PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Inter-tester reliability of clinical shoulder instability and laxity tests in subjects with and without self-reported shoulder problems
AUTHORS	Eshoj, Henrik; Ingwersen, Kim Gordon; Larsen, Camilla; Kjær, Birgitte; Juul-Kristensen, Birgit

VERSION 1 – REVIEW

REVIEWER	Nuno Sampaio Gomes Orthopaedics Department. Hospital Forças Armadas - Polo do Porto. Portugal
REVIEW RETURNED	01-Aug-2017

GENERAL COMMENTS	L 83: ...and laxity are diagnosed... L 111: Ref 24 is from Int J Nursing Studies. The same article has been published in J Clin Epidemiol. 2011 Jan;64(1):96-106. doi: 10.1016/j.jclinepi.2010.03.002. Epub 2010 Jun 17. I believe that would suit better considering the authors background and the type of research. L 118, 119, 122-3, 135, 140, 245: "shoulder cases" and "shoulder healthy" should be renamed. Suggestion: "affected shoulders" and "normal shoulders". L 131 ...(women and men, aged 18-60 years)... L 149 ...non-invasive / non-treating... L 157 ...whereas relief... L 172,3 ...Gagey tests... sign tests... Table 1: Sulcus sign description is not accurate as to whether external or internal rotation of the arm is used. Performing it in different rotations may have contributed to variations in the results. The same for the Gagey. The scapula is stabilized and prevented from elevating, not exactly depressed. Measurement is done as the scapula starts moving. Not "evalate" but "evaluate". L 194 Furthermore, both... L 241 "equivalent" instead of "equivocal"? L 254 ...also found for... L 256 would change "respectively" for "i.e." L 262,4 Please rephrase last sentence of the paragraph. Interesting study and definitely needs revision by a statistics specialist. That would confirm the results, namely the validity of the conclusions as it seems to be a small study group.

REVIEWER	Robert Manske Wichita State University
REVIEW RETURNED	17-Aug-2017

GENERAL COMMENTS	Thank you for allowing me to review this manuscript. I believe that reliability of various shoulder instability tests are very important so that as practicing clinicians we are able to more accurately determine consistently over time that a given condition is what we think it is. The introduction clearly describes the problem at hand and how physical exam tests are done for shoulder instability pathology. The last sentence clearly articulates what the objective of the study is. Improvements to this section could include clearly describing the study hypothesis. The methods section describe a strength of the study is that if followed GRASS guidelines. I am not familiar with those guidelines and am guessing most readership do not. A description of how those were developed and what the guidelines are may be nice. This might be able to be included as an appendix. I do commend the authors for blinding the testers from the whether the subject had instability or not. That is a positive about the study. The sample size seems large enough for a reliability study. IRB approval and oral and written consent is clearly described. The tables are a nice representation of the procedures and how the testing was performed. Results clearly describe what was found during testing. I am not a statistician by trade so not 100% sure what PABAK is telling me. It appears that PABAK takes into account some normal variation and will give a more favorable score than standard the kappa calculation? Is that generally the case? Is PABAK generally used for this type of study? If not why did you decide to use it here? Is there a potential it would artificially inflate the reliability scores? The discussion details each finding. However, it relies very heavily on comparing the Tzannes 2004 study of reliability. Is that the only study that has been done on similar shoulder reliability tests? This study did not describe limitation of using manual techniques for testing. The testers could have given more pressure during testing of each individual. There was not standard reference amount of pressure used during testing. Using a standardized force could have helped increase reliability. This may seem to be clinically unacceptable – however would have helped standardize the testing methods used.

