# The inter-rater reliability of clinical tests that best predict the subclassification of lumbar segmental instability: structural, functional and combined instability

**Faisal M. Alyazedi[1], Everett B. Lohman[2], R. Wesley Swen[2], Khaled Bahjri[3]**

[1]Prince Sultan Military Medical City, Kingdom of Saudi Arabia, [2]School of Allied Health Professions, Loma Linda University, USA, [3]School of Public Health, Loma Linda University, USA

Objectives: This study investigated the inter-rater reliability of three structural end range lumbar segmental instability tests with the highest positive likelihood ratio ( + LR) against flexion–extension radiographs, and three functional mid-range clinical tests that predict the success of lumbar stabilisation exercises in patients with recurrent or chronic low-back pain (R/CLBP). The study also investigated the reliability of lumbar segmental instability, subclassification as: functional, structural and combined instability.
Method: Forty adults with R/CLBP (30 men and 10 women), aged 21–71 years, underwent repeated measurements of specific clinical tests for structural or functional lumbar segmental instability.
Results: All functional-instability tests: the prone instability test (PIT), the aberrant motion test and the average passive straight-leg raise (PSLR > 91°) test showed a high percentage agreement (90, 97·5 and 95%, respectively) and a high kappa coefficient (0·71, 0·79 and 0·77, respectively). In addition, two structural tests: the lumbar flexion range of motion (ROM) >53° and the passive lumbar extension test (PLET) showed a high percentage agreement (82 and 73%, respectively), and a moderate kappa coefficient (0·48 and 0·46, respectively). The lack of hypomobility with the posteroanterior (PA) glide test was found to be unreliable (agreement=25%; $k=-0{\cdot}02$). Locating the pain-provoking segment, as the first portion of PIT, was found to be moderately reliable ($k=0{\cdot}41$). The subclassification categories of lumbar segmental instability (functional, structural and combined) were found to be significantly reliable (PABAK) 0·90, 0·70 and 0·95, respectively).
Discussion: All investigated tests (except the lack of hypomobility with the PA glide test), in addition to subclassifying the categories of lumbar segmental instability, were significantly reliable in the assessment of lumbar instability.

**Keywords:** Clinical prediction rule, Low-back pain, Physical examination, Reliability, Segmental instability

## Introduction

Low-back pain (LBP) is common, affecting up to 80% of the population in their lifetime,[1–3] with a 12-month recurrence rate ranging from 66 to 84%.[4] Previous researches show that one of the main causes of frequently recurring LBP is lumbar segmental instability.[1,5,6] During daily living activities, two spinal subsystems are chiefly responsible for controlling and preventing excessive motion between spinal segments: the neuromotor control subsystem and the osseoligamentous subsystem. The neuromotor control subsystem controls segmental motion during the mid-range. The osseoligamentous subsystem limits segmental motion at the extremes of lumbar motion.[6–8] The loss of neuromotor capability to control segmental movement during mid-range is defined as functional instability,[7] whereas the disruption of passive stabilisers, which limit the excessive segmental end range of motion (ROM), is defined as structural instability.[7,9] Panjabi[10,11] suggests that a loss of integrity within the passive subsystem may make segments unstable unless the neuromuscular subsystem compensates for that loss. The estimated prevalence of LBP due to lumbar segmental instability is about 33% for patients with functional instability,[12] compared to 57% for patients with evidence of structural instability, as indicated by positive flexion–extension radiograph.[13]

In 1944, Knutsson[9] recommended the use of flexion–extension radiographs to identify and quantify abnormal anterior–posterior translation of the segment at the end range of spinal flexion and extension.

