Author's response to reviews

Title: Inter-tester reproducibility and inter-method agreement of Beighton tests and criteria for Generalised Joint Hypermobility in primary school children

Authors:

Tina Junge (tjunge@health.sdu.dk)
Eva Jespersen (ejesper@health.sdu.dk)
Niels Wedderkopp (nwedderkopp@health.sdu.dk)
Birgit Juul-Kristensen (bjuul-kristensen@health.sdu.dk)

Version: 3 Date: 30 September 2013

Author's response to reviews: see over
Dear Editor

We are pleased to have the opportunity to respond to the useful comments raised by Claudia Celletti, Dariusz Czaprowski and Marietta van der Linden regarding the article ‘Inter-tester reproducibility and inter-method agreement of the Beighton tests and criteria for Generalised Joint Hypermobility in primary school children’ by Tina Junge, Eva Jespersen, Niels Wedderkopp and Birgit Juul-Kristensen.

The comments and questions will clearly help to improve the presentation and the precision of the key points in the article.

We hereby address the comments giving a point-by-point response to the concerns as well as a revised manuscript.

Reviewer: Claudia Celletti

1) Firstly is not clearly described why they chose these two methods to evaluate BT; furthermore is not clear if these two methods were compared with the standard description of the Beighton test.

   Answer: The following has been added to the Methods section: ‘The two methods of BT were both in accordance with the original text of Beighton et al (Beighton, 1973), which has a rather imprecise description. The tests were performed with slightly different starting positions and benchmarks as this reflects daily clinical practice’. If a difference is evident, then it should have an impact on future research and clinical practice as the most “correct” starting position would need to be determined.

2) Method section is not clear as the figure 1.

   Answer: We will modify the Methods section as below to make it clearer:

   **Inter-tester reproducibility**

   For the inter-tester reproducibility studies, a standardised protocol for clinical reproducibility studies was followed, including a three-phase study with a training phase, an overall agreement phase and a test phase (Patijn, 2007) for each of the two different test batteries, Method A and Method B (Figure 1).

   **Phase 1:** The training phase was performed in an open study in order to discuss and standardise every detail of performing and interpreting the BT among testers, thus improving their ability to follow strict test procedures. In this phase, the testers were not blinded to GJH status or test results. The training phase was carried out in 10 adult cases (fellow physiotherapy students).

   **Phase 2:** Using a blinded study, the main aim of the overall agreement phase was to obtain an overall percentage agreement of at least 80% for finding ≥5 positive tests out of 9 as the criterion for GJH. In this phase, testers were blinded with respect to both GJH status and the other testers results. Two observers were responsible for the
randomisation of the test order, the selection of Method A or B and instructing the children not to comment on their status and the test outcome. A total of 38 children were included in Method A and 32 children in Method B, with an average age of 7.4 years.

Phase 3: In the test phase, the aim was to determine the kappa value (agreement adjusted by chance), using a blinded study, while ensuring an approximate 50% prevalence in order to optimise the kappa statistics validity [12, 13]. Knowledge about the children with GJH score ≥5 found in Phase 2 was used to select children in advance for the test phase (Phase 3), so as to recruit as many children with GJH as possible. As a result, 19 children with, and 20 children without, GJH from Method A and Method B, were sent to the allocated testers (Figure 2). The test phase consisted of 39 children, who were tested with both Methods A and B, and by all four testers. The average age was 9.6 years (Table 2).

Inter-method agreement
For the inter-method agreement study of the prevalence of GJH, the a priori choice of comparing data from Tester 1 with Tester 3, and Tester 2 with Tester 4, was arbitrarily used. The prevalence of GJH for both Methods A and B was compared using the criterion of ≥5/9 as a cut-off level.

The inter-method agreement study involved data from 103 consecutively recruited children, who had been tested in both Method A and Method B during the inter-tester reproducibility study. Six children were not a part of the inter-method analysis, as they unfortunately were only tested with one method. All together, 62 children (60%) represented 7-8 year olds and 41 children (40%) 10-12 year olds.

