Author's response to reviews

Title: Reproducibility of and sex differences in common orthopaedic ankle and foot tests in runners

Authors:

Maarten P. van der Worp (MvdWorp@gmail.com)
Anton de Wijer (Anton.deWijer@Radboudumc.nl)
J.Bart Staal (Bart.Staal@RadboudUmc.nl)
Maria W.G. Nijhuis-van der Sanden (Ria.Nijhuis-vanderSanden@RadboudUmc.nl)

Version: 3
Date: 5 February 2014

Author's response to reviews: see over
Dear Dr. Bañas,

Thank you for consideration of our manuscript for publication. We would like to thank the reviewers for an excellent evaluation and the suggestions they made for improving our manuscript. We reviewed the manuscript according to the reviewer’s comments (please see below).

We are submitting the revised version of, *Reliability of common orthopaedic tests used to evaluate foot position and foot function in runners* for publication as an original research article in BMC Musculoskeletal Disorders. We changed the title in a more appropriate name: *Reproducibility of and sex differences in common orthopaedic ankle and foot tests in runners*.

**Competing interest:** No sources of funding were used to assist the preparation of this study. The authors have no conflict of interest that is directly relevant to the content of this study.

**Authors’ contributions:** All the authors have made substantial contributions to the design of the study, analyses and interpretation of the data and drafting of the manuscript. MvdW did the data collection and statistical analyses. The material within has not been and will not be submitted for publication elsewhere except as an abstract for symposium and/or congress. All authors have critically revised and approved the final manuscript.

Yours sincerely on behalf of the co-authors,

Maarten van der Worp
**Reviewer:** Takashi Nagai

**MAJOR COMPULSORY REVISIONS**

1. First, the authors have tried to address several research questions in this manuscript. I would suggest the authors to simplify and focus on a few research questions. For example, if the title is ‘Reliability of…’, why did the authors analyze differences based on gender (males/females), NDT comparison (stance/single limb-stance), three tests, etc. Based on the writing, the descriptive results and ICC/SEM results included both males, females, and all subjects. The authors should have included the rationale for gender comparisons. It might not be the best manuscript; but, I published one manuscript (Nagai T, Sell TC, Abt JP, and Lephart SM, Phys Ther Sport, 2012, 233p) that combined reliability and gender comparisons. If you have two research questions, it is ok to combine and address both questions systematically in one manuscript.

We agree with the reviewer that we combined more than one research question in this manuscript without an explicit rationale and without a clear description. To clarify this we now included two study aims/research questions in our manuscript and gave an explicit rationale for these two aims and used the article by Nagai et al. (2012) as guidance:

*the aim of this study was to develop and assess the intrarater and interrater reliability and agreement of the NDT, AJD test and extension MTP1 test in weight-bearing position in healthy runners. Secondly, we wanted to compare outcomes of these tests between female and male runners.* (Page 7, line 106-109)

2. Second, the interrater and intrarater reliability should be explained in methods. It was confusing to see how the authors actually conducted the reliability. Among the four days (listed in the methods), did they visited three separate days (day 1 with the examiner A, day 2 with the examiner A, and day 1 with the examiner B)? The examiner A performed all tests for all subjects for their day 1 and day 2 for intrarater (and between-day) reliability? If so, the data for day 1 by the examiner A and B can be used to interrater reliability. Please clarify. Some researchers to prefer to measure the same subjects by two examiners in the same day (~30-60mins apart) or separate day.

In the method “procedure” we changed and edited some sentences in order to clarify that all measurements of one runner were taken on one single day and that we needed 4 consecutive days of measurement to test 42 runners. Also, we included a flowchart (figure 1) of the measurement procedure.

