Author's response to reviews

Title: Inter-rater reliability of the evaluation of muscular chains associated with posture alterations in scoliosis

Authors:

Carole Fortin (carole.fortin@umontreal.ca)
Debbie E Feldman (debbie.feldman@umontreal.ca)
Clarice Tanaka (cltanaka@uol.com.br)
Michelle Houde (mhoude@santepub-mtl.qc.ca)
Hubert Labelle (hubert.labelle@recherche-ste-justine.qc.ca)

Version: 3 Date: 26 March 2012

Author's response to reviews: see over
March 2012

Doctor Dana Lawrence
Editorial member
BMC Musculoskeletal Disorders

Dear Dr Lawrence,

Thank you for giving us the opportunity to respond to your comments for our manuscript entitled “Inter-rater reliability of the evaluation of muscular chains associated with posture alterations in scoliosis”, reference number: MS: 6768700226277520. The main issue that appears to be problematic is that of validity. As the editor explained, agreement does not ensure establishing validity. We agree with this and have therefore removed the term validity from our manuscript and refer to agreement with experts. This was actually recommended by the reviewers. Since we are measuring agreement with experts, we used ICCs and not Pearson correlation coefficients.

Muscular chain impairment is evaluated by observation. Therapists are trained to do this evaluation, although there is no gold standard for comparison. Therefore, we chose to use expert evaluation (i.e. persons who are accredited to teach Global Postural Re-education - GPR) as our comparison group. This is not unique to GPR method, as *Sahrmann (2002) and *Kendall (2005) also use observation to evaluate muscle involvement in abnormal posture and movement patterns.

Regarding the relationship between muscular chain impairment and muscle tone, we responded to the reviewer that according to Souchard (the developer of GPR method) the pathophysiology of tonic muscles is muscle retraction whereas pathophysiology of phasic muscles is weakness. We prefer these terms (retraction or weakness) rather than hyper or hypo tonicity, since the latter imply neurologic involvement which is not at all the case in the present study of patients with adolescent idiopathic scoliosis.

We have responded to the reviewer who listed 14 major compulsory revisions (see attached). Further, we also made the required revisions in the text, as suggested. Regarding the two comments that the editor pointed out (numbers 2 and 4), we have responded as follows:

**Question 2):** Make a case for the validity of the muscle chain impairment construct. Include evidence for the existence of muscle chain impairment and association of posture patterns with muscle dysfunction patterns. Otherwise, there is no clear point for the current study.

**Answer:**
We modified the second paragraph of the *Introduction* section. Although there is a need for further evidence regarding the association between muscle chain impairment and posture, we have found two studies that support this: *Specific posture patterns caused by*
muscle chain retractions have been associated with lower back or neck pain among elite athletes in muscular power competitions [12] and functional disabilities in an adult with hemiparesis [13].

**Question 4):** Clinical relevance: The primary reliability analysis should be identified as that for the muscle chain impairment. Posture assessments are an intermediate step and reliability of any intermediate step is secondary.

**Answer:** Muscular chain impairment evaluation is an analytical way to evaluate the observed posture so that the therapist can develop the necessary treatment to address the problem. In order to determine the muscle chain impairment, one must first observe the posture pattern and then identify the muscles responsible for this pattern. In the methods, we state the following: “The first step was to note the presence or absence of posture alteration for each of the 23 posture indices. When the posture alteration was present on both sides (for example protracted shoulder), the PT had to indicate if the alteration was equal or greater on one side. Secondly, the PT had to determine if the posture alteration was attributable to retraction of muscles in the anterior muscular chain (1), posterior muscular chain (2), anterior and posterior muscular chains (3) or unable to be evaluated (4) (see Appendix for more details).”

We modified the order of presentation of the data in order to respond to the reviewer who preferred that the primary reliability analysis should be identified as that for muscle chain impairment.

Below, please find the responses to the 14 points brought up by the reviewer. The corrections in the manuscript are written in boldface font.

Thank you for considering our work.

Carole Fortin, Ph.D.
Corresponding author

*References:
Answers to the 14 major compulsory revisions:

Question 1). Patient selection: Identify the study population more completely. How were the five patients selected besides having scoliosis?

Answer:
The following information have been added under Participants in the Methods section: “Youths with idiopathic scoliosis were chosen because they typically display posture asymmetries [18,23]. These youths (4 females, one male) were recruited from a previous study on posture assessment performed at the Sainte-Justine University Hospital Center in Montreal. Mean age of youths was 15.6 ± 2.2 years and average weight and height were 53.6 ± 11 Kg and 161.9 ± 13.8 cm, respectively. Two youths had a right thoracic scoliosis (33° and 36°), two had a double major scoliosis (18°-15° and 23°-24°) and one had a left thoraco-lumbar scoliosis (38°). We selected youths with different morphological characteristics (anterior, posterior or mixed) with clear photographs.”