Participants
Participants were public school children from two different grades: first grade (7-8 years) and fourth grade (10-12 years).
Exclusion criteria were pain in the involved joints on the day of testing and movement restrictions, such as mild cerebral palsy, which would affect the results of the tests.

3) Participants came from two different grades; which is the reason? Why you have not evaluate children from the second and third grade? I think that are not representative of the young population and it should be better to evaluate a population from 7 to 12 years (from first to fourth grade school), also in accordance also to previous studies that showed a different percentage of hypermobility between children aged 4-9 years and at the age of 10-12 years. (van der Giessen, 2001). This in order to confirm in the younger the reproducibility of the BT already showed in adults (Juul-Kristensen, 2007)

Answer: You are right, that to represent a young population, all the grades from 1.- 6. Grade must be tested. The purpose of this study was to exemplify reproducibility of the BT in the youngest and oldest grades – why the text will be changed to:
'The grades are representing the youngest and oldest children in the CHAMPS Denmark part 1- The Childhood Health, Activity and Motor Performance School Study Denmark, a longitudinal cohort study of 1300 children in the Municipality of Svendborg.'

4) In the background specify that according to Fairbank, et al, a positive Beighton test occurs when ROM exceeds mean +3 SD; in a biological context, however, abnormality is generally considered present when the measured parameter exceeds mean +2 SD.

   Answer: You are right, that Fairbank (1984) describes that Beighton is using about 3 standard deviations from the mean in the joints that is selected, although this is not described in the original article by Beighton (1973). However, also according to Fairbank (1984), 2 standard deviations from the mean are usually recommended as an appropriate criterion when a scoring system is used, which is the argument for using this criterion in this study.

5) The standard protocol is not present in the web site showed in reference [9]; if is not correctable please insert more detail of the protocol.

   Answer: Thank you for that information – the website has been changed since the reference was inserted. The method is also described and applied in the reproducibility study of GJH in adults by Juul-Kristensen: ‘Inter-examiner reproducibility of tests and criteria for generalized joint hypermobility and benign joint hypermobility syndrome’.

6) In the discussion is described the high prevalence of GJH in the group; are there differences between male and female?

   Answer: You are right, that the prevalence between males and females could have been described, but in a reproducibility study as this is, gender differences is of minor importance, for which reason we do not find it relevant to present these values.

Reviewer: Dariusz Czaprowski

Methods

1) Phase 3: what does it means `almost 20 children`? Please be precise

   Answer: Yes, this is imprecise and will be corrected to ‘As a result, 19 children with, and 20 children without.’

2) Participants

   I don’t understand the information that for the Phase 2, 38 children were included in Method A and 32 in Method B (overall 70 children). Then we have information that in Phase 3, 39 children were included. What was the inclusion criterion to include these children for Phase 3? Why 31 children were excluded?

   Answer: We are sorry, if the Methods section is misleading, and we will try to simplify this. When carrying out an inter-tester reproducibility study according to a
standardised protocol format (Patijn, 2007), 3 phases must be executed: 1) a training phase (Phase 1), 2) an overall agreement phase (Phase 2) and 3) a test phase (Phase 3). For the overall agreement phase (Phase 2), 38 children were included in Method A and 32 children in Method B. For the test phase (Phase 3) 39 children were included. In each of the three phases different children were included (Patijn, 2007). It is recommended that about 40 arbitrarily chosen children are needed for the test phase (Phase 3), why 30 children were not tested in this phase.

3) The 3rd Paragraph starts from the sentence: `The inter-method agreement study involved 103...children who had been tested in both Method A and B during inter-tester reproducibility study`. In paragraph 2 the authors gave information that the inter-tester reproducibility study was carried out in 10 adult cases. I don’t understand that.

Answer: Again, we are sorry, if the Methods section is misleading, and we will try to simplify this. When carrying out an inter-tester reproducibility study according to a standardised protocol format, 3 phases must be executed: 1) a training phase (Phase 1), 2) an overall agreement phase (Phase 2) and 3) a test phase (Phase 3). For the training phase (Phase 1) 10 physiotherapy students were tested. We hope, that the description of the Study design in the Methods section is now describing this thoroughly: `The training phase (Phase 1) was performed in an open study in order to discuss and standardise every detail of performing and interpreting the BT among testers, thus improving their ability to follow strict test procedures, whether these are on adults or on children`. These results do not influence the reproducibility of the overall agreement and the test phases on the children.