*After giving written informed consent, each runner completed a baseline questionnaire about his/her running status and injury history. The height and weight were determined. The runners were randomly assigned to the two examiners (MW and HD). To determine intrarater reliability, runners were measured twice by examiner HD (HD$_1$ and HD$_2$). FIGURE 1 shows a flow chart of the procedure. Both examiners completed all three static tests once for both legs and feet, randomized in test order by computer, with minimally 10 minutes between measurements of examiner HD. All measurements of one runner were taken on the same day.* (Page 8, line 138-145)
MINOR ESSENTIAL REVISIONS
- Here are detailed comments. In my opinion, the manuscript should have major revisions and the change of theme (focusing on the gender difference and reliability). Therefore, some sentences may or may not be included. But, I will list them for your information for your scholarly work/writing.

Thank you for the detailed comments. We made revisions according to your suggestions as described below:

ABSTRACT
1. Abstract: Methods: line 3 (running minimally...): please remove redundant words and make a clean sentence.
   We removed the redundant words and made a clean sentence.

   ‘The three orthopaedic tests were administered by two sports physiotherapists in a group of 42 (22 male and 20 female) recreational runners.’ (Page 2, line 10-12)

2. Abstract: Results: please separate intrarater and interrater reliability.

   We split the results of the NDT Stance, NDT Single Limb-Stance, AJD-test and MTP1 in intrarater and interrater reliability.

   ‘The ICC’s of the NDT were in the range of 0.37 to 0.45 (SEM 2.5 to 5 mm) with a 95% LOA range of 11.4 to 13.9 mm. The AJD-test had an ICC of 0.88 and 0.86 (SEM 2.4° and 8.7°), with a 95% LOA of -6.0° to 6.3° and -5.3° to 7.9°, and the MTP1-test had an ICC of 0.42 and 0.62 (SEM 34.4° and 9.9°), with a 95% LOA of -30.9° to 20.7° and -20° to 17.8° for the interrater and intrarater reproducibility, respectively.’ (Page 2, line 16-21)

3. Abstract: Results: line 3 (SEM): please consider one decimal instead of two decimals. Also, please check the SEM for the MTP1 test. The upper range seems to be too large.

   We use one decimal instead of two and did a recalculation on the SEM of the MTP1-test (but this had no consequence for the outcome)

INTRODUCTION
4. All references: please check your references. Please try to use the original studies to support your sentences. For example, in the second paragraph (Decreased ankle dorsiflexion, increased...(ref7-10)), these references are not studies that evaluated these three tests for people with and without running injuries, prospective studies, etc.

   We apologize and we did a check on our references and made appropriate references with the original studies.
5. Introduction: First paragraph: last two sentences (the prevalence...): please be specific on running injuries (acute, chronic, joints, etc). Also, please separate the prevalence and incidence. I think you could expand these sentences.

We edited this part of the introduction also based on the comments of the other reviewer, who remarked that this information is not essential for the research question. Therefore we excluded the prevalence and included the most common site of lower injuries, because we think it’s relevant for the tests we used in this study.

‘However, running can cause injuries, especially of the lower extremities, with incidences varying between 20% and 92% and with the knee as most common site of injury followed by lower leg and foot [3].’ (Page 4, line 35-37)

6. Last paragraph: first sentence (Although a physiological amount of pronation...): please specify what the normative ranges are.

Because we edited the introduction and focused now more on our two study aims, this sentence was changed and used later in the introduction in relation with the findings in the literature as a predictor for running related injuries:

‘Bennett et al. [5] and Buist et al. [6] found in their prospective studies a positive navicular drop test (NDT > 10mm) as predictor for running related injuries.’ (Page 4, line 45-46)

7. (However, there are no data in the literature...): Please remove the references. It is your problem statement. Or please reword these sentences.

We removed the references.

8. (Moreover, videos on internet, ...): please remove the sentence. In general, you won’t need to include any videos.

We removed the sentence about videos on internet, although we think these videos showed the large variation between tests taking behavior. However, we think it is important for the reader to see how the measurements in this study were taken and were slightly adapted from those reported in the literature, so we decided to present the link to the videos in the manuscript, so readers can find this if wanted.

METHODS

9. Participants: first sentence: Please insert the table for their demographics.

We inserted one table with the age, length and weight and also included some running characteristics so there is a clear description of our study sample which facilitates interpretation and comparison with other populations.
10. Participants: second sentence: (Potential participants were…): please rewrite.

