Editor

- The reviewers offer useful recommendations on your manuscript which if addressed will improve its readability, and its subsequent impact. Could you please address their comments. In addition please consider the following observations,
    - Thank you for your time and consideration of this manuscript. The authors will address each of the reviewers' concerns below as well as the listed observations. Please note, the line numbers referred to in the rebuttal will be from the revised manuscript. We have submitted our responses to the reviewers in a point-by-point format, with our responses in red text and the in-manuscript edits inserted with yellow-highlights. Track changes were left on, as per requested.

- The manuscript focusses on reliability and validity; the concept of validity from a statistical perspective is not well established in the literature, but reliability is. In the manuscript what is meant by validity? Is this not simply accuracy?
    - In the manuscript, validity is referring to the accuracy of the automated technique compared to the manual technique. To further clarify that the use of validity in this manuscript is referring to the accuracy of the automated technique when compared to the manual counterpart, we have revised the secondary purpose to reflect we are evaluating the accuracy of the automated tool (lines 112-114)
        - A secondary purpose of this study was to assess the accuracy of the novel, automatic ultrasound analysis tool against the manual analysis equivalent.

- The manuscript reports measures of muscle architecture, with the measure of the of the number of sarcomeres in parallel identified as pennation angle, but this is really a proxy would muscle thickness be a better measure?
    - For the purpose of this manuscript, we solely wanted to compare fascicle angle and length as determined by the automatic tool versus then manual measurements of fascicle angle and length. Due to the automatic tool only selecting fascicle angle and length, we chose to only measure fascicle angle and length. Future work might consider measure muscle thickness.

Reviewer 1

- **Basic reporting**
- Overall the article is well written and able to be understood. I would add one thing to help clarify the abstract slightly. In the methods the low and high frequency images are explained (probe frequency) however, this is not explained in the abstract which can be confusing. Perhaps just noting it is probe frequency is changing would help.
    - Thank you for complementing the article. We appreciated the time and effort put into this review. After reading your comment, we agree that the clarity of the abstract may be improved upon and have added additional in the revised abstract (line 51-52). Specifically:
        - Images were taken at both lower (10 mHz) and higher frequency (12 mHz).

- **Experimental design**
- Overall, the methods are fine. The one issue I had was with the repeated measures stats. It was difficult to understand what measure was repeated. The authors have a pre and a post measure in their data tables and they have two probe frequencies and two scans with a 10 minute break in between. Were the repeated measures between the first and seconds scans (with the 10 minute break in between) or were they between measurements where the individual analyzing the data analyzed the same image twice? If the pre post is between the two scans then I am very concerned about measuring the same fascicle each time.
    - We noted your concerns with fascicle measurements and apologize for any confusion. To begin, the "pre" and "post" variables within the data table is specifically referring to the first and second scan session, with the 10- minutes break in between. We revised the data table and manuscript to denote the first scan session as trial 1 and the second scan session as trial 2. We analyzed the trial 1 and trial 2 images in lower frequency and then separately analyzed the trial 1 and trial 2 images in higher frequency. The location each image was taken was identical along the length of the muscle.
    - We agree with the reviewer, the average fascicle length is valuable and may minimize variability. However, we opted not to use the average fascicle length in the present investigation due to the fact the automated program only measures the length of one fascicle. Therefore, our goal was to mimic the automated program by selecting one fascicle and measuring it manually.
    - Lastly, it is noted as a limitation (lines 310-312) that we cannot guarantee the same fascicle and pennation angle was captured in both trial 1 and 2 (pre- and post-break). It is important to note, however, that this limitation remains valid in all studies utilizing B-mode ultrasound to assess muscle architecture.
        - The limitation section now includes the following (lines 315-319):
        - While averaging fascicles lengths and fascicle angles may reduce the variability of the measure, the automated program used in the current study only selected a single fascicle length and angle to measure. Therefore, the authors manually measured a single fascicle and corresponding angle. Future investigations may wish to average multiple fascicles for their analysis in order to reduce variability.

- Previous research (Infantolino and Challis, 2014) has shown pennation variability within one image so this variability alone could explain the pre-post differences. If the scans

change one would imagine this would become worse. Perhaps measuring multiple pennation angles per scan and averaging would help to reduce this issue.

  - Thank you for providing an insightful reference. As mentioned above, we agree. There may be added variability within the technique by utilizing only one fascicle vs. average of multiple. However, we feel that by doing so would not allow for the most appropriate analysis of our research question: comparing the performance of this automatic tool compared to an investigator. The automated program did not average multiple angles or fascicles, so we did not either. Again, this was clarified in the limitation section (lines 315 – 319).