Correspondence to: Faisal M. Alyazedi, Loma Linda University, USA. Email: faisalyasedi@hotmail.com


DOI 10.1179/2042618615Y.0000000002

This imaging modality has become the diagnostic standard for structural lumbar segmental instability.[7,8,14,15] However, there is no diagnostic standard to quantify the functional instability around the neutral position.[6,8,15] A number of studies have attributed this functional instability to the lack of neuromuscular control of the joint during activities of daily living.[6–8,10] Hicks *et al.*[12] studied the clinical tests that might predict the success of stabilisation exercises that have been developed to improve spinal motor control (stiffness) around the spinal neutral position. They came up with four predictors that together form the clinical prediction rule (CPR) for lumbar stabilisation exercise.[12] However, the reliability of the CPR in patients with recurrent or chronic low-back pain (R/CLBP) has yet to be established.

The purposes of this study were (1) to examine the inter-rater reliability of the six most valid structural and functional lumbar segmental instability tests, defined by the highest positive likelihood ratio ( + LR) in the literature; (2) to explore the inter-rater reliability of the segmental instability subclassification (functional, structural and combined instability) for those who suffer R/CLBP (Fig. 1).

## Methods

### *Study participants*

Forty adults (mean age 35 ± 11·3 years) who had R/CLBP were recruited from the San Bernardino community between February and June of 2013. Some of them were chronic patients, who had taken part in other LBP studies at the Loma Linda University research laboratory. Other 'convenient samples' observed the study poster that had been placed around the University campus and contacted the data collector through the phone number provided on the poster. Demographic information is illustrated in Table 1.

### *Inclusion and exclusion criteria*

The inclusion criteria for this study consisted of participants that had (1) a new episode of LBP; (2) experienced a similar episode of LBP before, with the first episode of back pain occurring at least 3 months before the date of recruitment; (3) currently experienced persistent LBP for at least 3 months.[16]

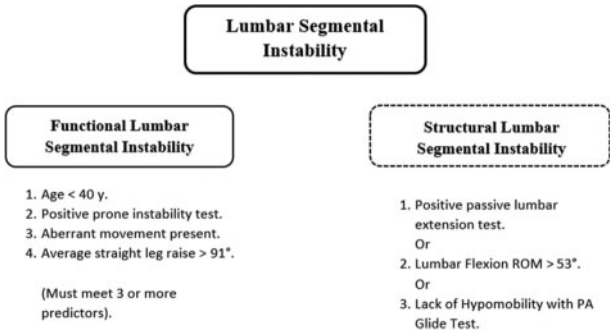


**Figure 1 Main lumbar segmental instability categories.**

**Table 1 Demographic information**

| Variables | Outcomes |
|---|---|
| Age (years) | |
| Mean | 35 (± 12.22)* |
| Range | 21–71 |
| Gender | |
| Male | 30 |
| Female | 10 |
| Modified ODI score | |
| Minimal disability (number, range) | (28) 0–20% |
| Moderate disability (number, range) | (12) 21–40% |
| FABQ score range | |
| Physical activity (median, range) | 3.50 (0.25–5.25) |
| Chronic LBP (number of participants) | 6 |
| Recurrent LPB (number of participants) | 34 |

ODI: Oswestry Disability Index; FABQ: Fear-Avoidance Belief Questionnaires; LBP: low-back pain.
* Value represent mean (standard deviation).

The participants were excluded if they had (1) undergone previous spinal-fusion surgery; (2) a history of traumatic fracture of the spine that resulted in a permanent neurological deficit; (3) scoliosis greater than 20°; (4) pregnancy; (5) inability to actively flex and extend the spine adequately to permit an assessment of segmental motion due to pain or muscle spasm; and (6) medical 'red flags', such as caudaequina syndrome, tumour and systemic inflammatory conditions.

### *Ethical issues*

Participation in this study was voluntary. Informed consent was obtained from all participants. This study was approved by the University Institutional Review Board (Clinical Trial Registration: ISRCTN18037677).

### *Examiners*

This study was performed by three physical therapists, who received 30 minutes of training regarding all written and clinical procedures of the study. The training included performing each of the tests on each other and filling in the clinical test forms indicating positive or negative test results. Information about the test's prescription and interpretation was given to the examiners 3 weeks before the training day (Table 2). One of the examiners [a data collector with 13 years of musculoskeletal clinical experience and who held a Doctor of Science (D.Sc.) degree in physical therapy] recorded the participants' baseline data. The role of the remaining two examiners was to examine, interpret and record the various clinical test results of each participant. One of the examiners was a D.Sc. and the other was a Doctor of Physical Therapy (DPT); both were certified as Orthopaedic Clinical Specialists (OCSs) with more than 20 years of musculoskeletal clinical experience.