REVIEWER	Kevin Plancher USA
REVIEW RETURNED	21-Aug-2017

GENERAL COMMENTS	The following recommendations could strengthen this study. Please review the manuscript for grammatical errors. Abstract • Page 2, Line 40 – How did the authors determine their sample size? • Page 2, line 41 – Please describe the inclusion criteria for each group. • Page 2, Lines 45-47 – Please clarify the meaning of this sentence. Are you suggesting that this battery of tests in conjunction with patient symptoms has the highest positive predictive value? • Page 3, line 51 – How will the results of this study improve clinical practice if “further standardization” is needed? Additionally, if the authors haven’t established validity of the tests, then what is the importance of establishing reliability? Furthermore, the reliability of these tests has been previously established in the literature, however, the authors do not sufficiently develop the rational and need for the present study. • Page 3, line 61 – Why couldn’t a 50/50 prevalence of positive and negative test be accomplished? Introduction • Page 4, Lines 87-89 – Please provide a reference for this sentence. • Page 5, Lines 81-93 – The relevance of this paragraph is suspect. Inclusion of pain was not a diagnostic factor in the present study. Furthermore, this study relates to anterior instability but the methodology doesn’t discriminate the type of shoulder instability in question, which may contribute to inconsistencies in the results. • Page 5, lines 96 - 99 – What do the authors expect to find in conducting this study? How will this study add to the literature on clinical shoulder instability and laxity? • Page 5, line 102 – How was sports-active individuals defined? Methods • Page 6, line 113-117 – If the first phase of the study was to train the 2 physiotherapist to be mutual agreeable in performing and interpreting each of the tests, then it is unclear how the results of the study would be applicable to the general clinical setting. It appears that you trained the clinicians to be reliable before conducting the study. The physiotherapists, as described, are novice clinicians and perhaps the purpose of the study should be related to inexperienced clinicians. Reliability of these tests in the hands of trained clinicians perhaps is drastically different. The rationale for the study should be developed further. • Page 6, line 117 – How was sample size selected? Why was there an unequal distribution of subjects? • Page 6, Lines 120-121 – Given that there was a minimum level of agreement before moving into the final study phase, why do the authors believe there wasn’t more agreement between the raters? It would seem plausible that this would be due to the lack of experience of the raters calling into question the clinical meaningfulness of the data. • Page 6, Lines 108-125 – What was the time period between each of the testing periods? Were the subjects used in the preparation and training phase, overall agreement phase, and study phase the same or different subjects?

	• Page 7, lines 136-137 – It is unclear how asking subjects these two questions and only requiring a yes in 1 of the 2 would allow you to obtain your injured subjects of people with instability and/or laxity related should problems. It is plausible that one could have a shoulder injury without shoulder instability but they would still fall within this category. • Page 7, Lines 142-148 – What is the relevance of collecting these data and how were they used overall in establishing interrater reliability? • Page 8, Lines 153-155 – Was the order of the clinical exam test the same or different between raters and between subjects? Was the order randomized? • Page 8, Lines 160-162 – The reasoning behind the decision to report only the direction with the most glenohumeral translation must be developed further. As written, it appears the decision was only to enhance the study results which is suspect. • Page 8, Lines 162-163 – Please report the measurement error of the ruler. Results • Page 13, Table 2 – It is unclear why patients in the healthy shoulder group had pain in their shoulder at rest and with activity. Please clarify. Furthermore, one subject had a previous shoulder injury and 3 had subjective shoulder instability. Please clarify how these patients were included in the healthy group. • Page 15, Lines 226-228 – Please clarify the meaning of this sentence. How were these the “most frequently used”? • Page 15, Table 5 – Please describe what the * denotes for Table 5. Discussion • Page 16, Line 254 – Found is spelled incorrectly. • Page 17, Lines 265-266 – Please discuss what is meant by “current poor prevalence index’. • Page 18, Lines 280-282 – As previously mentioned, this warrants further discussion. • Page 18, Lines 293-295 – This doesn’t appear to be a limitation but rather a nonsignificant finding. Furthermore, it appears that prior to the final phase of the study, the investigators had stronger agreement. Please clarify. • Page 18, Lines 297-298 – While the patients may have shoulder problems, it is unclear whether the injured group had instability symptoms. • Page 19, line 300 – Please clarify why the authors believe that determining reliability before establishing validity is clinically meaningful. What is the impact of the study results to the practicing clinician. Perhaps the investigators are reliably unreliable. Without establishing the diagnostic accuracy including PPV and NPV, the study results and meaning is questionable. Conclusion • Page 19, line 318 – How will the findings of the study impact clinical practice? References • 81% of the references are greater than 5 years old.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name
Nuno Sampaio Gomes

Institution and Country
Orthopaedics Department.
Hospital Forças Armadas - Polo do Porto.
Portugal

1)
Please state any competing interests or state ‘None declared’:

Answer+Action: None declared

2)
Please leave your comments for the authors below
L 83: ...and laxity are diagnosed...

Answer: Thank you for this grammar correction.

Action: Line 89, changed to: ‘..and laxity are diagnosed’

3)
L 111: Ref 24 is from Int J Nursing Studies. The same article has been published in J Clin Epidemiol. 2011 Jan;64(1):96-106. doi: 10.1016/j.jclinepi.2010.03.002. Epub 2010 Jun 17. I believe that would suit better considering the authors background and the type of research.

Answer: Thank you for this very relevant comment.

Action: Line 127/reference list: Reference no 24 changed according to reviewer´s suggestion

4)
L 118, 119, 122-3, 135, 140, 245: "shoulder cases" and "shoulder healthy" should be renamed. Suggestion: "affected shoulders" and "normal shoulders".

Answer: Thank you for this very relevant suggestion.

Action: Throughout the manuscript: “Shoulder cases” changed to “affected shoulders” and “shoulder healthy” changed to “normal shoulders”.

