Author’s response to reviews

Title: Inter- and intra-observer reliability of clinical movement-control tests for marines

Authors:

Andreas Monnier (Andreas.Monnier@ki.se)
Joachim Heuer (jocke.heuer@gmail.com)
Kjell Norman (Kjell.Norman@mil.se)
Björn O. Ång (Bjorn.Ang@ki.se)

Version: 4 Date: 22 September 2012

Author’s response to reviews: see over
Dear Prof Francesco Violante,

Thank you for your mail of June 29. We were pleased to know that our manuscript is acceptable for publication in *BMC Musculoskeletal Disorders*, subject to adequate revision and response to your and the reviewers’ comments. We would like to express our sincere thanks to the reviewers. They identified areas that needed correction or modification.

We have now revised the manuscript in the light of the reviewers’ comments. Attached, please find our responses. In addition, a professional native-English language reviewer has examined the text, and we believe now that the text has improved.

We hereby submit the revised manuscript for your reconsideration for publication in *BMC Musculoskeletal Disorders*.

Sincerely yours

Björn O. Äng
Reviewer's report
Title: Inter- and intra-observer reliability of clinical movement-control tests for marines
Version: 1 Date: 27 May 2012
Reviewer: Hannu Luomajoki

Reviewer's report:
Title: Inter- and intra-observer reliability of clinical movement-control tests for marines
Version: 1 Date: 27.5.2012
Reviewer: Hannu Luomajoki

Reviewer's report:
General
The paper is well organised and written. The topic has actuality. Although similar studies have been published (and they are acknowledged in referencing), there have been no papers on high load tests and on soldiers or sports people (which might be similar on their loading levels). However, I would like to address some critical points and help to improve the paper.

Major Essential Revisions

Reviewer's comments
1. You call this test battery “movement control tests of lumbar spine and lowerlimb”. However, I can see only one test regarding lower extremity SLKB + LL. And also by this test, it seems that only the alignment of the lower back is controlled. So, I do not think there is much of movement control tested of the lower extremity. If so, I would like to see the lunging from front and assess whether an adduction of the knee is happening and whether the pelvis stays in horizontal etc. And of course tests like one leg standing knee bending, squatting and some jumping would be more adequate for movement control of the legs.

Reply
Two tests (SLKB + LL and DLL-ALE) assess control of both lumbar spine and hip/lower extremity. The photos in the figures illustrate examples of views examining these parts during the test, now clarified in the figure text. In addition, we have now changed the test descriptions to “movement control of the back and lower extremities” to “movement control of the back and hip” throughout and have added an appendix (Appendix A) exemplifying the detailed test protocol with the exact start position, test movement instructions, observation criteria and analysis (dichotomization) for three of the tests (with permission from the copyright holder): SLKB + LL, DLL-ALE and DSLL. The other tests are described in detail elsewhere, and we refer to that source (as we didn’t get permission from the copyright holder).

Further, our test protocol and the tests were selected from various clinical tests used for prevention and management of injured marines, based on discussions with clinicians with fairly long experience of assessing this population. This is now clarified under Method; Clinical tests, page 8, lines 13-15.

Reviewer’s comments
2. I have also some difficulties with all these abdominal tests (BKFO and DLL-L;
DLL-ALE; DSLL). Four of the six tests are actually more or less testing the abdominal muscles. Can we call them movement control tests? Are they not testing more the ability of the abdominals to control for the stability of the pelvis and lumbar spine?