- **Validity of the findings**
- Overall the findings are valid based on the stats. I would reiterate the issue I mentioned in the methods as a possible issue with the results in general. The other thing I noticed while looking at the raw data is that the automatically measured pennation angles were in many cases well outside the realm of reported pennation angles in the literature (>60 degrees). Is it possible to constrain the computer program to only look within a range of pennation angles (<60 degrees)? If so, would this greatly influence the accuracy of the results?
  - This is an astute comment and something the authors have discussed. Having carefully reviewed the capabilities of the SMA Macro, we are not aware of a "filter" for pennation angle within the automated program, but we agree that such large pennation angles are inconsistent with physiological outcomes; however, we chose to follow the guidelines set forth by the automatic tool to remain as "real world" as possible. We also added that future automated tools should filter for pennation angles that are too large and are found outside the range of angles reported in the literature (line 319 - 322).
    - Lastly, the SMA macro procedure produced some muscle architecture measurements which may not be physiological (e.g., PA > 60°). Future investigations of automated image processing techniques may benefit with the inclusion of a "filter" so that outlying fascicle angles would be removed. However, these values are left in the final dataset of the present investigation to maximize the current applicability of the SMA algorithm for the reader.

- **Additional comments**
- I like the idea behind this study and agree it is very important work. My two issues are the confusion with the repeated measures and the pennation angles that are well outside the realm of possibility.

- Thank you again for taking the time to review the manuscript and we are appreciative of your feedback. We hoped we have adequately clarified our position, and credit the feedback provided by both reviewers. If additional suggestions come up, we welcome additional revisions.

Reviewer 2

- **Basic reporting**

The authors did a great job in addressing the questions and significantly improved the article. I have only provided minor comments/suggestions to further improve the manuscript.

- General comments
  Thank you for the opportunity to read and comment on the manuscript entitled "Reliability and validity of manual versus automatic ultrasound analyses". The present study reports the reliability and validity of manual and automated methods to determine vastus lateralis architecture using B-mode ultrasonography. The manual analyses showed good reliability and the automatic analyses poor to moderate. However, absolute errors for both analyses were quite large, particularly in the automated analyses. The differences in outcome measures from the automated vs manual were very high and put into question the validity of the automated measurement program. These results are important for the field. Overall, the manuscript is well written, although there are sections requiring further improvement/clarification. I believe if they are met, it will form a much stronger manuscript. I also have some concerns regarding the methods used and more clarification on these are needed. Please refer to my specific comments below.
  - Thank you for your time and consideration of this manuscript. Your comments and concerns have been addressed point by point in the following rebuttal. We hope to clarify the manuscript based off the feedback from both reviewers.

- Specific comments
  Tittle: It should be immediately clear that you're testing the validity or previously published algorithms/automated analyses. I was under the impression that this was a newly developed coding/programming. This should be clear upfront.
  - Thank you for pointing out this confusing issue. We do not wish to detract from the original work nor claim it as our own. Seynnes and Cronin (2020)

deserve said credit for their hard work in creation of this script. After consideration, we decided to change the original title to "Reliability and Accuracy of Manual Versus Automatic Ultrasound Analyses". We have also revised the manuscript to reflect how we assessed accuracy of the automatic tool compared to the manual technique.

R: Thanks for clarifying this. I would suggest including the automatic tool used, i.e., Simple Muscle Architecture Analysis, because multiple automatic analyses exist, and this will deceive the reader to believe you compared all of them. I'd narrow it down.

- Abstract: The authors should refer to fascicle angle instead of pennation angle. While pennation angle is generally used, it is conceptually wrong. Pennation angle is the angle between the fascicle and the muscle's force line of action – tendon. Given that in pennate muscles, the aponeuroses are not necessarily in line with the tendon, we measure the angle of the fascicle instead of the angle of pennation. Please consider changing this throughout.
    - We have changed each use of pennation angle and its corresponding acronym of PA to fascicle angle and FA throughout the manuscript.

- In the conclusion, I believe it's better to provide a more general statement. What are the implications of these findings to the field? Why are they impactful? I also don't see any inferences regarding the validity and the large errors in the automated analyses. Please consider including.
    - Thank you for the feedback. We have tried to revise the conclusion to have a more general statement in the abstract. The new sentences are added in lines 64-71. Specifically (new sentences highlighted in yellow):
        - The findings overall show that manual analyses had good reliability and low absolute error, while demonstrating the automated counterpart had poor to moderate reliability and large errors in analyses. These findings may be impactful as they highlight the good reliability and low error associated with manually analyzed ultrasound images and validate a novel automatic tool for analyzing ultrasound images. Future work should focus on improving reliability and decreasing error in automated image analysis tools. Automated tools are promising for the field as they eliminate biases between analysts and may be more time efficient than manual techniques.

R: Can you conclude whether, from your results, you would recommend performing one technique over the other? Perhaps the lower reliability and large errors from the automated analyses mean any research using it should perform more trials and analyses more fascicles given the large errors?

- Introduction:
  L91-92: This sentence is confusing. Please consider re-writing it
    - Thank you for pointing this out and the opportunity to clarify. The sentence was rewritten as follows in lines 91-92:
        - Previous studies have shown that muscle structure which influence muscle size (i.e., FL and FA) are important to whole muscle function and force production (12,15).
- L94: this sentence is also unclear. Assessment of muscle architecture doesn't directly imply number of sarcomeres in parallel and in series. It's just that in pennate muscle we can fit more fascicles within the same space than a non-pennate muscle and that in lesser pennate muscles fascicles are often longer.
    - We agree, the previous iteration was, perhaps, too strongly worded. The sentence in line 93-94 has been revised to:
        - Specifically, muscle architecture can provide insight, via proxy, regarding the total amount of sarcomeres in parallel (i.e., fascicle angle) and in series (i.e., fascicle length) (13,20).