You will not need to state the actual calendar dates. Instead, please focus on what subjects had to do. (if I understand right, they made three visits, correct?)

The runners had one visit during which all the measurements were taken. We edited this paragraph and included a flowchart to clarify this.

‘After giving written informed consent, each runner completed a baseline questionnaire about his/her running status and injury history. The height and weight were determined. The runners were randomly assigned to the two examiners (MW and HD). To determine intrarater reliability, runners were measured twice by examiner HD (HD$_1$ and HD$_2$). FIGURE 1 shows a flow chart of the procedure. Both examiners completed all three static tests once for both legs and feet, randomized in test order by computer, with minimally 10 minutes between measurements of examiner HD. All measurements of one runner were taken on the same day.’ (Page 8, line 138-145)

11. Examiners: Please remove the section and insert one sentence in the procedures.

We removed this section and inserted two sentences about the examiners, because we wanted to explain the experience of the examiners.

12. Procedures: second sentence: (After giving written informed…): Please remove the Table on running status. It does not address your research questions. As stated earlier, you can simply state about the minimal running frequency, distance, and health (injury free) for inclusion/exclusion criteria.

Because reliability is a characteristic of performance of an instrument in a certain population it is important to describe some runner’s characteristics to emphasize that the study population is comparable to a regular runner’s population. Although, we agree that this table was too detailed, we removed a part of this table and moved another part (years of running, weekly training frequency and weekly training distance) to the demographic characteristics presented in table 1.

13. Procedures: third sentence: (The height, weight, body mass index…):

Please remove the body fat measurement. It is not a part of demographics. Also, if you would like to include, you should explain in details why you decided to use impedance and its reliability as a part of study. The demographics table should include their height.

We removed the body fat measurement and included the height. See table 1.
14. Procedures: 4-7th? sentences: (The runners were randomly assigned...): Please re-write these sentences to make it clear. Were all subjects tested by the first examiner twice for intrarater reliability? If so, how many days (or if it is within-day, how many hours apart?) were separated between the first and second assessments?

The runners made one visit where all the measurements were taken. The runners were measured by examiner MW and twice by examiner HD in a random sequence. We edited this paragraph and include a flowchart (see figure 1) to clarify this.

15. NDT: first sentence: please remove the video and citations.

We removed the citation, but kept the video as stated before.

16. NDT: the first sentence: Please remove the figure and replace with the figure with a measuring rod from the side view.

We replaced this figure with a picture of measuring the most medial aspect of the navicular bone with a measuring rod (see figure 2).

17. NDT: as the authors later stated in the discussion, the authors modified measuring tools (measuring rod instead of a note card). Please include statements why the authors chose to use the measuring rod.

We used a measuring rod/ruler so the navicular drop (distance) could be measured directly. We included a statement in the introduction and method:

‘In our protocol a ruler was used instead of a blank card [12] and the sitting position was used as neutral position of the foot instead of palpating the talar head [13].’ (Page 8, line 148-150)

18. AJD: first sentence: please remove the video.
19. MTP1: first sentence: please remove the video

We kept the video as stated before.

20. Statistical analysis: First sentence: please expand on the rationale and what was performed. Please include reference to support.

We edited the first sentence, removed “were used to satisfy the assumption of independence of data” and moved the sentence to the last part of the paragraph. Furthermore, we edited the first paragraph by focusing on the baseline characteristics.

‘The Shapiro-Wilks test was used to assess whether the baseline characteristics were normally distributed. The mean and standard deviations were given for the normally distributed baseline characteristics, as the median and range were given for the non-normally distributed data. The data from the right and left leg and feet were used separately in all calculations.’ (Page 11, line 203-207)
21. Statistical analysis: the first paragraph: for reliability analyses, this section is not necessary. But, I would strongly suggest the authors to try to incorporate gender comparisons. Please see the previous comments.