**Reply**

In contrast to many other protocols that include abdominal activation, all our tests included an active movement component. This was so as to have a certain loading of the hip/lower back, i.e. loading directed to a weak link in the marines’ musculoskeletal system, following our clinical/empirical experience with marines. We believe this reflects important discrepancies compared to simply testing the lumbar spine and pelvis in a static or semi-static position, e.g. with the abdominal hollowing maneuver (Richardson C, Jull G: Muscle control–pain control. What exercises would you prescribe? Manual Therapy 1995, 1:2-10.) or the side bridge (McGill SM, Karpowicz A: Exercises for Spine Stabilization: Motion/Motor Patterns, Stability Progressions, and Clinical Technique. Archives of Physical Medicine and Rehabilitation 2009, 90:118-126.). Further, four test include two or more movement directions (e.g. flexion to extension) in which the lumbar spine and pelvis are observed. In considering this, we would therefore like to define our testing as movement control tests. We have now added information to clarify this in under Method/Clinical test, page 8, lines 18-23.

**Thank you.**

**Reviewer’s comments**

3. I also wonder, how functional these tests are? Are marines experiencing difficulties with their back while lying supine? Are these functions (the test positions) typical in all day duty by marines?

**Reply**

As indicated above, the present tests were initially selected and then further evaluated to reflect marines’ functional requirements related to their work and the critical points related to the physical demands upon them. For example, marching with heavy equipment may after several hours/days cause some marines to “lose their movement control” and move somewhat aberrantly (while others do not), and they may soon experience hip or back difficulties. In testing such insufficiency, however, one is reluctant to ask marines to do such long-term activity in a clinical testing/screening program. We simply need efficient clinical convenient tests. Here we have learned that certain tests may possibly reflect/test these, field-observed, weak links adequately and straightforwardly – hence the present tests. Thus these tests are performed in ‘functionally related positions/situations’, rather than in ‘functionally-specific’ situations. We have now added information to clarify this under Background, page 5, lines 3-7, and Method; Clinical test, page 8, lines 21-23.

Notably, empiric observation/testing are indeed performed over several days when initially recruiting marines (e.g. coastal rangers). The tests we are suggesting are, rather, designed for testing/screening in an existing operational population that will operate for some years.

**Reviewer’s comments**

4. If movement control is the focus, why not use tests which are already published but adapt them to more high load?

**Reply**

As indicated, these tests were initially selected (adequacy with a certain test, and the tests together) through years of clinical experience and empirical observations with marines. We
now make this clearer under Background, page 5, lines 3-7, and Method section, page 8, lines 13-16. We are well aware that our “high-load tests”, as detailed herein, have not yet been published. However, our clinical experience with these tests indicates that they adequately cover critical points of movement control associated with musculoskeletal disorders in back and lower extremities in this group, at least to a level that warrants them further evaluation as we have done here. Other tests may be considered in the future, but are outside the scope of the present paper.

Reviewer’s comments
5. So, I think you might rethink the name of the testbattery and the abilities tested and accordingly rethink the title of the paper?

Reply
As outlined above, we have now clarified our definition of a movement control test and specified the regions tested in the Background and Method sections (not, alas, accurately defined in our initial manuscript). “Movement control of the back and lower extremities” has now been changed to “movement control of the back and hip”. We believe that with these changes, the present title and tests now reflect the contents.

Reviewer’s comments
6. I would like to see more information describing the exact instructions to the participants in the study. How do you tell the participants what to do? Can you publish the standardized instructions?

Reply
As suggested, an Appendix A now details test protocol, covering start position, test movement instructions, observation criteria and analysis (dichotomization) for three of the tests (with permission from the copyright holder) to supplement the references for the test. Regarding detailed instructions for the other three tests, we need to refer to the text references where they are clearly described (since we lack publishing permission from the copyright holder). However, readers thus have access to details of all tests.

Reviewer’s comments
7. Figures: there is always a notion “reproduced with permission from Movement Performance Solutions”. Does that refer to the picture or to the test itself? I cannot imagine that some could “own” a certain test where you would need a permission. So, I think that permission should be omitted.

Reply
These tests, with adherent instructions, are not yet published (except in copyrighted materials), and our permission from the copyright holder to reproduce this material requires the text given. Our pictures are from our group.

