Reviewer's report

Title: Inter-rater reliability and validity of the evaluation of muscular chains associated with posture asymmetries

Version: 1 Date: 17 January 2012

Reviewer: Mitchell Haas

Reviewer's report:

This reliability/validity study has a unique design, including a sample of 50 physical therapists and only five subjects. Most published reliability studies use only a few raters to evaluate a considerably larger subject sample. A positive for the study is the better generalizability to the provider population, but this is at the expense of generalizability to the patient population. Appropriately, a generalized kappa was used to compute overall reliability coefficients. The authors did their homework on reliability.

Of primary concern is the issue of validity. The case for the existence of the muscle chain impairment is not clear in the introduction. Of particular interest is the assumption that posture patterns are related to hypertonicity and not hypotonicity. Agreement with a consensus of two experts does not ensure the validity of identifying the proper muscle chain impairment, unless the skill of the experts was first validated by a real gold standard. Evidence of clinical relevance of the muscle chain impairments would go a long way in supporting the relevance of this study. In addition, the method for computing the validity statistic is not included, both the intraclass correlation coefficient used and how it was applied to the expert and clinician data.

Reliability and validity are only meaningful when applied to clinically relevant entities. The primary reliability investigation is that for the “muscle chain impairments.” The text says that the reliability of “muscle chain impairments” was computed only for cases of agreement on posture asymmetry. It is not clear how this was implemented, and all ratings should have been included in the analysis at any rate (exception noted below for indeterminate categorization). The small sample may also make kappa ineffective in some cases; there could be a problem with homogeneity (restriction of range across categories of categories (underrepresented categories) leading to kappa instability. This may explanation of how PTs appear to agree with experts but not with themselves.

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

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.

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.

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.

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.

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

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).

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).

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.

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.

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
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.

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.

Minor Essential Revisions:
1. None.

Discretionary Revisions:
1. Title: Add in scoliosis. The generalizability is limited to this population.

**Level of interest:** An article of limited interest

**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.