5)
L 131 ...(women and men, aged 18-60 years)...

Answer: Thank you. Individuals inserted.

Action: Line 155; ‘Sixty-five individuals (women and men (aged 18-60 years)) were…’

6)
L 149 ...non-invasive / non-treating...

Answer: Thank you very much. “/” replaced with an “and”

Action: Line 186: ‘due to the non-invasive and non-treating study design’

7)
L 157 ...whereas relief...

Answer: Thank you very much for this grammar correction.

Action: Line 195; “relieve” changed to ‘relief’

8)
L 172,3 ...Gagey tests... sign tests...

Answer: Again. Thank you very much for this grammar correction. An ‘s’ is added to “test” for plural

Action: Line 217, 218: Apprehension, relocation, surprise and Gagey tests were dichotomous variables whereas the load and shift and sulcus sign tests were…

9)
Table 1: Sulcus sign description is not accurate as to whether external or internal rotation of the arm is used. Performing it in different rotations may have contributed to variations in the results.
The same for the Gagey test. The scapula is stabilized and prevented from elevating, not exactly depressed. Measurement is done as the scapula starts moving. Not "evalate" but "evaluate".

Answer: Very relevant comment. Thank you for that. Test descriptions of sulcus sign and Gagey tests are changed according to reviewer´s suggestion.

Action: Table 1, page 11:
Sulcus sign test changed to:
Individual sitting upright. Shoulder in neutral position (0 degree rotation)….

Distance from the top of the humeral head and the acromion is evaluated with a ruler.

Gagey test changed to:
The shoulder girdle is stabilized by examiners forearm preventing the shoulder girdle to elevate while the individuals arm is passively moved into end range in horizontal abduction.

10)
L 194 Furthermore, both...

Answer: Thank you very much for this grammar correction. Further replaced with furthermore,

Action: Line 265: ‘Furthermore, both groups..’

11)
L 241 "equivalent" instead of "equivocal"?

Answer: Thank you very much for this grammar correction. Equivocal replaced with

" equivalent”

Action: Line 325; ..or equivalent to,…

12)
L 254 ...also found for...

Answer: Thank you very much for this grammar correction. Fund replaced with found.

Action: Line 338: …was also found for the…

13)
L 256 would change "respectively" for "i.e."

Answer: Thank you. We agree that this sentence could be misunderstood. Therefore, the sentence is changed.

Action: Line 346: force produced to translate the humeral head in either posterior (relocation test) or inferior (sulcus sign test) direction...

14)
L 262,4 Please rephrase last sentence of the paragraph.

Answer: Agree. Sentence changed according to reviewer´s suggestion.

Action: Line 352: However, due to the presence of systematic bias in both the relocation and sulcus sign test, PABAK did not affect the overall reliability much.

Interesting study and definitely needs revision by a statistics specialist. That would confirm the results, namely the validity of the conclusions as it seems to be a small study group.

Reviewer: 2
Reviewer Name
Robert Manske

Institution and Country
Wichita State University

1)
Please state any competing interests or state ‘None declared’:

Answer+Action: None declared

Please leave your comments for the authors below
Thank you for allowing me to review this manuscript. I believe that reliability of various shoulder instability tests are very important so that as practicing clinicians we are able to more accurately determine consistently over time that a given condition is what we think it is.

The introduction clearly describes the problem at hand and how physical exam tests are done for shoulder instability pathology. The last sentence clearly articulates what the objective of the study is.

1)
Improvements to this section could include clearly describing the study hypothesis.

Answer: As is usual for reliability studies, the study hypothesis will be that reliability is satisfactory in a standardized design with standardized procedures. However, hypotheses are usually not part of reliability studies.

Action: No changes made.

2)
The methods section describe a strength of the study is that if followed GRASS guidelines. I am not familiar with those guidelines and am guessing most readership do not.

A description of how those were developed and what the guidelines are may be nice. This might be able to be included as an appendix.

Answer: Thank you for pointing this out. A short description of GRASS is inserted to the manuscript.

Action: Line 126: …a consensus document on how to report reliability and agreement studies…

I do commend the authors for blinding the testers from the whether the subject had instability or not. That is a positive about the study. The sample size seems large enough for a reliability study. IRB approval and oral and written consent is clearly described. The tables are a nice representation of the procedures and how the testing was performed.
Results clearly describe what was found during testing. I am not a statistician by trade so not 100% sure what PABAK is telling me.

3)
It appears that PABAK takes into account some normal variation and will give a more favorable score than standard the kappa calculation? Is that generally the case? Is PABAK generally used for this type of study? If not why did you decide to use it here? Is there a potential it would artificially inflate the reliability scores?