### *Data collection*

The first examiner collected all baseline data consisting of informed consent, demographic information, self-reported history and self-reported outcome measures. Subsequently, the two clinical examiners, blinded to

**Table 2 Lumbar segmental instability tests**

| | |
|---|---|
| 1) Aberrant motion[13,15] | If any of these movements are observed during forward bending – such as painful arc of motion, an instability catch, thigh climbing or a reversal of lumbopelvic rhythm – the test is considered positive. |
| 2) PIT test[13,15] | The participant lies in prone position on the edge of the examining table with the feet on the floor. Examiner performs PA mobility testing on each lumbar segment while the participant's trunk muscles are relaxed (relaxation phase); if a painful segment is identified, the participant is asked to lift the legs slightly off the floor (co-contraction phase). Then, the examiner applies the same amount of pressure to the painful segment. If pain is provoked at the relaxation phase and subsides at the co-contraction phase, the test is considered positive. |
| 3) Average PSLR $>91°$ test[12,15] | From supine position, the bubble inclinometer is positioned at the tibial crest. The leg is then passively raised to the maximum tolerated level; then the ROM degree is recorded, and the examiner repeats the same process on the second leg. If the average reading of both legs is $>91°$, the tests are considered positive. |
| 4) Lumbar flexion ROM $>53$[2,13] | From standing position, the bubble inclinometer is used to record the baseline reading of T12-L1 and S2 reference point. Then, after the participant has bent forward, the end range of T12-L1 is recorded; then, the S2 reading is recorded. The true lumbar range is a result of the subtraction of sacral ROM from thoracolumbar ROM. If the result is $>53°$, the test is considered positive. |
| 5) PLE test[8,21] | With the participant in prone position, both legs are passively raised about 30 cm from bed level and then pulled gently. If the subject experiences severe LBP, or if there is a feeling of heaviness on the lower back or a feeling as though the lower back were about to 'come off', the test is considered positive. |
| 6) Lack of hypomobility with PA glide test[7,13] | Participant in prone position. Examiner performs PA glide on the lumbar spinous process. If all lumbar segments are judged not to have stiffness (hypomobility), the test is considered positive. |

PIT: prone instability test; PSLR: average passive straight-leg raising; ROM: range of motion; PLE: passive lumbar extension; PA: posteroanterior; LBP: low-back pain.

each other's test results, performed the clinical tests and determined the test results for each participant. We allowed at least 15 minutes between the two sets of examinations to eliminate any possible change in clinical presentation due to replication of the examination procedure.

Each participant completed three self-reported outcome questionnaires: the Numeric Pain Rating Scale (NPRS),[17] the Modified Oswestry Low-back Pain Disability Questionnaire (OSW)[18] and the Fear-Avoidance Beliefs Questionnaire (FABQ).[19]

The NPRS assessed the severity of LBP on an 11-point (0–10) scale.[17] The modified OSW has 10 sections; one section is for pain severity, and the other nine represent various functional activities.[18] This questionnaire indicated the degree of LBP-attributed limitation in the specified activities. The FABQ assessed the level of fear-avoidance beliefs associated with LBP.[19] It consists of four items on physical activity (FABQ-PA) and seven items on the scale of work (FABQ-W).

After filling out the assessment tools, the participants underwent a series of specific tests as illustrated in Table 2.