- L94-97: Please revise the sentence.
    - Lines 94-98 were revised to state:
        - Previous works have shown that in pennate muscles (e.g., vastus lateralis, gastrocnemius), accounting for fascicles which are arranged at an angle relative to the tendon may be valuable as changes in fascicle angle can increase or decrease the mechanism of force generation within the muscle (8,11,15,20).

- L106-107: The authors suggest there is insufficient data regarding reliability and validity of automatic ultrasound analyses. I believe that published new automated analyses package/software are always validated at first glance. See Cronin et al 2020 (deep learning), Seynnes & Cronin 2020, Trackamte, UltraTrack (Farris & Litchwark

2016), etc. However, most of these manuscripts use single images with limited field of view, ie, they do not assess them using extended field of view.

- We agree with this reviewer, most previously published works introducing new techniques have included indices of validity. However, our study attempts to add to the literature by including both reliability and validity metrics. Specifically, for the purpose of the present manuscript, we wanted to specifically determine the reliability of an automated image analysis tool (i.e., SMA macro in Fiji) and validate it against manual techniques. We revised the sentence (lines 108-110) to highlight that reliability and validity statistics were lacking between automated and manual techniques. Specifically:
  - While the automatic image analysis programs may bring many benefits to future studies, insufficient data exists regarding reliability and validity of automatic ultrasound analysis tools compared to manual techniques.

R: Do you mean accuracy? Again, I don't think you're testing whether the research is valid or not. I believe that in order to test validity, you would need to compare your analyses to cadavers to validate the measurement.

- L109: the authors suggest that only fascicle angle and length are part of the architectural measurements. However, muscle thickness is also very important and is strongly associated with maximal strength capacity and estimates of hypertrophy and its changes with training. I wonder why this has not been included in your analyses?
  - Indeed, the authors agree that muscle thickness is an important variable for muscle function and measure of muscle size; however, the aim of the present study was to compare this specific automated tools' ability to assess muscle architecture (i.e., FL, FA) and not muscle size, which could be defined a number of ways (e.g., muscle thickness, muscle cross-sectional area). Therefore, we *a priori* elected to only include measurements of FL and FA and exclude considerations of muscle thickness.

- L110-11: there's a problem with this statement. While it is generally accepted that automated analyses are compared against the manual ones to determine validity, this comparison will be highly dependent on the rater's experience. If you have a sonographer that has no experience, the images will unlikely to be valid. But if the sonographer is

highly experienced, it not only captures better, more valid, images but it also analyses them much better. However, one could also argue that the automated analyses from images obtained from both a non-experienced and experience sonographers and provides equally good/bad results, it will prove to be a valid coding/software. Haven said all of this, I what is the sonographer's experience, how many images and analyses has the sonographer collected prior to this study, and why have you not included another rater to perform to capture the images and do the analyses? I also wonder if these analyses were blinded? All of that would have strengthen the methodological design of the paper.

- This concern is well received, underdiscussed in the field, and a great point to discuss. In fact, this was recently reported on by Carr et al., 2021 (*J. of Functional Morphology and Kinesiology*). This investigation (Carr et al., 2021) reported that experience with image acquisition and analysis has small effects ($d < 0.30$) and good-excellent interrater reliability for measures of muscle size (i.e., thickness, cross-sectional area). The investigator who acquired and analyzed the data in the present study has separately acquired and analyzed more than 1,000 B-mode ultrasound musculoskeletal scans. Collectively, the three investigators have completed multiple studies using B-mode ultrasound using the techniques detailed in the present work. For the purpose of comparing reliability and validity of our manual technique and the automated tool, we felt a single rater and sonographer was sufficient and was not likely to significantly impact the research question. Additionally, the images were unblinded during analysis. As this investigation is not longitudinal in nature (*i.e., not looking at an intervention effect*) we do not feel this would have significantly improved upon the methodological strengths of the present investigation. With that said, we believe this is a point worthy of the reader's attention and have added the following sentence to our limitation section (lines 312 – 314):
  - The present work utilized a single, unblinded investigator to analyze all ultrasound images. Future investigators may wish to consider the effect of multiple raters on the reliability and validity statistics presented.

- In your hypothesis statement, can you provide an estimate of reliability beforehand? That, is, would you expect the reliability to be poor or good? And what sort of reliability are you referring to? This should be similarly applied to validity. These need further clarification….