Answer: Thank you for this very relevant comment. PABAK is used where the distribution between positive and negative test results is not equal. Using PABAK makes the reliability of the tests independent of the included sample provided the inclusion criteria are representative for the target population. When testing reliability of 6 clinical tests it is impossible to obtain a 50% distribution of the prevalence (which is the optimal study condition) for each of the clinical tests, why the optimum way of handling reliability statistics is to use PABAK (Sim and Wright 2005). Also, for transparency reasons, PABAK need to be reported along with the kappa coefficients, as is also done in this manuscript.

Action: In the statistics section the following has been inserted:

Line 225: in prevalence and bias…(e.g. if a 50/50 distribution of positive and negative tests cannot be accomplished) the use of PABAK calculation is a valid supplement to the original kappa values.

Line 228: PABAK calculation is performed by adjusting for high or low prevalence by computing the average of cells a and d in a cross table, substituting this value for the actual values in those cells. Similarly, an adjustment for bias is achieved by substituting the mean of cells b and c for those actual cell values (Sim and Wright 2005).

The discussion details each finding. However, it relies very heavily on comparing the Tzannes 2004 study of reliability.

4)
Is that the only study that has been done on similar shoulder reliability tests?

Answer: Very relevant comment. However, we have strived to include the newest literature; and, to our opinion, this is the newest reference of relevance to the current study, that could be found in relevant databases.

Action: None.

5)
This study did not describe limitation of using manual techniques for testing. The testers could have given more pressure during testing of each individual. There was not standard reference amount of pressure used during testing. Using a standardized force could have helped increase reliability. This may seem to be clinically unacceptable – however would have helped standardize the testing methods used.

Answer: Thank you for pointing this out. In the current manuscript, we have already addressed this point in the discussion section (line 317-320) regarding the sulcus sign test as one of the reasons for the low reliability. And this could also apply for the remaining tests. Therefore, a general sentence has now been inserted in the limitation section to further address this.

Action: Line 393: ‘‘Firstly, the lack of standardized measurement of the amount of force exerted by the two testers during especially the relocation and sulcus sign test may have limited the inter-tester reliability in the current study.

Reviewer: 3
Reviewer Name
Kevin Plancher

Institution and Country
USA

1)
Please state any competing interests or state ‘None declared’:

Answer+Action: None Declared

Please leave your comments for the authors below
The following recommendations could strengthen this study.

2)
Please review the manuscript for grammatical errors.

Answer+Action: An English native speaker has now corrected the paper.

Abstract

3)
Page 2, Line 40 – How did the authors determine their sample size?

Answer: Thank for this very relevant question. The reference for designing reliability and validity studies, as used in many studies, suggests about 10 subjects in the training phase, 20 subjects in the agreement phase and about 40 subjects in the study phase (Patijn J 2004). The sample size is a pragmatic suggestion, which in other reliability studies of several clinical tests have shown to be adequate for performing satisfactory reliability studies (Juul-Kristensen et al. 2007) (Vind et al. 2011).

Action: None

4)
Page 2, line 41 – Please describe the inclusion criteria for each group.

Answer: Thank you very much for pointing this out. However, due to abstract limitations we have modified the sentence to meet the reviewer comment.

Action: Line 41: with self-reported shoulder instability and laxity…'

5)
Page 2, Lines 45-47 – Please clarify the meaning of this sentence. Are you suggesting that this battery of tests in conjunction with patient symptoms has the highest positive predictive value?

Answer: Thank for pointing this out. The primary aim was not to study predictive values. However, the aim was to study reliability and mutual dependency, meaning the highest frequency for each tester to characterize self-reported shoulder instability conditions. In the statistics section, line 183-185, mutual dependency is defined and described how it was calculated: 'The relationship between the individual tests and the classification (mutual dependency) by self-reported shoulder problems was tested by Cohen´s kappa (k) coefficients…'

Action: Line 48 …characterize self-reported shoulder instability conditions.'

Line 231: The relationship for each tester between the individual tests and the classification (mutual dependency) of self-reported shoulder instability and laxity was tested by Cohen´s kappa (k) coefficients and the characterization of the groups was tested with Fischer's exact tests.

6)
Page 3, line 51 – How will the results of this study improve clinical practice if "further standardization" is needed? Additionally, if the authors haven't established validity of the tests, then what is the importance of establishing reliability? Furthermore, the reliability of these tests has been previously established in the literature, however, the authors do not sufficiently develop the rational and need for the present study.

Answer: Thank you for this very relevant comment. Reliability has previously been established, however, with large variations in results, and with limited methodological quality – as also described P5, Line 104-107. Also, as reliability is a necessary first condition for a measurement to be considered valid and responsive to change, we believe that reliability is the most relevant aspect to address firstly (Mokkink et al. 2010). However, we very much agree with the reviewer that the next step in the process of improving clinimetric properties is to study validity, which has also been addressed in the conclusion section (Page 20, Line 403).