Discretionary / minor Revisions

Reviewer’s comments
You collected both a symptomatic and asymptomatic population. However, the level of pain was only NRS >1/10, which is very little for the subjects to be labelled as “patients”. I wonder whether Sensitivity and specificity are the correct labels here, as they would be used when you have a clear reference standard. Here the reference is only, if the second group has more pain than 1/10 NRS. So, basically for a diagnotis test you do not need this test battery (as the reference standard is pain or not – you could just ask if they have pain – without testing). I
wonder whether Odds ratios would be more adequate to display the differences between the groups.

Reply
Our clinical experience with this group is that they, without exaggeration, tend to underestimate their level of pain. In our communication with other military communities/nations, this seems to be common in this/similar types of population (e.g. Carragee EJ, Cohen SP: Lifetime asymptomatic for back pain: the validity of self-report measures in soldiers. Spine 2009, 34:978-983.). We have certainly considered this phenomenon, and believe that rating any pain (i.e. >1 NRS) simply means pain experience in this group. We believe therefore that our categories pain or no-pain subjects are fairly clear. Further, our experience with military personnel is that early recognition of deficits in movement control, possibly related to pain, may sometimes be more reliable/informative than simple asking about pain experience (the present pain-ratings were anonymous), particularly in preventive management. We believe therefore that our clinical tests are an important complement to “asking if they have pain”. This has now been stressed under Background, page 5, line 25- page 6, line 3.

Importantly, the present sample was recruited from one and same population, i.e. marines on active duty, and not from two samples with symptomatic and asymptomatic subjects, respectively (and note also that there is no labeling of “patients” as commented by the reviewer). Marines on active duty simply constitute subjects with and without pain experience, and this is actually the fundamental basis for the study. We believe that recruiting related to subjects seeking care is not very useful if one is really interested in musculoskeletal pain (rather than in behavior).

Reviewer’s comments
Otherwise the tables and statistics are clear and carefully documented. Statistical review: Page 10. Last paragraph of analysis. I think “the lower the AIC value generated the better the fit of the model”: I think this should be “the higher”. Please check with a statistician.

Reply
It is correct as written, i.e. “the lower the AIC value generated the better the fit of the model” (e.g. Maindonald J, Braun J: Data Analysis and Graphics Using R: An Example-based Approach. 3 edn. New York: Cambridge University Press; 2010). This can also be checked by observing the formula presented, i.e. “AIC= 2k – 2L, where k is the number of variables (or tests) in the statistical model and L is the log likelihood of the model.” We believe such an approach is important since the addition of any variable (or clinical test) could otherwise improve the log likelihood, even though it is not really associated with the dependent variable. Here, the AIC equation aims to “penalize” or correct the addition of new independent variables/clinical tests not significantly associated with the dependent variable by adding twice the number of parameters to the model fit.

We thank the reviewer for the comments and for the time spent reviewing our work.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.
Declaration of competing interests:
'I declare that I have no competing interests'
Reviewer's report
Title: Inter- and intra-observer reliability of clinical movement-control tests for marines
Version: 1 Date: 3 June 2012
Reviewer: Harald Ekedahl

Reviewer's report:
This manuscript has the potential to improve clinical practice through analysis of reliability and validity of these tests. However, there are some major points to be addressed in this manuscript.

Major points:

Reviewer’s comments
- The logic of studying a cohort of marines

Reply
Musculoskeletal disorders in marines are certainly a significant military problem, reducing operational efficiency before and during deployment; and injured marines with such disorders show very little likelihood of returning to operational duty (Cohen SP, et al. Diagnoses and factors associated with medical evacuation and return to duty for service members participating in Operation Iraqi Freedom or Operation Enduring Freedom: a prospective cohort study. The Lancet. 2010;375(9711):301-9.). Beside marines, our results could – we believe – be of interest for similar expeditionary military units and tactical athletes with reported high incidence of musculoskeletal injuries. Clinical tests are regularly conducted in these groups for monitoring health and operational status, and reliability and validity of clinical tests are commonly discussed nationally as well as internationally. Our background in the present manuscript is now, however, more comprehensive, and we believe therefore that the logic of studying this population should now be clear.