4) Authors write that `not all children participating in the inter-tester reproducibility study took part in the inter-method study`. Why? Please, give the reasons. Additionally the authors have written in the Paragraph 2, that in the inter-tester study there were adults.

Answer: The following will be added to the manuscript: `The inter-method agreement study involved data from 103 consecutively recruited children, who had been tested in both Method A and Method B during the inter-tester reproducibility study. Six children were not a part of the inter-method analysis, as they due to lack of time were only tested with one method.

For the training phase (Phase 1) 10 physiotherapy students were tested, as the purpose of this phase is to discuss and standardise every detail of performing and interpreting the BT among testers, thus improving their ability to follow strict test procedures, whether these are on adults or on children.

5) Exclusion criteria: the authors focus only on the pain in the involved joints in the day of testing, movement restrictions and mild cerebral palsy. What with the exclusion of children with systemic diseases or e.g. with arthralgia for longer than 3 months in 4 or more joints what could suggest hypermobility syndrome?

Answer: The population was a healthy children population all attending a normal
school, meaning that no children had systemic diseases. Therefore, pain was considered a factor that could possibly influence reproducibility, which was the reason for excluding those children. Therefore, if a child had had arthralgia for longer than 3 months in 4 or more joints, they would be excluded from the study, as one of the exclusion criteria was pain in the involved joints on the day of testing.

6) I have the problem with the evaluation of the inter-method agreement by different testers. We don’t know if the potential differences are not caused by the different interpretation of the tests by testers. Although in the Phase 1, the testers discussed the Beighton test (probably the original version), they did not analyze the Methods A and B. Therefore, we can not a priori assume that the interpretation of the particular test will be the same in Methods A and B by different testers.

In my opinion better is to conduct the comparison of two methods by one observer on the same group of children. However, the period of time between tests should be sufficiently long to avoid the situation when the observer remembers the results of the first observation.

Answer: There are 2 studies: The reproducibility study (of both methods) and comparison of the 2 studies. Since the reproducibility was satisfactory in both methods the risk for observer bias was minimal. For the comparison study, the tests were performed on the same group of children and by the same testers as described in the Methods section.

7) The Method section is difficult to understand. I suggest improving readability of this section.

Answer: We are sorry, if the Methods section is misleading, and we will try to simplify this:

**Inter-tester reproducibility**

For the inter-tester reproducibility studies, a standardised protocol for clinical reproducibility studies was followed, including a three-phase study with a training phase, an overall agreement phase and a test phase [9] for each of the two different test batteries, Method A and Method B (Figure 1).

Phase 1: The training phase was performed in an open study in order to discuss and standardise every detail of performing and interpreting the BT among testers, thus improving their ability to follow strict test procedures, whether these are on adults or on children. In this phase, the testers were not blinded to GJH status or test results. The training phase was carried out in 10 adult cases (fellow physiotherapy students).

Phase 2: Using a blinded study, the main aim of the overall agreement phase was to obtain an overall percentage agreement of at least 80% for finding ≥5 positive tests out of 9 as the criterion for GJH. In this phase, testers were blinded with respect to both GJH status and the other testers results. Two observers were responsible for the randomisation of the test order, the selection of Method A or B and instructing the children not to comment on their status and the test outcome. A total of 38
children were included in Method A and 32 children in Method B, with an average age of 7.4 years.

Phase 3: In the test phase, the aim was to determine the kappa value (agreement adjusted by chance), using a blinded study, while ensuring an approximate 50% prevalence in order to optimise the kappa statistics validity [12, 13]. Knowledge about the children with GJH score ≥5 found in Phase 2 was used to select children in advance for the test phase (Phase 3), so as to recruit as many children with GJH as possible. As a result, 19 children with, and 20 children without, GJH from Method A and Method B, were sent to the allocated testers (Figure 2). The test phase consisted of 39 children, who were tested with both Methods A and B, and by all four testers. The average age was 9.6 years (Table 1).