Action: None

7)
Page 3, line 61 – Why couldn't a 50/50 prevalence of positive and negative test be accomplished?

Answer: Very relevant question. When testing reliability of 6 clinical tests for shoulder instability it is impossible to obtain 50% positive results for all tests. We aimed to include 50% patients with self-reported shoulder instability, thereby assuming 50% positive results. However, as patients not necessarily tests positive on all clinical tests this was not possible. As also described in the manuscript line 411.

Action: None.

Introduction

8)
Page 4, Lines 87-89 – Please provide a reference for this sentence.

Answer: Thank you for pointing this out. Sentence revised and references moved

Action: Line 95: References 14,15,16 now inserted at the end of the sentence.

9)
Page 5, Lines 81-93 – The relevance of this paragraph is suspect. Inclusion of pain was not a diagnostic factor in the present study. Furthermore, this study relates to anterior instability but the methodology doesn't discriminate the type of shoulder instability in question, which may contribute to inconsistencies in the results.

Answer: Thank for your commenting on this matter. The study was not only addressing anterior shoulder instability, but self-reported shoulder instability in general, as also described by the question and clinical tests according to the inclusion criteria (line 146):. Cases answering yes to at least one of two questions ('Do you have a sense of shoulder instability?' and 'Have you ever had a shoulder injury?') were eligible for a clinical shoulder examination. Cases were then included if they present with at least one positive clinical shoulder test out of the following; apprehension, relocation, surprise, load-and-shift, sulcus sign or Gagey.

Action: None, since inclusion criteria are described in line 156.

10)
Page 5, lines 96 - 99 – What do the authors expect to find in conducting this study? How will this study add to the literature on clinical shoulder instability and laxity?

Answer: Thank for pointing out this very relevant aspect. Due to few studies and poor methodology this study adds with high quality results by a three-stepped design for reliability studies, and a clear descriptions of the 6 selected tests for shoulder instability. It highlights that 4 of the clinical tests (apprehension, surprise, load-and-shift and Gagey) can be reproduced with satisfactory reliability, while the remaining 2 clinical tests (relocation and sulcus sign tests) still need to be further standardized. Furthermore, it addresses the importance of studying the validity in future studies.

Action: None – since this has already been addressed in the conclusion section, line 396

11)
Page 5, line 102 – How was sports-active individuals defined?

Answer: Thank you for addressing this relevant point. As already described in the manuscript Sports activity was self-reported by questioning the subjects about how many hours/week they were performing sports-related activity. However, the information was used only to describe the included population.

Action: None.

Methods
12)
Page 6, line 113-117 – If the first phase of the study was to train the 2 physiotherapist to be mutual agreeable in performing and interpreting each of the tests, then it is unclear how the results of the study would be applicable to the general clinical setting. It appears that you trained the clinicians to be reliable before conducting the study. The physiotherapists, as described, are novice clinicians and perhaps the purpose of the study should be related to inexperienced clinicians. Reliability of these tests in the hands of trained clinicians perhaps is drastically different. The rationale for the study should be developed further.

Answer: Thank you for addressing the very important aspect of performing clinical reliability studies. However, we believe that a well-defined and standardized protocol should make it possible to follow by both novice and experienced clinicians. Also, there are no consensus whether experienced clinicians are more reliable in following such protocol; in fact some studies have shown the opposite, that experienced clinicians have more difficulties in following a standardized protocol since they are unable to ignore their general clinical experience when performing these tests (Remvig et al. 2009). Furthermore, by using a 3-stepped design, as in the current study, the subjective evaluation and performance procedures (variance between individual raters) should be eliminated, and thus only the reliability of the clinical tests will be evaluated, as also described in the guidance document for performing clinical reliability studies (Patijn J 2004) as also referred to in the manuscript.

Action: None.

13)
Page 6, line 117 – How was sample size selected? Why was there an unequal distribution of subjects?

Answer: Very relevant question. As also described in the answer to question 3 (reviewer 2, page 6 in this document), the study is a strict 3-phased study following guidance for clinical reliability studies (Patijn J 2004) and the suggested sample size is a pragmatic approach, which has been satisfactory in many previous reliability studies. Regarding the skewed distribution of subjects, this is a clear limitation, since we opted for a 50/50 distribution. However, due to a relative short recruitment period besides difficulties in recruiting subjects with shoulder instability and laxity only thirteen subjects with an affected shoulder were included. Naturally, this also affected the prevalence of positive and negative test findings. However, to overcome this, PABAK calculation was used and reported along with kappa, to show transparently how data would have been with equal distributions of positive and negative test results.