After performing all the clinical tests, each examiner classified the participants into one of the three instability subcategories (structural, functional or combined). Participants were classified as structurally unstable, if they tested positive for any of the following: passive lumbar extension test (PLET) (Fig. 2),[20] lack of hypomobility with posteroanterior (PA) glide or the lumbar flexion ROM ($>53°$) test (Fig. 3A and B).[13] Based on the work by Kasai *et al.*[20] and Fritz *et al.*,[13]

**Figure 2 Passive lumbar extension test (PLET).**

if either the first or second test is positive, then the participant is approximately nine times more likely to have positive radiographic instability,[13,20] and about 4·8 times more likely if the third test is positive.[13]

The participant was considered functionally unstable if three out of four predictors of functional instability (CPR) were present: (1) age $<40$ years; (2) positive prone instability test (PIT; Fig. 4A and B); (3) aberrant motion present; and (4) average passive straight-leg raise (PSLR) ($>91°$). If three out of four predictors are present, then the likelihood of success with lumbar stabilisation exercises is + LR 4·0.[12]

The participants were considered to have combined instability, if they met the criteria for both subcategories (structural and functional instability).

### *Data analysis*

Data were analysed using the Statistical Package for Social Sciences (SPSS IBM Corporation 1989, 2011; Version 20).

(a) (b)

**Figure 3 (A) Lumbar flexion range of motion (ROM) >53° test, stage (1) and (B) lumbar flexion ROM >53° test, stage (2).**

(a) (b)

**Figure 4 (A) Prone instability test (PIT), stage (1) and (B) PIT, stage (2).**

The inter-rater reliability for the various lumbar instability tests were evaluated using kappa correlation coefficients in order to establish that the inter-rater reliability were greater than chance agreement. In addition, the inter-rater reliability of the instability subcategories (structural, functional and combined instability) were calculated by using kappa and adjusted kappa (PABAK) for all subcategories.

The kappa results were interpreted using Landis and Koch's[21] suggestions: <0·0 is poor; 0·0–0·20 is slight; 0·21–0·40 is fair; 0·41–0·60 is moderate; 0·61–0·80 is substantial and 0·81–1·0 is almost perfect. However, the prevalence-adjusted, bias-adjusted kappa (PABAK) values were calculated using Sim and Wright's[22] methodology and reasons. The main purpose for calculating the PAPAK is to reduce the influence of prevalence and bias factors on kappa magnitude. The prevalence index describes the proportion of agreements on the positive classification that differ from the negative classification. Therefore, in a $2 \times 2$ table, it is expressed by the absolute difference of $(|a-d|/n)$, the bias index is the extent to which the raters disagreed on the proportion of positive (or negative) cases and is reflected in a $2 \times 2$ table. Therefore, it is expressed by the absolute difference of $(|b-c|/n)$.

## Results

The percentage agreement and kappa value for all of the CPR tests [the PIT, the aberrant motion test and the average PSLR test (>91°)] showed a high percentage agreement (90, 97·5 and 95%, respectively) and substantial kappa coefficients (0·71, 0·79 and 0·77, respectively).

The lumbar flexion ROM >53°and PLET showed high percentage agreement (82 and 73%, respectively) and moderate kappa coefficients (0·48 and 0·46, respectively).

The lack of hypomobility with the PA glide test was found to be unreliable, with a low percentage agreement and a low kappa coefficient (25% and −0·02, respectively). However, locating the painful segment,

which is the first phase of the PIT test, was moderately reliable, with kappa=0·41. The intra-rater reliability for all clinical tests is shown in Table 3.

The functional and combined lumbar segmental instability categories were found to be perfectly reliable (PABAK=0·90 and 0·95, respectively), whereas the adjusted kappa for structural instability was found to be substantially reliable (PABAK=0·7). The subclassification reliability is shown in Table 4.