Inter-method agreement
For the inter-method agreement study of the prevalence of GJH, the a priori choice of comparing data from Tester 1 with Tester 3, and Tester 2 with Tester 4, was arbitrarily used. The prevalence of GJH for both Methods A and B was compared with the criterion of ≥5/9 as a cut-off level.

The inter-method agreement study involved data from 103 consecutively recruited children, who had been tested in both Method A and Method B during the inter-tester reproducibility study. Six children were not a part of the inter-method analysis, as they unfortunately were only tested with one method. All together, 62 children (60%) represented 7-8 year olds and 41 children (40%) 10-12 year olds (Figure 1).

Participants
Participants were public school children from two different grades: first grade (7-8 years) and fourth grade (10-12 years).
Exclusion criteria were pain in the involved joints on the day of testing and movement restrictions, such as mild cerebral palsy, which would affect the results of the tests.

The grades are representing the youngest and oldest children in the CHAMPS Denmark part 1- The Childhood Health, Activity and Motor Performance School Study Denmark, a longitudinal cohort study of 1300 children in the Municipality of Svendborg [14]. The Committee on Biomedical Research Ethics for Southern Denmark approved the experimental protocol (jnr. S-20080047 HJD/csf). Written information for participation in the study was provided to the parents or guardians of the participating children according to the Declaration of Helsinki [15].
Methods

The two methods of BT were both in accordance with the original text of Beighton et al (Beighton, 1973), which has a rather imprecise description. The tests were performed with slightly different starting positions and benchmarks as this reflects daily clinical practice [4](Appendix 1). Besides variation in starting positions and benchmarks, the test batteries also differed in whether the tests were performed active or passive, how they were influenced by gravity and whether the surrounding soft tissue was in a stretched or relaxed position (Appendix 2). The current authors (TJ and EJ) made detailed descriptions regarding starting positions and benchmarks for the two different BT batteries (Appendix 2).

The BT started with a visual demonstration by the tester of the single test along with an oral instruction on how to perform the test before the children performed the test themselves. In the two methods, the children were asked to bring the joint to the most extreme position according to Methods A and B, tested consecutive by four different testers with approximately half an hour between testing sessions. All tests were performed in a random order with respect to right and left sides and to the test sequence.

A positive single test in the BT counted as 1 point, giving a maximum of 9 points, as previously described by Beighton [4]. A cut-off level for classification of GJH in children is internationally not established, as the predictive validity of GJH, for this time being, is not known. Due to the lack of predictive validity, an a priori cut-off level of ≥5/9 for GJH was chosen in the current study. Earlier studies have suggested different cut-off levels for classification of hypermobility in a child population: ≥4/9, ≥5/9 and ≥6/9 [8, 16, 17].

The same four testers evaluated the two different test batteries; two testers (Tester 1 and Tester 2) for Method A and two testers (Tester 3 and Tester 4) for Method B (Figure 2). The testers were physiotherapy students on the last year bachelor program, well trained in the performance and the interpretation of the BT.

8) Please add the gender distribution separately for children aged 7-8 years and 10-12 years.

Lack of information how many girls and boys were included in Method A and B, separately. Due to the fact that the prevalence of GJH is connected with a gender, the analysis regarding gender-agreement between groups A and B is needed.

Answer: It is right, that separately gender distribution as well as the prevalence between males and females could have been described. However, in a reproducibility study as this is, separately gender distribution and gender differences are of minor importance, since the 50%-prevalence method was used as recommended (Patijn, 2007), and thus an a priori specific gender distribution was not intended.
Tests

9) You have written that for knee and elbow the hyperextension should exceed 10°. Please explain, how was measured range of motion in knees and elbows? I'm afraid that it was not possible to measure range of knee and elbow motion in the evaluation showed in Appendix 2.