Question 2). Introduction: Make a case for the validity of the muscle chain impairment construct. Include evidence for the existence of muscle chain impairment and association of posture patterns with muscle dysfunction patterns. Otherwise, there is no clear point for the current study.

Answer:
We modified the second paragraph of the Introduction section to address muscle chain impairment construct. However, as already mentioned, although there is a need for further evidence regarding the association between muscle chain impairment and posture, we have found two studies that support this. The following sentences have been added in the Introduction section: “Souchard [6] describes muscles as being organized into two main static postural chains: the anterior and posterior muscular chains. Muscular chains are an ensemble of muscles defined according to their localization as well as their functional role which can explain posture alterations and movement dysfunctions [6,10-12]. Specific posture patterns caused by muscle chain retractions have been associated with lower back or neck pain among elite athletes in muscular power competitions [12] and functional disabilities in an adult with hemiparesis [13]. Despite the lack of studies linking muscular chain impairments to abnormal posture patterns and dysfunction, it seems that global muscular chain stretching is more effective than analytic muscle stretching to improve function and quality of life for several pathologic conditions including respiratory, musculoskeletal and neurological problems [7-9,13,14,15].”
**Question 3**). Expert agreement needs to be elaborated under methods, since there are two methods of measurement that are confounded with rater. This is also a limitation of the study.

**Answer:**
We added the following sentence under Procedure of Methods section to clarify this point: *"The two experts completed the same procedure. If there was disagreement between the two experts, they discussed their results to reach a consensus. If a consensus could not be reached, a third expert (CF) made the final decision. After consensus, agreement between PTs and experts was calculated."*

**Question 4**). Clinical relevance: The primary reliability analysis should be identified as that for the muscle chain impairment. Posture assessments are an intermediate step and reliability of any intermediate step is secondary.

**Answer:**
As already mentioned, in order to determine the muscle chain impairment, one must first observe the posture pattern and then identify the muscles responsible for this pattern. We modified the order of presentation of the data in order to respond to the reviewer who preferred that the primary reliability analysis should be identified as that for muscle chain impairment. We did the modification in the Results section and we modified the Tables as well.

**Question 5**). Validity: The term implies agreement with a clinical entity identified by a gold standard. This study is looking more at agreement with expert performance. Do not use the term validity in the title or text, and make it clear that no true gold standard was used.

**Answer:**
As suggested by the two reviewers, we have replaced the term *validity* by *agreement* with experts throughout the text. We also added the following sentences under Participants in the Methods section to make clear that no true gold standard was used: *"Two physical therapists instructors in GPR served as experts for determining muscular chain impairments associated with posture alterations, in the absence of any objective "gold standard" criterion for this assessment. Muscular chain impairments were determined by the two experts according to standards taught in GPR which can be found in GPR literature [6,7,13,14].”*

**Question 6**). The "no answer" category: The problem with this category is that if it predominates, then you can have a high reliability/validity index that merely indicates muscle chain impairment is indeterminate. This must clearly be explained, since you have highly positive findings that may support that lack of clinical usefulness of the index.
This needs to be addressed separately from the reliability-validity of muscle chain impairment categorization.

**Answer:**

We agree with the reviewer that addressing this point separately will be clearer for the reader. We added a specific section entitled “Muscular chain impairment associated with posture alterations”. We also modify this paragraph p.11-12: “The muscular chain could be determined for the majority of posture alterations (Table 1, fourth column). However, both the physical therapists and the two experts did not attribute specific muscular chain impairment for head lateral bending, head rotation and knee flexion. No muscular chain impairment has been identified for elevated shoulder and adducted scapulae since no such asymmetries were reported in the five cases that were evaluated and the term “no answer” is thus written in the table under these posture indices.”

We also added under Study Limitations in Discussion section, the following sentence: “Moreover, some posture alterations were not present among these youths and thus muscular chain impairment could not be determined.”

**Question 7).** No known meaning of the global mean score is given. This ad hoc index should be excluded.

**Answer:**

The global mean score have been removed.

**Question 8).** Kappa: Identify the kappa used more clearly. It is a generalized kappa for multiple raters and multiple categories. The reliability derivation in the appendix is unnecessary. Those who are interested will have the reference or can go to the cited online site. Make it clear that the Wikipedia is presenting the Fleiss kappa.

Take into consideration the possible lack of heterogeneity of the data and resulting kappa instability (kappa can vary wildly in this case).

**Answer:**

We better identified the Kappa used in our study under Data analysis of the Methods section. Please see the following sentences: “We used Fleiss’ Kappa coefficients (for categorical data) and percentage of agreement to assess inter-rater reliability of muscular chain evaluation and associated posture alterations (objective 1).” and “Fleiss’s Kappa is used to measure the overall agreement between several raters and is adapted for nominal scales with multiple categories [28-30].”

This appendix has been removed.