    - We rewrote our hypotheses to state that the automatic analyses would demonstrate good reliability and less error when compared to the manual counterpart (lines 116-118):
        - It was *a priori* hypothesized that the novel, automatic ultrasound analysis tool for the assessment of muscle architecture would have good reliability ($ICC_{2,1} > 0.75$) and less error compared to the manual counterpart.
- Methods:
- Why did the authors perform analyses in 23 participants only? Do you consider that enough for reliability and validity purposes? And why? Can you please justify your choice
    - Following previous literature for reliability work (See Mota et al., 2015, Marzilger et al., 2018, among others), a convenience sample of 23 participants were believed to be enough to complete and execute both reliability and validity statistics. Furthermore, this sample was over 50% larger than that of original work on the SMA script. For this project, we weren't looking, or powered, for an effect. Instead, we wanted to strictly evaluate the reliability and validity of a novel, automated tool compared to manual image analyses.

- Also, why was the study conducted on this specific population? If you're determining the reliability and validity of a measurement, would you not want to test it in a population with very heterogenous characteristics? This would have allowed the results of the present study to be extrapolated to a broader range of population
    - As described above, the present work recruited a convenience sample of 23 participants from the university's community. While the sample may be considered somewhat homogenous in nature (i.e., young, physically active), this sample included both males and females while demonstrating considerable variability in demographic variables (range age [18 – 31 years], mass [49.9 –110.3 kg]). Additionally, it is important to note that there is still considerable range in the reported physiological outcomes (10mHz; trial 1; manual FL = 4.7 – 10.5 cm, FA = 12.4 – 29.0 °). Also, this study was collected during the COVID-19 pandemic. During this pandemic, our university set forth guidelines and recommended only low risk individuals who were already on campus should participate in research.

- The experimental design needs further information. I only understood that Trials 1 and 2 were used for reliability half-way through the methods. I think the sentence 124-125 is confusing: two independent data collection trials…. I think this could be improved to suggest you collected vastus lateralis images twice, 10 min apart, and that the images obtained from each time point were used to calculate fascicle angle and length and then estimate the within testing session reliability.
    - Thank you for helping to make the experimental design clearer. Lines 135 – 136 now reads:
        - The independent data collection trials consisted of identical methodologies and were subsequently used for the calculation of reliability statistics (see below).

- L126: Technically, ultrasound was used to capture images of the muscle of interest. The images allowed then for architecture to be calculated. Be clear.
    - We clarified the sentence to explain that ultrasound was used to take images of the vastus lateralis muscle architecture was later assessed with image analysis programs (lines 130-133).
        - During each trial, Brightness mode (B-mode) ultrasound was used to take images of the vastus lateralis (VL), where muscle architecture (i.e., fascicle angle [FA] and fascicle length [FL]) of the VL was later assessed with open-source image analysis programs.

- L139: is this a secondary aim? Why was it done and why is it important? Please clarify that in the introduction. The reader won't know why this was done.
    - To help with clarity, we added a sentence in the last paragraph of our introduction to explain our tertiary purpose (lines 114 – 116):
        - Lastly, a tertiary purpose of this study was to examine the difference in reliability between images taken at a lower frequency (10 mHz) and higher frequency (12 mHz).

R: Thanks for clarifying this. However, it is still unclear why you have done that. Please provide a rationale to the reader.

- L142-143: Two images per trial? So what did you do - took the average of fascicle angle and length for both? Sorry this is unclear
    - We revised this portion of our work to explain two images were taken and single the clearest image was used for analysis (lines 152-156):
        - During data collection, a minimum of two quality images were taken per ultrasound frequency. A quality ultrasound image is defined as an

image where the outer borders of the muscle of interest are clearly seen as well as fascicle that is clearly captured from superficial to deep aponeuroses. During offline-image analysis (see below) the single clearest image was used for processing.

- L143-144: What does "outer borders of the muscle" mean? The upper and lower apo? And what if you clearly see them but your probe is off angle and therefore fascicles aren't clearly delineated? I'd assume the main interest here is to see "fascicles" and therefore you wanted to guarantee fascicles were clearly visible.
    - We edited the sentence to clarify what a clear image referred and that the aponeuroses and fascicles could be seen in the image (lines 153-155):
        - A quality ultrasound image is defined as an image where the superficial and deep aponeuroses of the muscle of interest are clearly seen as well as fascicle that is clearly captured from superficial to deep aponeuroses.

- L144: Please refer to the ultrasound images provided here.
    - "Quality image" was revised to state "quality ultrasound image" (line 153-155):
        - A quality ultrasound image is defined as an image where the superficial and deep aponeuroses of the muscle of interest are clearly seen as well as fascicle that is clearly captured from superficial to deep aponeuroses.

- L153-155: I'm concerned with this - vastus lateralis fascicles are not straight - they are curved and therefore you would be adding noise to your measurement if a straight line was considered. This will influence the validity of your measurement and affected the automated analyses. Please refer to Noorkoiv et al (2010, JAP).
    - Thank you for voicing your valid concern; however, it is important to note that Noorkoiv et al. 2010, states that ultrasound using an extended field of view function is valid and reliable method for assessing muscle fascicle length. In the present study, we followed similar methods to Koorkoiv et al. 2010 by capturing muscle fascicles with extended field of view technique. Also, previous papers in the field assess muscle fascicle length using ultrasound and a straight-line tool (Kawakami et al. 2001 [Canadian JAP]; Franchi et al. 2020 [MSSE]), so we felt the same methodology should be used. Lastly, and most importantly, the methods employed in the present study were

chosen to most closely match the SMA algorithm. Therefore, the two comparisons were done in as similar and controlled ways as possible. However, we believe this may be of worthy note to the reader and have therefore added a note to the limitation section (lines 308 – 310):

- In the present study, it could not be ensured that the fascicle selected for the manual and automatic technique was identical nor that either technique assessed potentially curved fascicles without bias.

