.17 (1.49)</td>
<td>[-3.08; 2.74]</td>
</tr>
<tr>
<td>-&gt; LBP = 0</td>
<td>0.98</td>
<td>[0.92; 0.99]</td>
<td>-0.75 (1.43)</td>
<td>[-3.54; 2.05]</td>
</tr>
<tr>
<td>-&gt; LBP = 1</td>
<td>0.98</td>
<td>[0.96; 0.99]</td>
<td>0.18 (1.44)</td>
<td>[-2.65; 3.00]</td>
</tr>
</tbody>
</table>
Legends to figures

Figure 1.
Bland and Altman plots with limits of agreement of the five tests, n=38 for SKE, n=40 for the remaining tests (LL in mm Hg, remaining tests in cm). A and B represent the examiners named A and B. Blue dots represent subjects without LBP, while red dots represent subjects with LBP on the day of examination.
Figure 1.
References


27. Harris-Hayes M, Van Dillen LR: **The inter-tester reliability of physical therapists classifying low back pain problems based on the movement system impairment classification system.** *PM R* 2009, 1:117-126.


32. Luomajoki H, Kool J, de Bruin ED, Airaksinen O: Movement control tests of the low back; evaluation of the difference between patients with low back pain and healthy controls. *BMC Musculoskelet Disord* 2008, **9**:170.