We edited the first paragraph, and incorporated sex comparisons in the statistical analysis paragraph:

‘For sex differences comparisons, means, standard deviations, mean differences and 95% confidence interval (CI) of the dependent variables of the three tests were calculated, for which the data of the measurement of examiner MW were used. Independent t-test with an alpha value of 0.05 was used to evaluate sex differences comparisons.’ (Page 11 and 12, line 224 - 228)

RESULTS

22. Participants: Please remove these paragraphs and tables. Table 2 can be inserted in the methods: participants to describe demographics (and please remove the body fat). Table 4 on running frequency, duration, training surface, etc. are not a part of the study.

Tables 2 and 4 were removed and the data on running experience, weekly training frequency and weekly running distance of table 2 were used in the methods: table 1.

23. Outcomes: Please remove this section and table 5. It is not a part of the study. If the authors include gender comparisons, it will make more sense.

We agree and we changed this.

24. Outcomes: second sentence: (The mean overall difference in NDT stance score…) Please remove. It is not a part of the study. If the authors would include, it would be the paired t-tests and Wilcoxon signed-rank test for stats. But, if you would like to include, it will be another research question. Frankly, why would you like to know the differences in two NDT tests? Would it be critical? Please add your rationale in the intro.

We agree, we don’t want to know the difference between the NDT tests as you stated, so we removed this part.

25. NDT single limb-stance: the second paragraph: (The ICC of the intrarater…) Please state the ICC/SEM and combine with the previous paragraph. Please remove all directional language (higher, lower than…). Instead, simply, please state values and the magnitude of correlation based on your definition of ICC: 0.0-0.25 little, 0.26-0.49 low, 0.50-0.69 moderate, 0.70-0.89 high, and 0.90-1.00 very high correlation.

We stated the ICC/SEM and combined the results with the previous paragraph en removed all directional language.
We don’t know if we understand you correctly but we disagree about the Bland & Altman plot, because we think these plots give a good graphical overview of the random error (95% LOAs) of the interrater and intrarater reliability, additional to the SEM. Moreover, this method is often used in this kind of studies. We included 95% LOAs and checked the font size, which was adapted into the “normal” size.

**DISCUSSION**

We included the statements/results of previous studies of SEM and 95% LOA and discuss how they differ from our results.

‘The SEM of 2.5 mm for the intrarater agreement in our study is in the range of 0.4-2.7 mm as reported in the literature [9,12,26,27]. The 95% LOA was however higher, 6.1 mm, as compared to the study by Evans et al [12], who found a 95% LOA of 5.2 mm. Our SEM for the interrater agreement (3.2 mm) was higher as reported in the literature with a SEM in the range of 1.4-2.7 mm [9,27,28]. Also the 95% LOA’s for the interrater agreement were wider than those in the study of Shultz et al. [9] who found values between 1.4 and 2.6 of the 95% LOAs by four testers. In addition, both ICC’s of 0.41 and 0.37 and SEMs of 5 and 2.5 mm for the interrater and intrarater reproducibility of our findings for NDT Single Limb-Stance differ from those of Vinicombe et al. [8], who reported a higher reliability (range of 0.33 to 0.76) and lower SEMs of 1.06 to 1.87 mm.’

(Page 13 and 14, line 274-284)

We agree, we removed the statements on 2SD or the smallest detectable change (SDC). Instead, please focus on the SEM and 95% LOAs. If the authors were interested in 2SD and SDC, please include rationale in the intro, calculations in the methods, and values in results.

We discuss the results of the SEM and LOA (Munteanu et al.) with our results and focus on the SEM and 95% LOAs as you stated.
32. MTP1: The authors should have informed readers that the MTP1 protocol by Hopson et al. is different from the current protocol. It should be addressed in the intro and methods. Frankly, why did you change the protocol? Please rewrite the section and include rationale in the intro and methods.

We addressed, in intro and methods, how and why we change the protocol. The sections were rewritten.

Introduction (Page 6, line 81-94)

‘In this study, the protocols in the literature of the NDT [8], AJD-test [11] and extension MTP1-test [10] were adapted for the use in our planned prospective cohort study of female runners (n= 433). This adaptation was necessary for practical reasons, which required that these orthopaedic tests are performed in maximal 10 minutes, on location and with a minimum of measurement tools and equipment.