Action: Inserted in discussion, limitation section:

Line 427: due to a relative short recruitment period besides difficulties in recruiting subjects with shoulder instability and laxity only thirteen subjects with an affected shoulder were included. Naturally, this also affected the prevalence of positive and negative test findings meaning that the prevalence of 0.50 in reliability studies in all six tests was not accomplished. However, to overcome this, PABAK calculations was used and reported along with kappa, to show transparently how data would have been with equal distributions of positive and negative test results.

14)
Page 6, Lines 120-121 – Given that there was a minimum level of agreement before moving into the final study phase, why do the authors believe there wasn't more agreement between the raters? It would seem plausible that this would be due to the lack of experience of the raters calling into question the clinical meaningfulness of the data.

Answer: Very relevant point. Thank you for that. However, this was why a training phase was used together with a thorough description of each test and how to interpret the clinical findings of each test. Further, as also described above, experienced clinicians may be unwilling to comply with a strict protocol and may be biased about their clinical findings. Therefore, novice clinicians were used to avoid this. Also, one can not expect that inter-tester agreements are improved from phase two to three, since there is no further training in performing and interpreting the tests between phase two and three. As also shown in table 4, agreements >80% were reached in four out of six tests, except for the relocation and sulcus sign tests, which proved to be statistically significantly affected by inter-tester difference. Thus, further standardization of these tests is needed as already described in the manuscript line 377.

Action: None

15)
Page 6, Lines 108-125 – What was the time period between each of the testing periods? Were the subjects used in the preparation and training phase, overall agreement phase, and study phase the same or different subjects?

Answer: Thank you for pointing this out. The time period between each test period was approximately 2 weeks, and new subjects were included for each phase. However, only the test phase (the actual reliability study) is reported in the current manuscript.

Action: Line 173: The time period between each test phase was approximately 2 weeks, and new subjects were included for each phase. However, only the study phase is reported in the current manuscript.

16)
Page 7, lines 136-137 – It is unclear how asking subjects these two questions and only requiring a yes in 1 of the 2 would allow you to obtain your injured subjects of people with instability and/or laxity related should problems. It is plausible that one could have a shoulder injury without shoulder instability but they would still fall within this category.

Answer: Thank you for this important comment. As is well known, at this moment there is no gold standard on how to classify shoulder instability. This is a general limitation for this patient population and this has also been addressed in the discussion section as one of the study limitations (line 378). It is anticipated (Dodson and Cordasco 2008) that having had a shoulder injury may change some of the shoulder structures thereby imposing instability or laxity, which in some situations may not directly be experienced by the subject.

However, the inclusion criteria were not only self-reported but also based on objective tests for instability, since besides answering yes to at least one of the two questions, subjects were also clinically examined. To be included as a case one had to have at least one positive test out of the six clinical instability and laxity tests .

Action: None

17)
Page 7, Lines 142-148 – What is the relevance of collecting these data and how were they used overall in establishing inter-tester reliability?

Answer: Thank you for addressing this aspect. These data were used to describe the population to be able to compare this study group with other studies for general representativity.

Action: None.

18)
Page 8, Lines 153-155 – Was the order of the clinical exam test the same or different between raters and between subjects? Was the order randomized?

Answer: Thank you for pointing out this very relevant issue. The order of tests was always the same and not randomized. This is a clear limitation and has now been addressed in the study limitations in the discussion.

Action: Line 397: Also, the current study did not randomize the order of the clinical tests. However, we do not believe this to have biased the reliability of the data, since the same order was used for both testers.

19)
Page 8, Lines 160-162 – The reasoning behind the decision to report only the direction with the most glenohumeral translation must be developed further. As written, it appears the decision was only to enhance the study results, which is suspect.

Answer: Very relevant point. Thank you. This phrase could be misunderstood. However, it was not the authors of this manuscript that chose only to report on the direction with most glenohumeral translation. It was the test procedure of the load and shift test that was reduced to being only reporting of the direction with most laxity. Therefore, examiners did only note the laxity of e.g. the anterior direction if this was the direction with most laxity.

Action: Line 199, Sentence reworded: ‘only the direction (anterior vs. posterior) with most glenohumeral head translation was scored.’

20)
Page 8, Lines 162-163 – Please report the measurement error of the ruler.

Answer: Very relevant comment. Unfortunately, we did not note the exact distance between acromion and the humeral head. We used the ruler only, as also suggested by Bahk and colleagues (2007), to clarify whether the distance between the humeral head and acromion was above or below 1 cm

Action: None.

Results

21)

Page 13, Table 2 – It is unclear why patients in the healthy shoulder group had pain in their shoulder at rest and with activity. Please clarify.