Reviewer’s comments
- The validity analysis and the dependent variables in this analysis

Reply
We have now added information on “Subjective measures of back and lower extremity pain and distribution to body regions” in Table 1. We have also added information under Methods regarding subjective measures of pain, page 7, lines 26 - 27 and page 8, lines 2-5. This is further addressed in the Discussion, page 14, lines 9-12, and we think that the dependent variables in this analysis are now clearly described, i.e. pain or no pain.

Reviewer’s comments
- Can the validity be supported although reliability is not supported

Reply
Even though we found test-retest reproducibility fair-to-moderate (i.e. modest intra-reliability κ-coefficients, probably due to learning effects), we believe the inter-reliability is acceptable (κ-coefficients of inter-observer reliability ranged between 0.49-0.89). However, limitations on inter-and intra-observer reliability and discriminative validity are now more thoroughly addressed in the Discussion, and more explicitly considered in the Conclusion.

Reviewer’s comments
- The recommendations for clinical practice and future research
Reply
We do recommend that the less accurate test-retest reproducibility should be considered when used in follow-up evaluations. However, this is not the same as not recommending the test for clinical use. Several factors could influence the kappa values, including systematic improvement. These are addressed in the Discussion, page 15, lines 7-22. More practice rounds could perhaps reduce the possible systematic improvement learned herein, hence increase kappa values.

Minor points
Reviewer’s comments
- The subject group is not adequately described regarding low back pain, disability and radicular pain.

Reply
We have now added information on “Subjective measures of back and lower-extremity pain and distribution to body regions” in Table 1.

Reviewer’s comments
- The method needs clarification

Reply
We have now added more information in all Method sections, pages 6-11. This includes information on pre-study training, subjects’ characteristics in relation to the population, pain intensity and distribution and how long it took to complete the test protocol. Further, the test-retest procedure is clearly described. Related limitations are further addressed in the Discussion, page 14, lines 4-6.

We have also added an Appendix A with detailed test protocol giving the exact start position, test movement instructions, observation criteria and analysis (dichotomization) for three of the tests (with permission from the copyright holder) to supplement the references for the test.

Reviewer’s comments
- Strength and limitation need to be discussed further

Reply
Limitations of the study are now further described in the Discussion, pages 13-16.

We thank the reviewer for the comments and for the time spent reviewing our work.

Level of interest: An article of importance in its field
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's report
Title: Inter- and intra-observer reliability of clinical movement-control tests for marines
Version: 1 Date: 8 June 2012
Reviewer: Gertrud Nilsson
Reviewer's report:
The purpose of this study is of great interest and many articles in the same area have been published. Both reliability and discriminative validity are of great importance. However, there are questions in this manuscript that have to be considered.

REVIEWERS' COMMENTS AND OUR REPLY
Reviewer Gertrud Nilsson:

Reviewer’s comments
Background:

Major Compulsory Revisions:
In the background section I think the authors should define and describe the concept “movement control”, “motor-control” and “movement-control tests” in order to make clear to the reader exactly what this study deals with.

Reply
We have now added descriptions on this, page 4, lines 20-21, and also a definition of movement control tests, page 5, lines 9-11. This is now hopefully clear to the reader.

Reviewer’s comments
Methods

Study design:
Line 5: Should it not be The observers were blinded…?

Reply
Changed as suggested, page 7, lines 1-2. Thank you.

Reviewer’s comments
Minor Essential Revisions

Study sample:
How were the marines enrolled in the study, consecutively as they volunteered, randomized from a larger population, in other ways? Were they representative for the population? In the equation in the Analyses the population (N) has been used. What was the figure for N?