For the NDT [8], in our protocol a ruler was used instead of a blank card [12] and the sitting position was used as neutral position of the foot instead of palpating the talar head [13], so the NDT could be determined directly and the measurement time was minimized. The performing times of the protocols of the AJD-test [11] and the extension of MTP1-test [10] were optimized by refraining from using a tape line and standardized step length, but extra attention was paid to maximal stretch of the posterior leg and MTP joint, respectively. Consequently, by deviating of existing protocols, the agreement (as a characteristic of the protocol and measurement instrument itself) of these three tests had to be determined as well.’

Methods (Page 10, line 182-184)

‘For measuring the mobility of the MTP1 the protocol, as used in the study of Hopson et al. [10], was slightly modified (not standardizing the step length) and used in this study, see video 3.’

33. Discussion: MTP1: please remove the video.

We removed the video.

34. Conclusion: please include limitations of the study.

We included the limitations on standardization and foot length in our study by including the following text phrases in the last part of the discussion:

‘This study has some limitations. First, the standardization of the NDT was not optimal. Concerning the NDT, in our protocol the sitting position was used as neutral position of the foot instead of palpating the talar head, so the ND could be determined directly and the measurement time was minimized. The measurement time of the protocol of the extension of MTP1-test [10] was optimized by refraining from using a tape line and standardized step length. However, to guarantee standardization, it was ensured that all participants reached the maximal stretch of the MTP joint with an extended knee, so step length did not influence the MTP1 extension. It was deliberately chosen to deviate slightly from the existing protocols in the literature to optimize the performing speed of the tests and to facilitate test
performance in practice. Given the fact that we were planning a large epidemiological study on risk factors for running injuries we needed tests that were relatively easy to administer (for logistical reasons). Furthermore, in the review of Menz [35] was stated that the navicular drop was possibly influenced by foot length. Nielsen et al. [36] found that foot length had a significant influence on the navicular drop in both men and women and that this could have been incorporated in the measurement protocol.’ (Page 18 and 19, line 396-412)

REFERENCES
35. References: Please review the format throughout. For example, the second reference contains ‘2007’ three times! The third and fourth references are both from the same journal; but, the one is abbreviated while the other is not. The ref #26 and 28 are books; but, the format is completely different between two.

We apologize and reviewed the references format and made it uniform conforming the guidelines of BMC.

TABLES & FIGURES
36. Table 1: Please remove. It is not a part of the study.

We removed.

37. Table 2: Please include height and remove percentage fat.

We included height and removed percentage fat.

38. Table 3-4: Please remove. They are not a part of the study.

We agree and removed.

39. Table 5: please explain why the authors decided to combine the right and left limb over three visits. Please be sure to include your rationale in the methods.

The data of the right and left limb were used separately in all calculations and we stated this in the methods. Because we now focus on the reproducibility and sex difference this table 5 was removed.

‘The data from the right and left leg and feet were used separately in all calculations.’ (Page 11, line 206-207)
40. Table 6: Please include 95%LOA values. MTP1 SEM values look extremely high. Please rerun the data and check all calculations.

We rerun the data, included the sex differences (independent t-test) and checked all calculations. Nevertheless, the SEM values of the MTP1-test stay very high and we discussed this now in the discussion.

‘The SEM for the interrater agreement of 34.4° is high although in line with the low reproducibility. Probably, it could have been helpful to calculate the intrarater reproducibility of examiner MW as well and so identify possible examiner inconsistencies [20], which could explain the high value of the SEM of 34.4°. However, we decided to provide only the intrarater reliability of examiner HD. We chose this option to limit the time involvement of the participating runners. The total time for the measurements of one runner was about an hour, including the breaks in between. If the other examiner had taken the tests twice for every runner, the randomization schedule had to be adapted and probably runners had to spend more than two hours while being measured. Furthermore, the SEM\textsubscript{consistency}, which not included the systematic error [24], gave a value of 9.1° and is more in the line with the findings of the intrarater agreement of this study. So, the high SEM\textsubscript{agreement} (34.4°) of the interrater agreement is possibly based on a systematic error.’ (Page 17, line 354-366)

41. Figure 1: Please replace with the nicer quality picture from the side view using a measuring rod.

We changed the picture, as stated earlier.