**Question 9.** Intraclass Correlation Coefficient: The ICC used is not identified. There are many of them with different applications. It is not clear if categorical data were used or some index vaguely alluded to in the text. It is not unknown how the average expert rating and PT ratings were included in the ICC used. An ICC is often a wrong choice for a validity statistic; commonly used ICCs are measures of agreement among equals (equal error variances assumed for all raters). This does not apply to experts assumed to have superior skill to the study population of PTs. Look for appropriate indices that allow for error-free measures such as likelihood ratios at Pearson’s r (through regression analysis which assumes the independent variable is error-free in its computation).

**Answer:**
We used ICC and not Pearson correlation coefficient because ICC is considered similar to coefficient of concordance for assessing agreement with experts and also because ICC is well applicable in a reliability study using a two categorical data model (agree or disagree) for a large number of raters. We added the following explanation under *Data analysis in Methods section* (paragraph 2): “For our third objective, we examined agreement between PTs and experts regarding muscular chain evaluation and their associated posture alterations, using intra-class correlation coefficients (ICC$_{3,k}$) for categorical data (agree or disagree). PTs’ answers were re-coded as agree versus disagree with experts’ answers (after consensus) and were averaged for each possible choice of posture indices of the grid (example: right knee flexum – R, left knee flexum – L, R > L or R < L; see Appendix).”

**Question 10.** Table 2. You need to explain how the PTs can agree with experts well but not with each other. If raters do not agree with each other well, then some cannot agree with any standard such as experts. This could be a problem with the choice of statistics or how it is applied.

**Answer:**
The difference comes from the impact of the number of categories taken in kappa and ICC analysis and because kappa coefficients are less favourably influenced by a large number of raters than are ICCs. As already mentioned, the PTs’ answers were re-coded as agree versus disagree with experts’ answers (after consensus) and were averaged for each possible choice of posture indices of the grid (79 items representing the 23 posture indices; example: right knee flexum – R, left knee flexum – L, R > L or R < L). In kappa analysis, all the physical therapists had to choose all of the same choices (2 to 4 choices) for each of the indices to have perfect agreement.

Please see the following explanation in *Study Limitations in Discussion:*
“Another limitation concerns the number of categories for each posture index and for muscular chain evaluation in our assessment scale. Increasing the number of categories in a measurement scale decreases the Kappa coefficients [28]. The analysis for the posture indices was done by combining two to four elements for each of the 23 indices. For example, in the case of knee flexum index, the PT had four choices on the grid (right knee flexum – R, left knee flexum – L, R > L or R < L; see Appendix). In order to have
perfect agreement, all the physical therapists would have had to choose all of the same choices (2 to 4 choices) for each of the indices. When we compare the grid for the two experts, there are only a few differences. For the 79 items (representing the 23 posture indices) the two experts are in perfect agreement for 59 of the 79 items; there is a difference for one youth on 17 items and on 3 items for two youths. Thus we have been much more conservative by choosing to analyze the grid by 23 posture indices instead of 79 individual items, and our coefficients are considerably lower as a result. Moreover, Kappa coefficients are less favourably influenced by a large number of raters than are ICCs. The Kappa coefficient is a conservative measure since it eliminates agreement by chance. This explains a Kappa coefficient of zero even if the two experts agreed for three youths out of five and had a percentage of agreement of 60% (Table 2).”

**Question 11).** The Discussion seems to stray a bit from reliability/validity to treatment strategy of the GPR technique. The discussion should focus on the level of reliability/validity and the relevance to practice.

**Answer:**
We modified the Discussion accordingly.

**Question 12).** Limitation: The study looked at the reliability of reading a photograph. It did not take into consideration positioning error. This would have required a second photograph.

**Answer:**
We agree that positioning can affect the perspective and yield to errors. However, the photographs of all youths were taken in the same position and distance from the camera. Also, all PTs and experts made their observation from the same set of photographs (7) for each youth.

**Question 13).** Limitation: There is no small sample standard error for a generalized kappa to my knowledge that can be used to perform a t-test. There is no standard error for constructing a confidence interval.

**Answer:**
We agree with the reviewer. We performed a chi-squared test on posture indices with Kappa coefficient ≥ 0.40 to determine if there was a significant difference between the 3 groups. Please see the following sentences in Data analysis in Methods section:

“To address our second objective, we divided the PTs into three groups according to their experience in GPR (Group 1: ≤ 2 years, Group 2: 2.5 to 9.5 years and Group 3: ≥ 10 years) and the analysis of the muscular chain associated with posture alterations was made for each of the three groups. We determined whether there were differences between the 3 groups for Kappa coefficients ≥ 0.40, using the chi-squared test (χ²).”
Question 14). Conclusion: It is not clear that reliability is greater with experience. No significance testing was conducted comparing independent kappa values between the three experience subgroups. It is apparent that experts are more reliable than non-experts for posture evaluation.

Answer:
As mentioned in point 13, we added a chi-squared test to verify this second objective and found that Group 3 has a significantly higher level of reliability than the other groups ($\chi^2$, $p=0.005$). This sentence has been added in Results section:

“Group 3 has a significantly higher level of reliability than the other groups ($\chi^2$, $p=0.005$).”