Answer: We agree that previous studies found the knee and elbows to have the lowest reproducibility of all the Beighton tests. This was, however, not found in the present study. Joint laxity was measured during clinical examination by (subjective) judgment by eye as described by Beighton (1973).

10) I have also problem with validity of the assessment of 5th finger. I see in the Discussion that authors showed the paper of Hansen et al. to justify their choice. In my opinion one reference for that, is not enough. There exist a lot of papers where we see suggestions that the objective measurement of range of motions is needed. Therefore, I suggest to include in the paper the Limitations section and to describe there the questions regarding assessment of knees, elbow and 5th finger.

Answer: You are right, that more objectively measurements of range of motion would be more valid, and should be used in a validity study. As the purpose of this study was to test the reproducibility of this particularly clinical test battery, joint laxity was measured during clinical examination by (subjective) judgment by eye in all test conditions as described by Beighton (1973). Based on this we anticipate the risk of bias in the current reproducibility study in both groups to be equal and minimal.

11) Does the manuscript adhere to the relevant standards for reporting and data deposition?
In my opinion the protocol of the study is correct. However, I have some doubts regarding the methods. The first one is the evaluation of joints range of motion. I don’t know how authors assessed hyperextension (10°) in knees and elbows, dorsiflexion of the fifth finger beyond 90°, shoulder position 90° in flexion or abduction and 90° of elbow flexion in assessment of fifth finger in Method B.
The picture that illustrated evaluation of the fifth finger in the Method B: the authors described this test: `...elbow in 90° flexion...`. I don’t think that we see this angle on the picture.

Answer: It is true, that the picture does not illustrate an angle of 90° elbow flexion, and this will be replaced by a new illustration.

12) The second question is why the testers were physiotherapy students? Why the tests were not physiotherapists? There were students of the bachelor or master program? On which year there were?

Answer: Testers were physiotherapy students on the last year bachelor program, which will be added to the text. By using this 3-phase study design for reproducibility studies, with a thorough training and overall agreement phase, it is shown, that it is possible even for physiotherapy students to obtain a satisfactory reproducibility of these tests. Other studies have shown that experienced clinicians are not necessarily the best in following a standardised research protocol.
13) Is the discussion and conclusion well balanced and adequately supported data?

Conclusions are supported with the results. However, I have previously mentioned doubts about the methods of the assessment of joint range of motion. Therefore, I hope that the authors will be able to improve this section and accordingly it comment in the Discussion section.

Answer: We agree with that, and hope that paragraph 4, 5 and 6 in the Discussion section covers this problem: ‘The current most difficult body parts to evaluate were the knees, the elbows and the fifth fingers when visually estimating range of motion (ROM) in degrees (≥10° for knees and elbows and ≥90° for the fifth fingers). This was in accordance with the study by Hansen [7], with kappa values of 0.68 for the elbows and only 0.44 for the knees, judged by trained rheumatologists. However, that study did not include an overall percentage agreement phase, which may be the main reason for the poor reproducibility. In a previous study, reproducibility of tests for the elbows and the fifth fingers for adults was correspondingly low (κ <0.61), but for the knees kappa was as high as >0.85, possibly due to a prevalence close to 0.50 for the knees [6].

Comparing visual judgements with goniometer measurements represents a general challenge, but visual judgement is part of daily clinical practice. This problem was illustrated in a child study [21], where goniometry was used to measure the passive bilateral hyperextension of the knees along with visual judgements. The children were placed into three sub-groups covering: the not hypermobile (Beighton Score 0-4); the children with increased mobility (Beighton Score 5-6); and the children being hypermobile (Beighton Score 7-9). These three sub-groups were used for analysis of concurrent validity presenting significant differences between the exact degrees by goniometry and the total scores classified as the three sub-groups. The visual judgment of ROM in degrees for the single test was not validated against goniometry, potentially biasing the results, as the presence of hypermobile knee joints in the third sub-group could be low.