Reply
Eligible subjects was recruited from a representative Swedish marine company including all marine arms branches studied, such as assault infantry, combat craft crews and coastal rangers. After an oral presentation of the study aim and method to the assembled company,
subjects were supplemented by written information and volunteers were then enrolled. We have now added information on this under “Study sample”, page 7, lines 19-20.

A comparison with internal military statistics (i.e. age, gender, rank and arms branches; details not for public presentation) indicates that our sample is representative. N = 250 (this number will though increase since the Swedish Armed forces are becoming “fully professional” and thus with no conscripts).

**Reviewer’s comments**

**Major Compulsory Revisions:**

**Subjective measures of back and lower extremity pain**

The medical history for the previous six months is assumed to include pain (it is not clearly expressed). As this type of pain turned out to be the main results of this study, it should be described more in detail. How was it measured, by numerical rating scales? Pain was defined as >1 which is very low. Was this level predetermined? In Table 1 the rated back pain and lower-extremity pain among the subjects has been described shortly as yes or no. It is of importance to know how much pain they had: the median value, range, interquartile range for example. How many had hip/thigh pain, knee pain and foot/ankle pain? Here the authors should bring more information.

**Reply**

We have now added information on medical history for the previous six months (under “Methods; Subjective measures of back and lower extremity pain”, page 7, lines 26 - 27), and this should now be clear. Note, however, discriminative validity including pain ratings is a secondary aim of the present study, not a second aim as we first, and erroneously, expressed it (reliability testing is the main aim). This is now changed, page 6, line 17. Pain was rated subjectively on numerical rating scales, i.e.’ pain at present’ and average and worst pain for “the previous six months”, which we have now clarified under Methods, page 7, line 26 – 27.

Our clinical experience with marines is that they clearly underestimate their level of pain (compared to civilian subjects). While a cut-off of >1 may be considered low/arbitrary, our experience is that a rating of “>1” means that a marine experiences pain, or ache/discomfort (as defined herein). A cut-off >1 has, for the same reason, also been used with military pilots in previous papers from our group. Also, previous research conducted in the marine community has used “any pain” to define musculoskeletal disorders (Hollingsworth DJ: The prevalence and impact of musculoskeletal injuries during a pre-deployment workup cycle: survey of a Marine Corps special operations company. J Spec Oper Med 2009, 9:11-15.). We now present our thoughts about this in the Discussion, page 14, lines 9-12. As suggested, we have detailed rated pain intensity according to body region in Table 1.

**Reviewer’s comments**

**Major Compulsory Revisions:**

**Clinical tests:**

What was exactly registered by the observers in the BKFO-test except for the < 4 mm HG of the PBU? The same question will be raised for the DLL-L test and the DSLL test. What were the criteria for the observations? The PBU is of course of great help to control the movement for the person being tested. How many passed or failed due to PBU criteria and how many due to observed movement control,
this should be clarified? In the SLKB+LL test: were the observation criteria just for the 5 s in the static position or also during the dynamic movement to take this position? I think the test protocol with the exact observation criteria for “pass” or “fail” for all the tests should be enclosed otherwise it is impossible for the reader or other researchers to understand what has been really observed and scored. It should be possible for others to reproduce the results from this study.

Reply
Regarding what was exactly registered by the observers: during the BKFO the subjects’ ability to prevent lumbar/pelvic rotation (no more than 4 mmHg pressure increase/decrease) during leg movement, was rated. Inability to do so was recorded as “fail”. During the DLL-L and DSLL the subject’s ability to prevent lumbar flexion (a pressure increase by 5 mmHg or more) or extension (a pressure decrease by 5 mmHg or more) during the leg movement was rated. Unfortunately, we lack separate data as to whether subjects passed or failed due to PBU criteria specifically, or whether the subject passed or failed in terms of observed movement control (the test protocol gives only fail/pass on certain criteria). These details were however not under investigation in this study. The observation criteria for the SLKB+LL test included both the dynamic movement and the ability to hold the static end position for five seconds.