## Discussion

We investigated the reliability of the most valid physical examination tests (highest+LR) in the literature to identify lumbar segmental instability. Participants were classified as having structural instability if any of the structural instability tests were positive. This includes lumbar extension test ( + LR 8·84),[22] lack of hypomobility with the PA glide test ( + LR 9)[13] and the lumbar flexion ROM >53° ( + LR 4·8).[13] The participants were classified as having functional instability, if three out of four of the stabilisation CPR were positive.[12] These include age (<40 years), average PSLR (>91°), aberrant motion test and PIT ( + LR 4·0).[12]

The PLET was validated by Kasai *et al.*[20] for patients (age range=39 to 88 years, mean=68·9 years) who had experienced chronic pathologies such as lumbar stenosis, lumbar spondylolisthesis and lumbar degenerative scoliosis. In this study, we included younger patients (age range=21–71 years, mean age=35 years) who had recurrent or chronic LBP. For these patients, the PLET test displayed acceptable inter-rater reliability even for the younger age group (with kappa=0·46). Recently, Rabin *et al.*[23] investigated the reliability of the PLET on general LBP subjects and found it to be substantially reliable (kappa=0·76).

The lumbar flexion ROM >53° test was found to be moderately reliable (kappa=0·48), with a high percentage of agreement (82·5%). This finding is in agreement with the findings of a previous study that also showed the test to be moderately reliable (ICC=0·60).[13] However, as far as we know, this was the first study that explored the reliability of the test as a categorical non-continuous measure. Furthermore, previous studies showed a high correlation between lumbar flexion ROM and the flexion–extension radiograph.[24] It is notable that this test replicates the first portion (lumbar flexion) of the radiographic procedure in a standing position.

Two previous studies have reported the segmental-mobility test to be unreliable in the prone position, with no more than chance agreement (kappa=−0·02 to 0·26 and −0·20 to 0·17, respectively).[15,25] Hicks *et al.*[15] have reported the judgement of hypomobility with PA glide to be unreliable (kappa=0·18), and the judgement of any hypermobility with PA glide to be fairly reliable (kappa=0·30). Conversely, Fritz *et al.*[13] reported the judgement of hypomobility with PA glide to be fairly reliable, and a hypermobility judgement to be moderately reliable (kappa=0·38 and 0·48, respectively).

The lack of hypomobility with the PA glide test is determined as having less than chance agreement (kappa ranging from −0·22 to 0·18). We believe that

**Table 3 The reliability coefficient for all lumbar segmental instability tests**

| Variables | Kappa (95% CI) | Percentage agreement | Examiner 1 −ve/+ve | Examiner 2 −ve/+ve |
|---|---|---|---|---|
| (1) PLET | 0·46 (0·20, 0·72) | 0·725 | 19/21 | 26/14 |
| (2) Lumbar flexion ROM >53° | 0·48 (0·16, 0·80) | 0·821 | 7/33 | 10/30 |
| (3) Lack of hypomobility | −0·020 (−0·22, 0·18) | 0·250 | 29/11 | 11/29 |
| (4) Aberrant motion | 0·79 (0·39, 1·19) | 0·975 | 37/3 | 38/2 |
| (5) PIT | 0·71 (0·45, 0·98) | 0·900 | 9/31 | 9/31 |
| (6) Average PSLR >91° | 0·77 (0·47, 1·08) | 0·950 | 35/5 | 35/5 |
| (7) Locating the pain-provoking segment* | 0·41 (0·18, 0·63) | 0·60 | 4/36 | 2/38 |

PLET: Passive lumbar extension test; ROM: range of motion; PA glide: posteroanterior glide; PIT: prone instability test; PSLR: passive straight-leg raising.
* Locating the pain-provoking segment is the first phase of PIT and not one of the six segmental instability tests.

**Table 4 Unadjusted kappa, adjusted kappa 'PABAK', prevalence and bias indices, and percent of positive and negative agreement**

| Instability categories | Unadjusted kappa | 95% CI for unadjusted kappa | Percent agreement | PABAK adjusted kappa | Prevalence index | Bias index | Percent of positive agreement | Percent of negative agreement | *n* |
|---|---|---|---|---|---|---|---|---|---|
| Structural | 0·19 | −0·21, 0·59 | 85% | 0·70 | 0·80 | 0·1 | 92% | 25% | 33 |
| Functional | 0·72 | 0·36, 1·09 | 95% | 0·90 | 0·80 | 0·0 | 75% | 98% | 3 |
| Combined | 0·84 | 0·55, 1·14 | 98% | 0·95 | 0·83 | 0·3 | 86% | 99% | 3 |