Concurrent validity between goniometer measurements in degrees and visual judgment of the score of the single test was also evaluated in a pilot study, with, in contrast, no significant difference in the prevalence of GJH (criterion ≥6/9) in a child population, evaluated by goniometer measurements in degrees and visual judgment [22]. However, when comparing the individual tests, the prevalence for the five single tests was dissimilar for the elbows and especially the knee, judged by goniometer and visual estimates (right knee 2% resp. 18%, left knee 6% resp. 18%). This difference was obvious by in experienced and non-experienced physiotherapists [22].

Without using test standardisation, reproducibility of knee extension by goniometer and visually, with or without test standardisation, was confirmed in a systematic review, where the reproducibility of measurements by goniometer varied from Kappa
(PABAK) -0.02 (pre-standardisation of test) to 0.88 (post-standardisation of test) by rheumatologists [23, 24]. In general, both goniometric measures and visually estimated measures were above ICC 0.59 for adults with or without diagnoses in the aforementioned systematic review including seven studies for knee extension measures [24].

14) Paragraph 12: the authors found the prevalence of GJH in 31% and 35% for Method A and B, respectively. The authors compare these results with other authors with suggestion that they used the same cut-off level and analyzed children at the same age. I disagree with this comparison as the authors of the present study used other method (not Method A and B from this paper) to assess GJH. Furthermore, Mikkelsson et al. [8] used cut-off #6 points, and Juul-Kristensen et al. [16] analyzed only children at 8 years of age.

Answer: It is right that the BT was not performed in exactly the same way; but all the studies claimed that their methods were following the original description by Beighton (1973), which shows that interpretation was ambiguous and rather imprecise.

15) Are limitations of the work clearly stated?

No. In my opinion, the separately section Limitations is needed where the authors will comment lack of objective evaluation of joints range of motion. I’m afraid that it is very difficult for young children to stay in e.g. abduction 90° in shoulder when they have to stretch elbows. Of course, I see that in the Discussion section the authors analyzed the subjective and objective evaluation. However I’m afraid that is not sufficient to justify the choice of the methods and it should be described as a limitation. I hope that the authors will be able to explain this methodological problem

Answer: The purpose of the study was not to compare different objective measurements, or to compare objective with subjective measurements. The purpose of the study was to study reproducibility and compare performance of the BT tests, as described by Beighton, 73, according to the interpretations in clinical practice. Since the clinical interpretations are varying, due to the imprecise description of the Beighton tests (Beighton, 1973), we have selected different ways of performing these tests for reproducibility, and also compared the different test performances with each other, to see the influence on the prevalence. We believe that paragraph 4, 5 and 6 in the Discussion section covers now this problem, as the purpose of this study was to examine an often used, subjective evaluation for range of motion.

16) Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?

Yes, but I suggest to include in the Discussion section an information what is the difference between generalized joint hypermobility and hypermobility syndrome. The authors used in the Reference list the position, which are focus on the hypermobility syndrome [3 and 20]. Therefore, I suggest including a comment to distinguish GJH and HS

Answer: Thank you for this suggestion, but since this is a reproducibility study of the Beighton tests, and these tests are the same in both conditions, this will not influence the data.
Do the title and abstract accurately convey what has been found?

Abstract

17) I would like to see one sentence to explain Method A and B to facilitate the reader to understand the difference between studied groups.

Answer: The text has now been revised and changed to: “...the inter-tester reproducibility of the tests and criteria for classification of GJH of two different Beighton test batteries (Methods A and B) with a variation of starting positions and benchmarks between methods, and 2) the inter-method agreement for the two batteries.”

18) I suggest changing the title to show that the authors assessed two different methods/variants of original Beighton scale

Answer: Thank you for that suggestion, we will change the title to

“Inter-tester reproducibility and inter-method agreement of Beighton tests and criteria for Generalised Joint Hypermobility in primary school children

– An evaluation of 2 different test batteries for performing the Beighton tests”.


Answer: It is right that different cut-points have been suggested for girls and boys as well as for different age and ethnic groups, but there is no consensus on a specific cut-point. Final cut points for classification of GJH is although not possible to establish, as the predictive validity of GJH is not yet known.