As suggested, we have now added Appendix A with detailed test protocol showing the exact start position, test movement instructions, observation criteria and analysis (dichotomization) for three of the tests (with permission from the copyright holder). This is to supplement the references for the test. The other tests are described in detail elsewhere, and we refer to that source (as we lack publishing permission from the copyright holder). This should clarify what was registered by the observers.

Reviewer’s comments
Discretionary Revisions
Test protocol
All subjects performed at least one trial with feed-back from the test leaders and if needed further visual and oral instructions etc. This means, as I can understand it, that the subjects were instructed in detail how to perform the tests almost perfect. In order to evaluate discriminative validity, would it not be more relevant to examine how movements could be controlled without these learning effects? Will it be possible to apply these tests as they are described in this study as screening tests? Some individuals can probably perform the tests easily while for others it will be a great effort but they can manage due to feedback from the observers and from the PBU.

Reply
The reason for the practice run and feedback was to have the subjects understand the test performance, and thus to ensure that possible aberrant movement during the test reflected an inability to control movement rather than a simple misunderstanding at a cognitive level. Clearly, even though you understand the test performance perfectly, you may not perform accurately (since the tests challenging movement control). We have now added clarification under Method; Test protocol, page 9, lines 20-22.
Yes, we believe it is possible to apply these tests as described (including such aspects as time consumption), at least judging from years of experiences with these tests. Other tests have not been as successful.

**Reviewer's comments**

Major Compulsory Revisions:

**Analyses**

Power: The sample size was based on the primary aim: to determine inter- and intra-observer reliability. Was this sample size adequate also to investigate the discriminative validity? In Table 1 one group was formed having both Back and LE pain. How did the authors handle this group in the regression model? Only back or LE pain has been presented in Table 5. This should be clarified.

**Reply**

Power calculation was based on the primary variable as is common (discriminative validity as secondary aim, now clarified according to reviewer 1’s comments). Our sample size may however be considered rather small with respect to our regression analyses, at least according to suggestions in statistical literature (Petrie A, Sabin C: Medical statistics at a glance. 3ed edn. Oxford: John Wiley And Sons Ltd; 2010) for such analysis. This limitation is now explicitly addressed in the Discussion, page 14, lines 6 – 9, - thank you.

A subject with both back- and lower-extremity pain was analyzed in both the back pain and the lower-extremity-pain group. Information now added under Method, Analysis, page 11, lines 6-7.

**Reviewer’s comments**

Results

Minor Essential Revisions

One person was lost to follow-up, which group did this person belong to?

**Reply**

This subject was lost at the time of re-testing, and was analyzed in the back pain group. This should be clear from the Results.

**Reviewer’s comments**

Discretionary Revisions

What are the authors’ comments to the differences of kappa values between those tests having high values and those having low kappa values? Were they more difficult to perform, more difficult to evaluate, more difficult to instruct, others? There seems to be a learning effect between the two tests. What is the authors’ opinion? Was this learning effect between observers or within subjects or both? Did the two observers discuss the tests, the protocol, the way of evaluating “pass” or “fail” or other things between the two measurements? This could influence the higher kappa values when repeating the evaluation.

**Reply**

There appears indeed to be a learning effect between initial testing and re-testing (for certain tests), and this is discussed in the Discussion. We have certainly discussed/reflect as to whether some tests may have high vs. low kappa values. There is however no clear
indicator/aspect that may explain the present poor/good kappa for certain tests. This is now further addressed in the Discussion, page 15, lines 11-15. This is interesting as it reflects the importance of research in clinical practice, in this case revealing important information to our experienced clinicians.

Observers did not discuss the tests during or between the two measurements. They were however able to do so during the practice sessions before study start, now described under Method, page 10, lines 1-4.