PABAK: prevalence-adjusted, bias-adjusted kappa; *n*: number of participants in each category.

inclusion of the normal category added confusion to the already poorly reliable test because this category is in the grey zone between the judgement of slight hypomobility and slight hypermobility. Additionally, we considered the lack of hypomobility glide to be an indirect test because the examiners had to assess the mobility of all lumbar segments first. If they did not identify a hypomobile segment, the test was considered to be positive. Therefore, this is a test of exclusion of the hypomobility judgement. Moreover, the presence of a hypomobile segment does not exclude the possibility of the presence of other hypermobile segments, which could be functionally unstable segments. Therefore, in the presence of a hypermobile segment, some instability participants might have been excluded based on the presence of one hypomobile segment.

It is important to note that we did not rotate the order of examiners. We did, however, allow at least 15 minutes delay between the two sets of examinations in order to minimise the potential change in clinical presentation due to procedural repetition. However, the variation in the positive test results between the raters (11 participants compared to 29 participants) might potentially indicate a carryover effect, in which the performance of the first palpation procedure may have altered the mobility of segmental motion prior to the second set of examinations. In the literature, Fritz *et al.*[13] came up with the highest inter-rater reliability for both hyper and hypo segmental-mobility tests (0·48 and 0·38, respectively). In their study, they did not switch the order of the examiners and adopted a shorter rest-time (5 minutes) between the two sets of examination procedures. Thus, the length of rest-time might be a potential confounding variable on the reliability studies of lumbar segmental-mobility tests. It may be reasonable to study this variable in future studies.

Locating the pain-provoking segment is the first portion of the PIT. Segmental pain provocation was found to be moderately reliable (kappa=0·41). This finding is in line with the findings of previous studies of pain provocation judgement.[13,15,25]

The reliability of the aberrant motion test was substantial (kappa=0·79), which is similar to that found by Hicks *et al.*[15] (kappa=0·60) and Rabin *et al.*[23] (kappa=0·64). The latter studies recruited participants who had general LBP, with recurrent LBP participants equivalent to 81–66%, respectively,[15,23] compared to 85% in this study.

The PIT was substantially reliable (kappa=0·71). This finding is consistent with reports from previous studies: Rabin *et al.*[23] (kappa=0·67) for general LBP patients and Fritz *et al.*[13] (kappa=0·69) for the patients who were referred for the flexion–extension radiographs due to suspicion of lumbar instability. Hicks *et al.*[15] (kappa=0·87), found almost perfect reliability. These frequent reports of high reliability of the PIT support its generalisability to a wide spectrum of clinical examiners.

The average PSLR in this study was substantially reliable (kappa=0·77). This result is similar to that found by Rabin *et al.*[23] (kappa=0·73), who repeated the test twice before recording the test scores at the third iteration. They performed the test as described by Hicks *et al.*[12] who used a description of the test provided by Waddell *et al.*[26] Neither description mentions the repetition of the PSLR procedure; instead, the examiner is required to record the test's result the first time the test is performed. This helps to avoid any chance of the participant's passing the 91° mark due to the stretch effect produced by repeating the test.

We divided the lumbar segmental instability into three categories: functional instability (dysfunction of neuromotor control), structural instability (disruption of passive stabilisers) and combined instability (dysfunction of both the neuromotor control and the passive subsystem).

The kappa coefficient for the functional-instability category was substantial (95%, kappa=0·72), slightly lower than that found by Rabin *et al.*[23] (93%, kappa=0·86). The combined instability result was almost perfect (98%, kappa=0·84). We found that most participants who had functional instability also had structural instability. This may be because young and flexible subjects were likely to pass both cutoff values for the ROM test: PSLR $>91°$ and lumbar flexion range $>53°$. This is especially true because passing the PSLR cutoff value increases the chance of subject allocation to functional instability, whereas passing the lumbar flexion ROM cutoff value directly allocates the subjects into the structural category.