- L163: How is your manuscript different form the authors of this macro – ref 21? They showed validity of their automated programming in the ref 21
  - The authors of macro listed as reference 21 explain their validity measurements are of pennation angle and fascicle length of the gastrocnemius medialis and tibialis anterior in two separate samples. The authors used 15 participants to take images at 9 MHz with a 96-element transducer. For the second sample, 15 different participants were used, and images were taken at 12 MHz with a 128- element transducer. In the present study, we evaluated the validity of fascicle angle and length measurements in the vastus lateralis with the same participants for each trial and the same ultrasound probe (i.e., (L4 – 12t - RS, 4.2-13 MHz, 47.1 mm field of view).

- L171-172: If the automated program cannot identify fascicle angle and length from your "best" chosen image, it suggests that either (1) your image had poor quality or (2) the software doesn't do what it should be doing. Please clarify on this. It is also important to report how many times this occurred.
  - To help clarify, we took a minimum of two quality images, so that all muscle aponeuroses and fascicles could be viewed. We do not understand why the automated tool's algorithm could not highlight and measure a correct fascicle angle and its corresponding length, as the SMA tool does not express this information. Therefore, we will not speculate. We followed the same methods and protocol for every image analyzed and the macro was successful for other participants and images. Please note, we did explain how many times this error occurred in lines 182-184 (i.e., only one time out of 92 analyses). Therefore, and with respect, we do not believe image quality was limiting factor based off the low amount of failed automated analyses.

- Statistical analyses:
- L177: On what basis was this ICC chosen?
  - The topic of ICC model selection is often under discussed, and we thank the reviewer for this opportunity. While there are many models available for

implementation (Shrout and Fleiss 1979; McGraw and Wong 1996), the ICC model 2,1 was chosen as it allows for usage with future investigations by other laboratory groups (via its usage of both systematic and random error within the calculations). In brief, this model assumes our rater was selected from a larger population of raters (i.e., one investigator chosen out of the larger population of investigators). Therefore, other researchers may assume that our reported reliability statistics (i.e., ICC, SEM) may be applied to their methodologies (if similar). The authors do not believe that other models (i.e., ICC 1,1 or 3,1) would be appropriate for our research question.

- L178-179 & 181 & 185-189: Clarify how all the reliability measurements were calculated. The reader doesn't need to check Weir's paper to know how these were calculated. This also applies to the validity statistics.
    - We appreciate this comment, but the authors believe it would not be feasible to go into detail about reliability (i.e., ICC, MD, SEM%) and validity (i.e., SEE, TE, CE) calculations as this would require a substantial word-count investment and is outside the aim of the present work. The references for the works explaining the calculations are listed for the curious reader. However, we leave this potential option to the editor's discretion and remain open to future requests to revise our work.

R: I am sorry, but I disagree with that. The purpose of your manuscript is to determine reliability and accuracy so the information as how you produced those statistics it is crucial. For example, MD is calculated using used SEM, and there are at least two ways of determining SEM (typical error) which produces different numbers. While this is often under discussed, it affects MD and the conclusions drawn from its information. It is thus important to clarify the methods used. Given that, I strongly suggest adding some basic information on how the calculations were performed, "e.g., SEM was determined as the mean/SD x 100 (%), minimal difference (MD)..". I see this information has been added to statistics related to accuracy, so why would you not inform that for the reliability?

- L187: Why trial 1 only? Why didn't you repeat these to increase the number of comparisons to validate the automated analyses?
    - Standard practice when evaluating validity or agreement is to compare techniques from a single time point/measure. Including two trials for the same individual would not add any "new" data to the model. This could potentially

add bias to the results as we would essentially be including the same participants data in the model twice presented as two completely separate data points which they would not be. Therefore, we only included results from one trial for agreement analyses.

- L193: ICC of 0.75 and SEM and MD of 14% and 7.4 degrees were reported. Does this suggest that, if an acute intervention was performed, you would need to change fascicle angle for at least 7.4 degrees for a change to be considered real? This is very impressive and practically impossible - changes in fascicle angle after fatiguing contractions or stretching are ~2.5 degrees. Do your data suggest they are possibly not real changes?
  This should also be discussed further in the Discussion section.
  - Thank you for concern with the changes in fascicle angle reported in our study. We chose to use the minimal difference to be considered real as the statistical calculation for the minimal difference to be considered real is used at the individual level, not the group level. In other work, fascicle angle has increased following a fatiguing bout from 23.5 ± 4.1 to 26.3 ± 2.2° (Mademli et al. 2005). The variability in the present study may have been less if the authors averaged fascicles. As mentioned previously in our responses, we chose to select a single fascicle for analyses to copy the SMA tool selecting a single fascicle for analysis.