Reviewer’s comments
Major Compulsory Revisions:

Sensitivity/specificity
In the BKFO-test 58 (88%) passed and 7 (12%) failed (observer A). Observer B had almost the same figures. In the regression analyses this test was found to be sensitive. In what aspect do the authors mean it was sensitive? As far as I can understand it this test does not meet the criteria to identify persons with back or LE pain. Furthermore the DSLL-test had low 95%CI: 0.16 (Table 2) (Test 1) and has although been included in the regression model and concluded to be of importance as a discriminative test. The authors had predetermined and defined the lower 95%CI to be > 0.2. In Table 5, LE pain: the values of observer A’s sens/spec was 59/81 while observer B had reversed figures 88/44, how should these figures be interpreted?

Reply
Thank you for your informative comments. The figures first referred to should, though, be 28 [85%] and 5 [15%], not 58 (88%) and 7 (12%), as listed in the initial manuscript. Also, we wish to stress that the BKFO test mentioned above was considered as sensitive when together with DSLL, i.e. these tests emerged as best-fitting model, and that the figures listed for sens/spec represent estimated levels with respect to the model. However, our point was that these two tests reflect two different aspects of movement control (low- and high-load conditions), and may together more accurately predict the outcome than a single test could. This remains so although the regression intends to find the minimal goodness-of-fit of the included variables/tests. This means that the AIC equation aims to “penalize” or correct the addition of new independent variables/clinical tests not significantly associated with the dependent variable. As we initially described it: “Our results support the use and interpretation of test combinations, rather than information from single tests.”

However, we agree with the reviewer that the low 95%CI of 0.16 (0.58 for test 2) for the DSLL test indicates an increased risk of measurement error: this limitation is outlined in the Discussion, page 15, lines 22-24.

Regarding the sens/spec values of 59/81 vs. 88/44 (Table 5), these figures represent BKFO and DLL-L (observer A) and SKLB+LL and DLL-L (observer B). They meet, or approach, the criteria of 60% as defined, although the specificity for observer B is indeed rather poor. We have now, therefore, more clearly discussed that there were differences between these two observers, page 16, lines 6-8 (now also explicit considered in the Conclusion). In short, we agree that with these tests, discriminative validity is somewhat problematic (although we believe validity-analysis is important), and the conclusion here is now more modest. Thank you.
Reviewer’s comments
Furthermore, one important step in the discriminative validity has been passed over. Before being included in the regression model, the capability of each test to identify individuals with back pain and LE pain should be analysed for observer A and B. Thereafter criteria for the regression model should be set up, that is the lowest kappa value and the lowest 95% CI as it has been done in the reliability evaluation.

Reply
Yes, selection criteria and pre-step sifting of tests could have been an option. We elected however not to process each test in such pre-filtering, partly since we have learned that the AIC autoregression rather accurately separates tests that do not really relate/contain properties with the dependent variable, and partly since discriminative validity was in fact a secondary aim and perhaps should not take up too much space in the present paper. We have, however, added information about this possible approach in the Discussion, page 16, lines 8-10. Thank you.

Reviewer’s comments
In conclusion:
Many articles in this area are now being published and I think the present study is an interesting contribution on this subject. Due to my opinion the results from the intra- and inter-reliability evaluation could be published after revision as suggested. The evaluation of the sensitivity of the clinical tests for discriminating musculoskeletal pain disorders have many weaknesses as described above and before these results will be published I think there are major revisions that have to be done. To my opinion it cannot be concluded from the data from this manuscript that both low- and high threshold tests had discriminative validity as there are too many ambiguities.
Best regards,
Gertrud Nilsson

Reply
We believe our replies and manuscript changes meet the reviewers’ requests, particularly since we consider data on reliability AND validity in clinical testing is important for the overall understanding and interpretation of such testing. We have learned and discussed this addition of validity analysis in reliability studies, in several recent workshops/congress panels. We hope therefore that our design may be useful for other researchers tackling reliability issues in clinical testing. Our discussion of whether low- and high threshold tests had discriminative validity is changed to more modest. Thank you.