Even though there was high agreement between the raters (85%), the kappa value for structural instability was poor (kappa=0·19). This phenomenon is known as the kappa paradox: the examiners agree more closely on the subjects who have the condition of interest (positive structural instability, percent of positive agreement=91%) than on the participants who do not have the condition of interest (negative structural instability, negative percent of agreement=25%). This imbalance in the percent of agreement between positive and negative ratings skewed the magnitude of kappa.[27] Furthermore, positive structural instability was common, as indicated by the high prevalence index and a high percentage of positive agreement. This increased the percentage of chance agreement and thus reduced the kappa value.[22]

One way to reduce the skewed influence of prevalence and bias indices is to calculate the PABAK or adjusted kappa.[22,28] Some statisticians recommend using adjusted kappa to eliminate the adverse effects of prevalence and bias on the true value of kappa derived from the study.[22] Considering the high prevalence indices of

all lumbar instability categories, we calculated PABAK to find the true value of kappa after adjusting the prevalence and bias indices. We found that all of the categories rounded up to about 0·18 and 0·11 for functional and combined instability categories, respectively. However, the kappa value of structural instability increased dramatically by about 0·51 to become substantially reliable (kappa=0·7). This indicated that the prevalence and bias indices adversely affected the structural instability category more than the other categories. Thus, the established adjusted kappa value was more representative of the observed high agreement between the raters.

We recommend that further research efforts be directed towards establishing the cluster of structural instability tests that can be used as screening tools to rule out structural instability among LBP patients. This can be accomplished by studying all structural tests with the inclusion of PLET and the lumbar flexion range $>53°$ test in one comprehensive valid and reliable study against the radiographic gold standard.[7,8,13,20]

Furthermore, in view of the lumbar mobility test's poor reliability in the prone position, we agree with previous research findings that recommend the exploration of the added effect of using a pressure/force device prior to the reliability study. In addition, we support the exploration of the reliability of other kinds of lumbar mobility testing, such as the side-lying lumbar mobility test.[15,25]

Finally, we would like to mention some of the limitations of this study: first, we examined the inter-rater reliability between two examiners; this might limit the generalisability of the study results in a wide range of examiners. Second, although we studied the inter-rater reliability of structural instability tests as part of the study, it is warranted to re-examine the reliability of the structural tests on participants who have already been examined by flexion–extension radiograph. Third, the 30-minute training session for the examiners was rather short and may have led to inconsistencies in the performance. Moreover, the 95% confidence interval for the kappa coefficient was noticeably wide and might have affected the kappa precision.

## Conclusion

We studied the inter-rater reliability of six clinical tests that might predict the radiographic diagnostic standard or the outcome of stabilisation therapy in 40 participants who had R/CLBP. The kappa correlation coefficient values of the functional-instability tests of the lumbar spine confirmed these tests to be substantially reliable. The lumbar flexion ROM and PLETs were also found to be adequately reliable. Conversely, lack of hypomobility with the PA glide test was found to be unreliable and, in many cases, worse than chance. Finally, the subclassification of patients into lumbar instability categories was adequately reliable, as depicted by their PABAK values.

## Acknowledgements

We would like to thank the following physical therapists for their substantial support in this study: Mazen Alqahtani, PT; Oscar Ramirez, PT; and Sharick Shamsi, PT.

## Disclaimer Statements

**Contributors** Faisal M. Alyazedi: main designer of the study, the baseline data collector, interpreter of the results and the writer of the article. Everett B. Lohman: the first examiner of the study and helped in the study design and revision of the first draft of the manuscript. R. Wesley Swen: the second examiner, helped in the study design and revision of the first draft of the article. Khaled Bahjri: the statistician of the study, helped in the design before the beginning of the study, conducted all statistical analysis and revised the drafts of the manuscript.