Due to this lack of predictive validity, an a priori cut-off level of ≥5/9 for GJH was chosen in the current study. Earlier studies have suggested different cut-off levels for classification of hypermobility in a child population: ≥4/9, ≥5/9 and ≥6/9 (Mikkelsson, 1996, Juul-Kristensen, 2009, van der Giessen, 2001)
Reviewer: Marietta van der Linden

1) The main issue I have with this manuscript is its clarity. For example, as it describes a rather complicated study design, a diagram with the different phases and comparisons would be useful, especially which children took part in which study. Although this has been explained, a diagram would be helpful.

   Answer: Thank you for that suggestion – a diagram with the different phases has now been added (Figure 1).

2) Another issue is the choice for the cut-off of # 5. In the methods it is stated that there is a lack of international consensus, but no rationale is given for 5. In the Discussion however, the authors write that ‘in order to follow the cohort over time,... a higher cut-off level is needed....as recommended by other authors. So why was no higher cut-off chosen for this study? Can the analysis be repeated using a higher cut-off?

   Answer: A cut point for classification of GJH in children is internationally not established, as the predictive validity of GJH, for this time being, is not known. Due to the lack of predictive validity, an a priori cut-off level of ≥5/9 for GJH was chosen for this inter-tester reproducibility study of 2 methods for assessing joint hypermobility. The 50%-prevalence method was used on the basis of this cut-point, meaning that it will not be relevant to reanalyse the data with another cut-point. When following a cohort of children with GJH over time, the cut point might be set higher, as the prevalence when using cut point ≥5/9 is high.

Background
3) Please describe briefly methods A and B (consider removing the second sentence, as this on its own does not explain the need for the study).

   Answer: Methods A and B has now been briefly described: ..’the inter-tester reproducibility of the tests and criteria for classification of GJH of 2 different ways of performing the Beighton test batteries (Methods A and B) with a variation in starting positions and benchmarks between methods’.

Methods
4) ‘three phase protocol’: the three phases are not described. Also, phase 1, did not include children but 10 physiotherapy students?

   Answer: The three phases have now been introduced.

   It is right, that Phase 1 was not including children, why this has been changed to ‘participants’.

Results
5) ‘overall agreement’, agreement of what? Are these numbers given kappa values?

   Answer: The result for overall agreement is the percentage agreement, not corrected for agreement by chance, which has been taken into consideration when calculating
kappa. It is not relevant to calculate kappa for the overall agreement, since this was not intended in the design, due to no controlled 50%-prevalence.

6) .. in the study phase...’ this is not clear
   Answer: The term study phase will be removed from the abstract to make it clearer.

Background
7) First sentence, ‘+2SD (1)’, should that not be ‘± 2SD of a normal age matched population?
   Answer: When focusing on hypermobility and not hypomobility, it is only +2SD.

8) Please add whether the same (four) raters involved in all three phases?
   Answer: The same four raters were involved in all three phases, and this will be added to the manuscript.

9) Page 4, ‘almost 20 with ‘, please state the exact number: 19?
   Answer: Yes, this is imprecise and will be corrected to ‘As a result, 19 children with, and 20 children without.’

‘Participants’
10) Any particular reason why none of the second grade pupils took part/were not invited?
    Answer: The purpose of this study was to exemplify reproducibility of the BT in 2 grades, representing the youngest and oldest children in a cohort study we are performing.

11) The inter-method agreement study involved ..’
For clarity, consider -adding 103 out of the 110 children who had taken part in the inter-tester reproducibility study’ (What happened to (the results of the) remaining 7 children)
Did these 103 children come back for further testing or were their results used for the inter-method agreement?
   Answer: The following will be added to the manuscript: ‘The inter-method agreement study involved data from 103 consecutively recruited children, who had been tested in both Method A and Method B during the inter-tester reproducibility study. Six children were not a part of the inter-method analysis, as they due to lack of time were only tested with one method.

12) For clarity please add diagram with study flow/participants
    Answer: We have now included 2 diagrams with study flow/participants that we hope you find illustrative: Figure 1: The inter-tester reproducibility study included a three-phase study with a training phase, an overall agreement phase and a test phase.
Figure 2: Flow-chart for the 0.50 prevalence index method, study phase (Phase 3) for Methods A and B.