We thank the reviewer for the comments and for the time spent reviewing our work.

Level of interest: An article of importance in its field
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
Reviewer's report

Title: Inter- and intra-observer reliability of clinical movement-control tests for marines

Version: 1 Date: 11 June 2012
Reviewer: annemarie vlaar

Reviewer's report:
The authors investigate the inter- and intra-observer reliability and the sensitivity and specificity of 6 clinical movement-control tests for low back and lower extremities in 33 marines. The study question and methods posed by the authors is well defined.

Major compulsory revisions:
1. the gold standard in this study is the question if the volunteer had pain the last six months or at the moment of the investigation. I think this is a suboptimal gold standard to calculate the sensitivity and specificity.

Reply
First, we would stress that the determination of inter- and intra reliability is the main question in this study (discriminative validity was a secondary aim), and this should now be clear after our revision. Regarding the pain ratings, we agree that they may be considered suboptimal, particularly when reflecting on “ordinary” civilian patients. However, we have learned that the present ratings (pain at present and pain under a certain time-period) are of great interest for clinical (at present) or preventive (pain in the past) action. Also, a cut-off of >1 may be considered arbitrary (see reply to Gertrud Nilsson). However, in our experience a rating of “>1” means that a marine experiences pain, or ache/discomfort (as defined herein).

Reviewer’s comments
2. the authors included only marines on duty and no marines who seeks medical help. They write in the discussion this is a strength of their study, however to calculate the sensitivity and specificity I think it necessary to include proven "ill marines" and "healthy marines".

Reply
Subjects were recruited from a population of marines on active duty. This means that subjects indeed could have sought medical attention, as only marines on full- or part-time sick leave were excluded (refer to our exclusion criteria). Further, as preliminary results from our research group show that the 12 month prevalence of musculoskeletal disorders, in this military group is as high as 70%, of whom only a fraction seek medical attention, we believe using “any pain” and marines “on duty” reflects a realistic screening situation for this population. In addition, recruiting subjects from the health-care system (where marines seek medical help) affords little information about illness in this population, or probably also in other similar populations. We have now added details of the present pain ratings to make this background clearer, see Table 1.

Reviewer’s comments
3. Test protocol: the inter-observer variability is only measured for a part of the procedure (only the performance). The oral instruction and the demonstration is only done by the instructing observer and not by the second observer. Only the performance of the test by the subject is scored by both observers at the same time. Was the instructing observer the same in the test and re-test situation? Otherwise this may be an explanation of the low intra-observer variability.
Reply
Yes, the instructing observer was the same in the test and re-test situation. This limits the inter-observer reliability to test performance only, and has now been explicitly addressed in the Discussion, page 14, lines 4-6.

Minor essential revisions:

Reviewer’s comments
1. title: please mention that not only the inter- and intraobserver reliability is studied but also sensitivity and specificity.

Reply
Discriminative validity is a secondary aim, as now clarified in the manuscript, we believe such a title would possibly take too much focus from the primary aim. We have however added sensitivity and specificity as keywords.

Reviewer’s comments
2. Abstract
A: in background: not only sensitivity but also specificity is studied

Reply
We have now added a few words about specificity under Background, page 2, line 10.

Reviewer’s comments
B: results: the values of sensitivity and specificity are not mentioned.

Reply
Since values emerging for sens/spec were a secondary aim (now clearly defined), we elected to use a more general description of those results in the Abstract.

Reviewer’s comments
3. only 1 female is tested against 32 male, no point about this is made in the discussion.

Reply
The sample was found representative of the population, based on comparison with internal military statistics (not for public presentation), also regarding the gender ratio.

Reviewer’s comments
4. Results: in table 2 the mean kappa coefficient of test SB is missing

Reply
Thank you for pointing this out. This is now inserted.

We thank the reviewer for the comments and for the time spent reviewing our work.

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 have no competing interests in relation to this paper.