Answer: Thank you for pointing this out. This was self-reported data, which may have affected the answers. When participants were asked whether they had “shoulder pain” during activity they may sense some activity-related soreness and, therefore, report this as pain. Nevertheless, since data are only used for describing the population in general, this information is not expected to influence reliability data as presented here.

Action: None.

22)

Furthermore, one subject had a previous shoulder injury and 3 had subjective shoulder instability. Please clarify how these patients were included in the healthy group.

Answer. Thank you for this very relevant comment. The reason may be that the healthy shoulder participants were recruited through public advertisement, and then asked through a telephone interview whether they had had or currently had any shoulder trouble at all. Patients who responded “NO” were invited to participate as subjects with normal shoulders. In the actual test phase, the subjects with normal shoulders were also asked to complete a baseline questionnaire regarding demographics and other shoulder-related questions (as shown in table 2). Apparently, three subjects answered yes to feeling subjectively shoulder instable and one to have had a previous shoulder injury. Since this is self-reported data, we do not know the real answer to why subjects answer to the first question did not comply with the baseline questionnaire on shoulder-related questions. One explanation may be that a subjective feeling of shoulder instability (as in the baseline questionnaire) is not always equal to perceiving any shoulder trouble (as asked during the telephone recruitment procedures). However, we have now addressed this in the discussion, limitation section.

Action: Study subjects: page 7, line 159: Individuals with normal shoulders were recruited through public advertisement and…

Discussion, limitations, page 19, line 412: Also, in the group with normal shoulders, one individual reported to have had a previous shoulder injury and three individuals reported subjective shoulder instability, which does not comply with the inclusion criteria for being regarded as shoulder healthy in the current study. At the clinical session, a self-reported questionnaire was completed regarding demographic data and historical information. Apparently, in the baseline questionnaire three shoulder healthy individuals answered yes to perceiving instability in their shoulder and one had had a previous shoulder injury, even though they all had reported no shoulder trouble during the telephone inclusion interview. However, as depicted in table 2, WOSI and pain scores in the group with normal shoulders seem not to be influenced severely by these four individuals. Also, re-calculations of demographic data and mutual dependency with the revised classification into affected/normal shoulders did not change the mutual dependency of the most frequently used tests for classification into affected/normal shoulders, and neither was kappa and demographics affected (data not shown).

23)

Page 15, Lines 226-228 – Please clarify the meaning of this sentence. How were these the “most frequently used”?

Answer: Very relevant. However, please look at our answer to question 5, where the terminology of mutual dependency is explained. As described, the aim was to study reliability and mutual dependency, meaning the highest frequency for each tester to characterize self-reported shoulder instability conditions. In the statistics section, line 220, mutual dependency is defined and described how it was calculated.

Action: Line 31, changed to: to describe the mutual dependency for each tester between the individual tests for identifying self-reported shoulder instability and laxity.

Line 110: changed to: 'Therefore, the objective of this study was to investigate the inter-tester reliability of commonly used clinical shoulder instability and laxity tests and secondly to describe the mutual dependency for each of the tester, in a group of sports-active individuals with and without self-reported shoulder problems. '

Line 231 has been changed to: 'The relationship for each tester between the…'

24)
Page 15, Table 5 – Please describe what the * denotes for Table 5.

Answer: Thank you for pointing this out. The * means that these tests are systematically biased.

Action: The meaning of the * has now been explained at the bottom of table 5.

Table 5, page 15: *Significant inter-tester differences

Discussion

25)
Page 16, Line 254 – Found is spelled incorrectly.

Answer: Thank you very much for this grammar correction.

Action: Line 288, …were found for...

26)
Page 17, Lines 265-266 – Please discuss what is meant by "current poor prevalence index'.

Answer: Thank you for this comment. To clarify what is meant the sentence has now been changed

Action: Line 356: …due to the current low prevalence index below 50%, which is the optimum prevalence in reliability studies….

27)
Page 18, Lines 280-282 – As previously mentioned, this warrants further discussion.

Answer: Very relevant. However, please look above at our answer to this same point in question 5 and 23

Action: none.

28)
Page 18, Lines 293-295 – This doesn't appear to be a limitation but rather a non-significant finding. Furthermore, it appears that prior to the final phase of the study, the investigators had stronger agreement. Please clarify.

Answer: Thank you for pointing this out. You are right, this is not a study limitation, but a non-significant finding. Sentence therefore deleted from the manuscript.

Action: Line 370 and the following has now been deleted.

29)
Page 18, Lines 297-298 – While the patients may have shoulder problems, it is unclear whether the injured group had instability symptoms.