42. Figure 4-7: Please remove.

We re-arrange the figures, but kept the Bland and Altman plots, as stated before.
Reviewer: Angela Evans

1. To my reading, it is not clear as to why the selected participants were chosen i.e. whilst acknowledging the influence of lower limb biomechanics in running, injury, performance etc.; it was not clear for the purpose of a reliability study – which assesses the repeat accuracy of the tests (not the participants) - as to why these subjects were so purposefully selected and detailed. If there was the expectation that the runners as characterized are known/expected to differ from other same age/gender subjects, then this needs stating and supportive information (in terms of AJ range, ND, 1st MPJ). There are four Tables for participant characteristic data - if there is expected difference in any tested parameters in subjects with differing body fat levels, or who run on particular surfaces - this needs to be stated and clarified, as otherwise the abundance of this information seems superfluous in the context of assessing clinical test reliability.

We have chosen in this clinimetric study for a runner’s population because the NDT, AJD-test and extension MTP1-test seem to be important to identify runners with higher injury risk. Furthermore, there is a risk profile difference for running injuries between male and female runners, but the exact reason is not yet identified. This difference in risk profile may be related to the possible difference in ND, AJD and extension of the MTP1 in male and female runners. So we edited the introduction and gave the rationale for choosing these tests and determining the difference in ND, AJD and extension MTP1 between male and female runners.

‘In conclusion, there are many studies of the reliability of the NDT, AJD-test and extension MTP1-test and these tests seem to be important to identify runners with higher injury risk and for prevention purpose in runners. Because reliability is a characteristic of performance of an instrument in a certain population the reliability characteristics of these tests need to be identified in a population of runners.’

(Page 5 + 6, line 76- 80)

‘Furthermore, we hypothesized that there is a difference between sexes based on several runner studies [6,14-17] which showed differences in risk profile between male and female runners. This sex difference, regarding the musculoskeletal system, can partly be explained by the difference in NDT, AJD- and MTP1 mobility.’ (Page 6, line 95- 98)

We agree that the aim of the study was not very clear and that the information about the participants/runners was too much. We changed the title of our study, edited the introduction as well as the results and we focus now on two aims of our study:

‘Hence, the aim of this study was to develop and assess the intrarater and interrater reliability and agreement of the NDT, AJD test and extension MTP1 test in weight-bearing position in healthy runners. Secondly, we wanted to compare outcomes of these tests between female and male runners.’

(Page 7, line 106- 109)
2. It was curious to me that the specific clinical tests selected for this study, were, and were used as detailed in the methods. The protocols for both the ND and 1st MPJ appear non-standardized, and differ from better available and published methods (as is immediately acknowledged by the authors when referencing the Sell study as discussion).

The use of clinical tests for both foot and lower limb which involve skin markings have long been shown to be less reliable, yet such methods are used for both 1st MPJ and NDT. Similarly the protocols for both of these tests (1st MPJ and NDT) were ill-defined (no standard step length; sit to stand vs ST palpation; measurement rods). It has also been shown in many previous studies that the reliability of measures which involves manual manipulation of foot posture (ie ND with ST position / sit-stand; 1st MPJ with variable step length) is reduced as opposed to those with standardized and non-manipulated attributes eg AJ rom - which has been well tested previously (e.g. Munteanu, Evans, Bennell) with the standard weight-bearing lunge method.

We fully agree with the reviewer that it is important to use measurement tools with high reproducibility and that they should be performed in an optimal standardized way. We have chosen to adapt these three orthopaedic tests slightly, because we want to use them for a future prospective study with 400 participants who participate in a run. This future study focuses on the identification of risk factors for running related injuries in female runners. Because navicular drop, degree of dorsiflexion of the ankle and/or extension MTP1 appear to be potential risk factors, tests to determine these variables have to be administered rather easily and quickly without using (expensive) equipment.