- L204: Somewhere it should be mentioned that the results show moderate reliability but very high errors!
  - We appreciate the excitement contained by this reviewer. A concluding sentence was added at the end of the reliability results section for both the manual (lines 211-212) and automatic analyses (lines 222-223).
  - Manual:
    - Further, the manual analyses for both FA and FL had good reliability with low absolute error.
  - Automatic:
    - Taken together, the automatic analyses for both FA and FL had poor to moderate reliability with moderate absolute error.

- L214-215: Do you consider these errors acceptable? They seem very high
  - We consider these errors to be acceptable as they are moderate errors, but more work needs to be completed to understand what errors are acceptable for

automated tools. We are not aware of a well-accepted reference that categorizes ranges for errors. However, if the reviewer is aware of a source, we would appreciate their insight.

R: That's a great point. I don't know either and was hoping the authors would. However, in one of my comments above I was hoping to stimulate a discussion around the effect of acute or chronic intervention on changes in architectural properties and the results produced here. That is, the changes in FA and FL observed after acute and chronic training are often within or below the range of errors and reliability reported here. Thus, I wonder if within the context of changes in FL and FA with training and detraining, overall, these results suggest errors to be unacceptable.

- L224: Can you speculate why the validity is worse for different frequencies?
    - The ability for ultrasound to penetrate and echo through deeper tissue is inversely proportional to its frequency. Yet, the image quality is noted as being proportional to frequency. That is to say that higher-frequency ultrasound will produce higher-quality (i.e., clearer) images in superficial tissues. Yet, when deeper tissues are targeted, lower frequency may be required to maintain image quality. When considering the literature as a whole, the authors of the SMA tool suggested the program can fail to detect FA and FL when there is not a stark contrast between light and dark tissues in the images (i.e., lacking heterogenous echogenicity). Perhaps the higher frequency caused less contrast between tissues; therefore, leading to more error and causing poorer reliability compared to lower frequency.

R: Thanks for addressing this. Can you include something like the above in the manuscript? I believe it fits well and strengthens your discussion. My apologies if you have included, though.

- Based on your figures, it seems to me that the only variable showing proportional bias is Figure A. Lower Frequency fascicle angle. All the others seem to increase the bias with increases in values.
    - The regression lines for Bland-Altman plots (B, C, D) were statistically significant, indicating proportional biases for said variables. For figure A, the regression line was not statically significant, indicating no proportional biases

(i.e., a flat regression line). The nonsignificant regression line indicates the method utilized did not impact FA measurements.

- Discussion:
- L230: You need to make clear here that this is not your coding/program but that it has been published before. I suggest "...novel, previously published, automatic...."
    - The opening sentence of the discussion was edited to reflect this suggestion in lines 244-245:
        - This study examined the test-retest reliability of a novel, previously published, automatic ultrasound analysis tool for the assessment of VL muscle architecture (i.e., FA, FL).

- L233-236: Please also refer to the SEM and other absolute reliability for this interpretation. ICCs only suggested that all subjects within the sample varied proportionally and therefore provides good between-subjects reliability; however, the absolute reliability is poor considering the expected changes in architecture with acute and chronic interventions.... this needs to be clarified.
    - The first paragraph of the discussion was revised to refer to the SEM and MD as well as ICC for both manual and automatic analyses in lines 247 - 255.
        - The findings from this study show that manual ultrasound analyses for FA and FL for both lower and higher frequency displayed good reliability ($ICC_{2,1}$ = 0.75 – 0.86) and low absolute error (SEM% = 9.99 – 13.69%). The MD for manual analyses for FL ranged from 2.16 to 2.48cm, while the MD for the FA ranged from 5.38 to 7.40°. However, automatic ultrasound analyses for FA and FL revealed moderate reliability ($ICC_{2,1}$ = 0.61 – 0.72) for the lower frequency images, poor reliability ($ICC_{2,1}$ = 0.16 – 0.27) for higher frequency images, and moderate absolute error for both lower and higher frequency images (SEM% = 18.75 – 31.38%). The MD for FA ranged from 8.79 to 12.12°, while the MD for FL ranged from 6.13 to 6.87 cm.

- L238: Tell the reader what TQ and SEE means.
    - Unfortunately, we are unaware of what TQ is and we are assuming the reviewer is referring to TE in the results section. We defined all acronyms the first time they were used in the manuscript. Total error (TE) and standard error of the estimate (SEE) were defined in the statistical analyses paragraph (lines 193-194).

- L241-242: What do they add to? What's the implications of these findings? Please expand on this
    - We added information regarding the implications of the findings in lines 259 – 262:
        - These findings add to an emerging body of literature examining automatic ultrasound analysis tools, specifically the validity and reliability of an automated tool when assessing the muscle architecture of the vastus lateralis.