Answer: Very relevant. As described, subjects were included based on subjective criteria as well as on clinical pre-screening using shoulder instability and laxity tests. Since the aim primarily was to study reliability and not differences within the case group, especially not with such small subgroups, we decided to report all results for the case group being one group.

Action: None.

30)
Page 19, line 300 – Please clarify why the authors believe that determining reliability before establishing validity is clinically meaningful. What is the impact of the study results to the practicing clinician. Perhaps the investigators are reliably unreliable. Without establishing the diagnostic accuracy including PPV and NPV, the study results and meaning is questionable.

Answer: Thank you for pointing this out.

We believe this comment is very much in line with our answer to question number 6:

(Page 3, line 51 – How will the results of this study improve clinical practice if "further standardization" is needed? Additionally, if the authors haven't established validity of the tests, then what is the importance of establishing reliability? Furthermore, the reliability of these tests has been previously established in the literature, however, the authors do not sufficiently develop the rational and need for the present study.

Answer: Thank you for this very relevant comment. Reliability has previously been established, however, with large variations in results, and with limited methodological quality – as also described P5, Line 104-107. Also, as reliability is a necessary first condition for a measurement to be considered valid and responsive to change, we believe that reliability is the most relevant aspect to address firstly (Mokkink et al. 2010). However, we very much agree with the reviewer that the next step in the process of improving clinimetric properties is to study validity, which has also been addressed in the conclusion section (Page 20, Line 403).

We agree that validity is important to study, including diagnostic accuracy etc. However, to study validity of these clinical tests, a gold standard is required, which is lacking and therefore a challenge for subjects with shoulder instability, which we have also addressed in the study limitations in the discussion section (line 383). We believe that before validity of the tests can be studied, reliability needs to be satisfactory.

We agree with the reviewer that the next step in the process of improving clinimetric properties is to study validity, which has also been addressed in the conclusion (line 408).

Action: None, since both the lack of gold standard for shoulder instability and the recommendation for the further development of clinimetric properties for these tests are addressed in the discussion section.

Conclusion
31)
Page 19, line 318 – How will the findings of the study impact clinical practice?

Answer: Thank you for raising this important aspect. We evaluate this study as the first step in establishing satisfactory clinimetric properties of these clinical tests. Since the aim was to establish reliability the message to the clinicians is that only 4 of these tests have been found to have a satisfactory reliability, while 2 of the tests will need further standardization and testing of reliability. We also recommend to further establish a gold standard method so that the validity and diagnostic accuracy can be established.

Action: None, since these points have been addressed already in the discussion section.

References
32)
81% of the references are greater than 5 years old.

Answer: Thank you for this comment, in which we do agree. However, we have strived to include the most recent literature from relevant databases on this subject. Though, and as previously described, not many studies exist within this clinical area. This is why we believe that this study provides new important insight into an under-investigated clinical area.

Action: None.

References

Dodson, C. C., and F. A. Cordasco. 2008. 'Anterior glenohumeral joint dislocations', Orthop Clin North Am, 39: 507-18, vii.
Juul-Kristensen, B., H. Rogind, D. V. Jensen, and L. Remvig. 2007. 'Inter-examiner reproducibility of tests and criteria for generalized joint hypermobility and benign joint hypermobility syndrome', Rheumatology (Oxford), 46: 1835-41.
Mokkink, L. B., C. B. Terwee, D. L. Patrick, J. Alonso, P. W. Stratford, D. L. Knol, L. M. Bouter, and H. C. de Vet. 2010. 'The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes', J Clin Epidemiol, 63: 737-45.
Patijn J, R. L. 2004. 'Reproducibility and Validity Studies of Diagnostic Procedures in Manual/Musculoskeletal Medicine. Protocol formats', Third edition.
Remvig, L., P. H. Duhn, S. Ullman, T. Kobayasi, B. Hansen, B. Juul-Kristensen, J. S. Jurvelin, and J. Arokoski. 2009. 'Skin extensibility and consistency in patients with Ehlers-Danlos syndrome and benign joint hypermobility syndrome', Scandinavian Journal of Rheumatology, 38: 227-30.
Sim, J., and C. C. Wright. 2005. 'The kappa statistic in reliability studies: use, interpretation, and sample size requirements', Phys Ther, 85: 257-68.
Vind, M., S. B. Bogh, C. M. Larsen, H. K. Knudsen, K. Sogaard, and B. Juul-Kristensen. 2011. 'Inter-examiner reproducibility of clinical tests and criteria used to identify subacromial impingement syndrome', BMJ Open, 1: e000042.

VERSION 2 – REVIEW

REVIEWER	Robert C. Manske Wichita State University Via Christi Health
REVIEW RETURNED	26-Oct-2017

GENERAL COMMENTS	Second review - no additional comments. Authors answered my previous questions and concerns.