**Funding** Loma Linda University.

**Conflicts of interest** There are no conflicts of interest.

**Ethics approval** All procedures of the study were approved by the Loma Linda University Institute of Review Board.

## References

1 Hauggaard A, Persson AL. Specific spinal stabilisation exercises in patients with low back pain – a systematic review. Phys Ther Rev. 2007;12(3):233–48.
2 Waddell G. Volvo award in clinical sciences. A new clinical model for the treatment of low-back pain. Spine (Phila Pa 1976). 1987;12(7):632–44.
3 Cairns MC, Foster NE, Wright C. Randomized controlled trial of specific spinal stabilization exercises and conventional physiotherapy for recurrent low back pain. Spine (Phila Pa 1976). 2006;31(19):E670–81.
4 Pengel LH, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. BMJ. 2003;327(7410):323.
5 Hides JA, Jull GA, Richardson CA. Long-term effects of specific stabilizing exercises for first-episode low back pain. Spine (Phila Pa 1976). 2001;26(11):E243–8.
6 Demoulin C, Distree V, Tomasella M, Crielaard JM, Vanderthommen M. Lumbar functional instability: a critical appraisal of the literature. Ann Readapt Med Phys. 2007;50(8):677–84.
7 Beazell JR, Mullins M, Grindstaff TL. Lumbar instability: an evolving and challenging concept. J Man Manip Ther. 2010;18(1):9–14.
8 Alqarni AM, Schneiders AG, Hendrick PA. Clinical tests to diagnose lumbar segmental instability: a systematic review. J Orthop Sports Phys Ther. 2011;41(3):130–40.
9 Knutsson F. The instability associated with disk degeneration in the lumbar spine. Acta Radiol. 1944;25(5):593–609.
10 Panjabi MM. The stabilizing system of the spine. Part II. Neutral zone and instability hypothesis. J Spinal Disord. 1992;5(4):390–396, [discussion 397].
11 Panjabi MM. Clinical spinal instability and low back pain. J Electromyogr Kinesiol. 2003;13(4):371–9.
12 Hicks GE, Fritz JM, Delitto A, McGill SM. Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. Arch Phys Med Rehabil. 2005;86(9):1753–62.
13 Fritz JM, Piva SR, Childs JD. Accuracy of the clinical examination to predict radiographic instability of the lumbar spine. Eur Spine J. 2005;14(8):743–50.

14 Leone A, Guglielmi G, Cassar-Pullicino VN, Bonomo L. Lumbar intervertebral instability: a review. Radiology. 2007;245(1):62–77.
15 Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. Arch Phys Med Rehabil. 2003;84(12):1858–64.
16 Abbott JH, McCane B, Herbison P, Moginie G, Chapple C, Hogarty T. Lumbar segmental instability: a criterion-related validity study of manual therapy assessment. BMC Musculoskelet Disord. 2005;6:56.
17 Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. Spine (Phila Pa 1976). 2005;30(11):1331–4.
18 Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. Phys Ther. 2001;81(2):776–88.
19 Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain. 1993;52(2):157–68.
20 Kasai Y, Morishita K, Kawakita E, Kondo T, Uchida A. A new evaluation method for lumbar spinal instability: passive lumbar extension test. Phys Ther. 2006;86(12):1661–7.
21 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
22 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85(3):257–68.
23 Rabin A, Shashua A, Pizem K, Dar G. The interrater reliability of physical examination tests that may predict the outcome or suggest the need for lumbar stabilization exercises. J Orthop Sports Phys Ther. 2013;43(2):83–90.
24 Mayer TG, Tencer AF, Kristoferson S, Mooney V. Use of noninvasive techniques for quantification of spinal range-of-motion in normal subjects and chronic low-back dysfunction patients. Spine (Phila Pa 1976). 1984;9(6):588–95.
25 Schneider M, Erhard R, Brach J, Tellin W, Imbarlina F, Delitto A. Spinal palpation for lumbar segmental mobility and pain provocation: an interexaminer reliability study. J Manipulative Physiol Ther. 2008;31(6):465–73.
26 Waddell G, Somerville D, Henderson I, Newton M. Objective clinical evaluation of physical impairment in chronic low back pain. Spine (Phila Pa 1976). 1992;17(6):617–28.
27 Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6): 551–8.
28 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993;46(5):423–9.