13) Page 6, line 4: The tests.. Consider adding ‘methods’
   Answer: We agree with that and have changed ‘the tests’ to ‘the methods’.

14) Please add a rationale for choosing 5 as a cut-off value
   Answer: The following has been added to the text: ‘A cut point for classification of GJH in children is internationally not established, as the predictive validity of GJH, for this time being, is not known. Due to the lack of predictive validity, an a priori cut-off level of ≥5/9 for GJH was chosen’.

Results
15) For clarity please add the table number at the start of the paragraph.
   Answer: We have now added the table numbers at the start: ‘In Phase 3, kappa values varied from 0.49-0.94 (Method A) and from 0.30-0.84 (Method B) for the nine single tests in the batteries (Table 2). In 8 out of 9 tests, Method A had the highest agreement and the largest kappa value with a mean percentage agreement of 87%, while Method B had a mean percentage agreement of 81%. The mean kappa value for all tests was 0.70 (Method A) and 0.59 (Method B).

   The body part with the highest agreement and kappa value was the first finger on the right hand for both Methods A and B (97%, κ 0.94 resp. 92%, κ 0.84) and the first finger on the left hand (95%, κ 0.89 resp. 92%, κ 0.82) (Table 2). The most difficult body parts to judge were the knees (mean 85%, κ 0.62 Method A, mean 68%, κ 0.37 Method B) and the elbows (mean 85%, κ 0.68 Method A, mean 79%, κ 0.57 Method B) (Table 2)’.

16) Please also check your table numbering. The text refers to mean kappa values in table 3 which I can’t find. Table 4 is referred to in the text but doesn’t exist.
Table 3 includes the results for BT#4 which is not referred to in the text.
   Answer: We thank you for drawing our attention to this and will correct it in the text.

Discussion
17) First sentence: The inter-tester reproducibility of the test items of Methods A and B.....
   Consider adding ‘of the test items’
   Answer: We agree and will add this to the text.

18) Second paragraph; ‘Only two studies.......Method A This paragraph is not clear; which study found which results?
Answer: That is right and has been changed to: `. .. the ones in the current study (0.69 (only four tests) (Hansen, 2002), 0.78 (Mikkelsson, 1996) and 0.70 Method A (current study))’.

19) Third paragraph
The forward ...but only moderate kappa values (0.64 and 0.84). due to low prevalence.
A kappa of 0.84 is actually classified as ‘substantial’?
Answer: Thank you for drawing our attention to this. We will correct this to: ‘The forward bending test had high overall agreement (95% resp. 97%) in the current study, but diverging kappa values from moderate to almost perfect kappa values (κ 0.64 resp. 0.84), affected by low prevalence’.

20) Page 9, Third paragraph ‘With standardised and detailed test protocols.... is likely to be attainable.[9]But could you say that this was not confirmed for all test items by the current study?
Answer: That is true – in general the aim is to increase the agreement of the outcome scores with standardised and detailed test protocols, which succeeded with moderate to substantial kappa values in the current study. Also, despite standardised test protocols, kappa values for reproducibility studies of tests for GJH are often not high, as the magnitude of kappa is affected by the prevalence of the condition in the population.

21) Page 9/10 ‘deceptively’ high.
Please add a possible explanation why the prevalence so high in this study compared to others using the same cut-off value?
Answer: The following has been changed to explain the prevalence found: ‘The prevalence found in this reproducibility study was deceptively high (31% Method A, 35% Method B) using a cut-off level of ≥5/9, however this cut-off level was chosen with the 50%-prevalence method for purpose’.

22) Last paragraph Line 4 onwards ‘Small differences ..the total BT score. This cannot be concluded from the results of this study.
Answer: We believe, that when finding the same prevalence of GJH by applying 2 different test batteries of performing the BT, it is fair to conclude that small differences in the way the BT is performed may not have an impact on the prevalence when using a relatively low cut-off level and when using this specific combination of test performance. But naturally there are differences between the individual tests with different starting positions and different benchmarks.
References