R: Thanks for addressing this. I wonder if the authors should refer it to accuracy throughout the manuscript instead of validity?

- L246-247: is this low? Which variables are these?
    - We modified the sentence to highlight the muscle architecture variables were FA and FL in lines 265 - 267:
        - In the current study, manual ultrasound analyses indicated good reliability when assessing muscle architecture (i.e., FA and FL) ($ICC_{2,1}$ = 0.75 – 0.86) with low absolute error (SEM% = 9.99 - 13.69).

- L249: The minimum provided here is significantly lower than the minimum SEM provided by the present study. Please discuss.
    - The previous work referenced (Kwah et al. 2013 [JAP]) had SEM% that ranged from 4.3 to 14.2. The present work had SEM% that ranged from 9.99 to 13.69% (line 267). The current studies range for SEM% lies within the previously reported ranges in the literature.

- L251: Does this fit with the criterion adopted here? 0.63 does not seem an excellent score reliability.
    - We stated the ICC reported showed moderate to excellent reliability and presented the range of 0.63 to 0.91. The moderate reliability is 0.63 and excellent reliability is 0.91. These still fit with the criterion we used (Koo and Li 2015 [Journal of Chiropractic Medicine]).

- L255-256: These errors are large (19-31%). This is within the range of muscle architectural changes with training and detraining.
    - As explained previously, the errors reported for the automatic analyses are moderate and we were also surprised by the results. The moderate errors could be due to the automated program only measuring one fascicle or not correctly

tracing a fascicle as described in our limitations section. To decrease errors in future work we have added this suggestion to our limitations section (lines 319-322:

- Lastly, the SMA macro procedure produced some muscle architecture measurements which may not be physiological (e.g., PA > 60°). Future iterations of automated image processing techniques may benefit with the inclusion of a "filter" so that outlying fascicle angles would be removed.

- L2625-263: Why did you only analysed one fascicle? If this information was known prior to the study being performed, then you would expect to do something similar so you can discuss the findings to that specific study.
  - We chose to analyze and measure a single fascicle in order to mimic the single fascicle that was measured by the automated program. The automated program did not average angles or fascicles, so we did not either. This was added to the limitations section to clarify why we only chose to analyze a single fascicle in lines 315 - 319:
    - While averaging fascicles may reduce variability of the measure, the automated program used in the current study only selected a single fascicle to measure. Therefore, the authors manually measured a single fascicle and corresponding angle. Future investigations may wish to average multiple fascicles for their analysis in order to reduce variability.

- L264-266: Why didn't the authors expand further and analysed multiple fascicles? If this is indeed quick, it shouldn't have taken longer to analyse another fascicle of the same image. Please justify.
  - As mentioned previously, the goal of this current project was to compare the manual analyses to the automated program. With the automated program only selecting a single fascicle, we chose to copy the automated program and manually analyze a single fascicle. While it may be logical to think to just "run the program again" to find an additional fascicle, the SMA algorithm will often select the same facile for subsequent attempts, therefore making it difficult to truly find an average from multiple fascicles based off of this method. To keep the comparisons as close as possible, we elected to ensure both manual and automatic processes included identical calculations.

R: Thanks for clarifying that. I would strongly suggest explaining that in the manuscript too. This might help the developers to make changes to the program and "teach it" to not select the same fascicle if a repeated images was to be opened.

- L269-271: This seems an unacceptable error and invalidates the automated program. On top of these differences between manual and automated analyses, you are also expecting a big between trial variation, further reassuring the inability of the automated program to accurately and reliability determine fascicle length and angle.
    - We disagree that this is unacceptable error as we suggest the error as low but agree that the automated analyses resulted in larger error than we hypothesized. As the goal was to validate the automated program, we were simply trying to assess the error and variability of the automated tool compared to the manual equivalent. We believe future reliability and validity work should be completed in automated programs.

- L274: What is a true validity calculation? Please be clear
    - We describe true validity calculations as constant error (CE), total error (TE), and standard error of the estimate (SEE) which help understand the accuracy of the measurement. To make it more clear to the reader, we can incorporate that CE is the mean difference and we can add the calculations for total error and standard error of the estimate. (Lines 293-294):
        - In some of these previous works which report validity, traditional validity metrics (i.e., CE, TE, SEE) were not employed.

- L281-283: Why have you not performed those calculations to make it, at the least, comparable with the literature?
    - We completed validity calculations that were common with validity literature. We were still able to compare with ultrasound and musculoskeletal literature because we used calculations such as total error, which is mathematically the same as root mean square error.

- L289-293: Why were they not identifiable? If the images are performed under the same conditions, by the same experience investigator, would you not be able to place an image side by side and tell whether the region of the fascicles chosen are matched and therefore select the same fascicle? I also think analyses of a single fascicle is

complicated. You may need to include that in your limitations. How can you guarantee the fascicle chosen was representative of the fascicles of the muscle?