In our future study, measurement time was limited to a maximum of 10 minutes for the three tests. In addition, the measurements had to be done on location and with minimal equipment. Because reliability is a characteristic of performance of an instrument in a certain population it was important to determine the reliability of these tests and to identified possible difference between female and male runners in the ND, AJD and extension MTP1. Furthermore, by adapting the existing protocols in literature we had to determine the extent of agreement.

The rationales for choosing these tests, adapt them and the reason to study their reproducibility are now stated in the introduction and method section.

‘For future etiologic cohort studies of runners it is important to identify whether (hyper-) pronation of the foot, decreased ankle joint dorsiflexion (AJD) and the degree of the extension of the MTP1 are risk factors for running injuries. To measure the extent of foot pronation, AJD and the extension of the MTP1, reproducible orthopedic tests are essential.’ (Page 4, line 40-44)

‘In this study, the protocols in the literature of the NDT [8], AJD-test [11] and extension MTP1-test [10] were adapted for the use in our planned prospective cohort study of female runners (n= 433). This adaptation was necessary for practical reasons, which required that these orthopaedic tests are performed in maximal 10 minutes, on location and with a minimum of measurement tools and equipment.

For the NDT [8], in our protocol a ruler was used instead of a blank card [12] and the sitting position was used as neutral position of the foot instead of palpating the talar head [13], so the NDT could be determined directly and the measurement
time was minimized. The performing times of the protocols of the AJD-test [11] and the extension of MTP1-test [10] were optimized by refraining from using a tapeline and standardized step length, but extra attention was paid to maximal stretch of the posterior leg and MTP joint, respectively.’ (Page 6, line 81-92)

3. Generally, tests involving the height of the navicular (static or dropped) are also normalized against foot length, such that foot morphology is made relevant (i.e. males often have longer feet than females) (e.g. Barton, Murley).

   We didn’t take the foot length into account and discussed this as part of the limitations of this study. The following text is included in the article:

   ‘Furthermore, in the review of Menz [35] was stated that the navicular drop was possibly influenced by foot length. Nielsen et al. [36] found that foot length had a significant influence on the navicular drop in both men and women and that this could have been incorporated in the measurement protocol.’ (Page 19, line 409-412)

4. The limited intra-rater involvement is curious, why was this data not collected by both examiners (and why only 10 minutes between sessions, more usual to separate sessions by hours rather than minutes). Whilst the inter-rater reliability is of wider interest, the value of the intra-rater data for all examiners is that it helps to identify examiner inconsistencies, between session differences - both/either of which may be particular for any/all of the tests being assessed.

   We agree that it is more usual to separate sessions by hours rather than minutes to avoid bias. However, we choose a minimum of 10 minutes between sessions for practical reasons and did not expect test differences based on runners differences but only on measurement differences. Furthermore, we expected that recall bias was minimized with a minimum of 10 minutes between measurement sessions because of the number of measurements and runner (3 tests, both feet and legs).

   We fully agree that it would be better to determine the intrarater reliability of examiner MW too. However, we decided to provide only the intrarater reliability of examiner HD. We have chosen this option to limit the time involvement for the participating runners. The total time for the measurements of one runner was about an hour, including the breaks in between. If the other examiner had taken the tests twice for every runner, the randomization schedule had to be adapted and probably runners had to spend more than two hours while being measured. We stated this point in the last paragraph of the discussion on the MTP1-test; page 17+ 18, line 349-385.

5. Statistical testing is adequate, but could include be improved with the addition of results/test for each session for each examiner (intra-rater and inter-rater).

   We included the results of each session for each examiner; see table 2.
Summary points for consideration:
Less focus on participant characteristics, unless justified.
Given the difficulty with navicular palpation, why is this test included?
The testing protocols need to be better managed for both NDT and 1st MPJ, as was the case for AJ rom (and based upon best existing methods e.g. Sell).
Does AJ rom really need to be re-examined, given the existing results from many studies?
The intra-rater for both examiners should be included.