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.

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.

Methods

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

Minor Essential Revisions

Study sample:
How were the marines enrolled in the study, consecutively as they volunteered, randomised 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?

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.
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.

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.

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.

Results

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

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.

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?

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.

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

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