- We are not sure what the reviewer is saying is "not identifiable". After performing a search in our original submission, we cannot find an instance where we use "identifiable" within our text. However, we will assume the reviewer is commenting on the fact that the investigators could not guarantee that the fascicle selected by the SMA program would match that chosen by the investigator. As we have expressed in other rebuttal comments and throughout our paper, we aimed to mimic the technique to assess a muscle fascicle between the manual and automated techniques. However, we did not let the fascicle selected by the SMA program to influence our selection process. In other words, we did not run the SMA program, then use said outcome to lead us to measure a similar fascicle. We have added to our limitations to express that the measurement of a single muscle fascicle may be a point worthy of consideration by future investigators in lines 315- 319:
  - While averaging fascicles and fascicle angles may reduce variability of the measure, the automated program used in the current study only selected a single fascicle length and angle to measure. Therefore, the authors manually measured a single fascicle and corresponding angle. Future investigations may wish to average multiple fascicles for their analysis in order to reduce variability.

- L303-304: If the errors were large, would you consider this valid? The concept of validity needs to be clarified.
  - We appreciate this comment. When considered in conjunction all other comments, we have made changes throughout the manuscript which may impact this specific comment. If the reviewer still has concerns, we are happy to address in a future revision.

- L306: Theres no mention of how these are applied to training, detraining, or acute interventions. A recent paper has shown that the changes in muscle architecture after disuse are method dependent (Sarto et al 2021 - 10.1249/MSS.0000000000002614). Considering the large errors showed here, this should be discussed!
  - The authors agree with this reviewer. Though application to training, detraining, and acute interventions are important, we felt going into detail about application would be outside the scope of this manuscript. The purpose

of the manuscript was to examine reliability and validity of a novel, published, automatic image analysis tool.

- Conclusion
- L298-299: Can you please provide an explanation for this? Can you speculate as to why this is impacting on your measurements?
    - As discussed above, higher-frequency ultrasound will produce higher-quality (i.e., clearer) images in superficial tissues. However, when deeper tissues are being assessed, lower-frequency ultrasound may be required to maintain image quality. Also, the authors of the SMA tool suggested the program can fail to detect fascicles when there is not a stark contrast between light and dark tissues in the images. With that said, the authors would prefer to reduce the amount of pure speculation without additional evidence and would like to therefore limit said speculation within our paper. If this reviewer and editor feel otherwise, we would be happy to revisit this point in future revisions.

- References:
- Please check your reference throughout so that tittles are standardized. See ref 3 vs 2 - tittle has capital letters in each word whereas the other doesn't.
    - The references were revised.

- Figures/Tables
- Figure 1: If A and C are the same participant, there are clear differences in where the fascicle starts and ends and how the aponeuroses are determined – this is important because it significantly affects fascicle length and angle measurements.
  It is unclear what the boxes in A and B are.
    - Indeed, we agree that there are clear differences in where the fascicle starts and ends in A and C. At its core, this is the point of the paper and we agree with this reviewer.
    - To clarify, these images are of the same participant. As described in the figure legend, **A** is the automated program's measurement of fascicle length and angle during trial 1, **B** is the automated program's measurement of fascicle length and angle during trial 2, while **C** is manual measurement of fascicle length and angle during trial 1, and **D** is manual measurement of fascicle

length and angle during trial 2. In A, the reader can see the fascicle measured is not a true, physiological fascicle and perhaps could lead to why there was poor reliability and moderate errors in the higher frequency images analyzed automatically. As explained in the original work (Seynnes and Cronin 2020 [PloS ONE]), the boxes are explained as regions of interest used in automatic calculations of FA and FL.

- Figure 2: The x-axis of Figures A-D should be the average of the variable name, e.g., "average fascicle angle (degrees)". Additionally, on the top of the figure, can you please write the variable name and in parentheses the US frequency?
    - We have modified figure 2 to now include the name of the variable with the US frequency in parentheses. However, at this time we respectfully disagree with the reviewer's request to the change of x-axis titles to "average fascicle angle (degrees)". We have written the x- and y-axis using traditional Bland-Altman conventions. Similarly, we would prefer to use "mean" over "average". Furthermore, we believe it is more appropriate to use º and "cm", as is done throughout our manuscript instead of spelling them out for consistency purposes. However, if the editor disagrees with this conclusion of the authors, we are happy to address this in future revisions.

R: Thanks for clarifying. The Figure 2 caption still contains PA rather than FA.

- Table 2: since it's unclear to me how you calculated any of these validity outcomes, it's hard to tell why this is in cm. The total error a can also be calculated as % value or arbitrary units, depending on how it is determined. Please clarify.
    - Our validity statistics were calculated for FA in degrees (as shown in Table 2) and FL in centimeters (as shown in Table 2) due to us and the automated tool measuring FA in degrees and FL in centimeters, respectively. For clarity, we have explained and shown the calculations for validity in the statistics section (lines 196-199):
        - The error calculations were completed as follows: constant error (CE = criterion (manual analysis) – comparison (automated)), total error (TE = $\sqrt{\sigma}$ [comparison-criterion]^2/n), and standard error of the estimate (SEE = $\sqrt{\sigma}$ [comparison-criterion] × $\sqrt{1 - r^2}$).