Author’s response to reviews

**Title:** Concurrent validity of the Fitbit for assessing sedentary behavior and moderate-to-vigorous physical activity

**Authors:**

Nicklaus Redenius (nicklaus.redenius@ndsu.edu)

Youngwon Kim (Youngwon.Kim@utah.edu)

Wonwoo Byun (won.byun@utah.edu)

**Version:** 1  **Date:** 10 Oct 2018

**Author’s response to reviews:**

We appreciate the thoughtful and thorough reviews provided by both reviewers. We have carefully read all of the reviewer’s comments and have responded to each recommendation. We believe our manuscript has become stronger as a result of the modifications outlined below. We thank the reviewers and the editorial staff for their contributions to improving our manuscript.

We would like to acknowledge that in addition to the point-by-point responses to the reviewers’ comments below, we performed some general editing of the text in order to further improve the quality of the manuscript. Further, in the process of addressing the reviewers’ comments (data normality check), we identified two (N=2) outliers and removed them from the all of statistical analyses in the revised manuscript. As you will notice, the exclusion of two outliers yielded no impact on the results, but just slight changes in numbers reported in the tables. We would like to also acknowledge that we added an additional author, Dr. Youngwon Kim, during the revision process and it is also reported in the ‘Request for change in authorship form’ in re-submission.

Please note that the authors’ responses are provided in blue font color throughout this document and modified texts are highlighted in yellow in the revised manuscript.

Reviewer 1

Comments to Author:

Overall comments:

1. Useful research question, defensible methods particularly with respect to handling of sleep and non-wear time, and analyzing only minutes that appear valid from both devices.

We appreciate the reviewer time for reviewing the manuscript and providing valuable comments!
2. The difference in MVPA estimates is huge, and I think the authors should highlight this a bit more. The Fitbit is giving more than 3x the MVPA as the vertical-axis GT3X+ cutpoints and more than 2x the MVPA as the vector magnitude cutpoint. That's a huge difference and an important finding.

Thank you for this comment. As both reviewer suggested, we’ve conducted additional statistical analyses including mean absolute percent errors, Pitman’s tests, and equivalence tests. We believe that results from those analyses highlight the significant difference/non-equivalence in estimating MVPA between the Fitbit Flex and the GT3X+. We’ve also made extensive edits throughout the manuscript to outline the observed non-equivalency between two devices.

3. Authors are careful to say that Fitbit overestimated sedentary time relative to the GT3X+, but in the abstract, it implies that Fitbit is overestimating the true MVPA, which is unknowable (although it seems likely here).

Thank you for this comment. In the revised abstract, we’ve now explained clearly that the estimates from the Fitbit was compared with the estimates from the GT3X+.

Background:

4. Should give reference for GT3X+ validity and clarify what makes it have "high feasibility."

Thank you for pointing this out. We’ve changed the term “feasibility” to “reliability” and now included additional references in the revised manuscript.

5. I would cite a study other than NHANES that used the GT3X+, since the NHANES 2011-2014 data is still not available.

Thank you for this suggestion. Additional relevant citations are included in the revision.

6. IC should be introduced as indirect calorimetry.

Updated as recommended.

7. I would suggest specifying the Freedson and Troiano MVPA cutpoints when they are first mentioned, or at least somewhere - I don't think they are currently given.
Thank you for the suggestion as this will improve the clarity of the methods. The cut-points are now defined in the data reduction section (page 8, first paragraph).

8. I don't think duration of activities would affect the cutpoints. The type of activities, but likely not the duration, unless longer duration leads to lower intensity, in which case the counts ought to decrease.

Thank you for this comment. We clarified this as only pertaining to activity type (page 3, second paragraph).

Methods

9. Using an unequal variance (i.e. Welch) two-sample t-test is fine, but unequal sample sizes is not a good justification. The choice should be based on whether you think the variances are equal in the two groups.

Thank you for this comment. In the revised manuscript, we’ve corrected that the choice of the test was based on observed unequal variances within gender and BMI categories.

Results

10. The p-values in Table 1 are unnecessary in my opinion, since testing whether males and females had different characteristics is not of scientific interest here.

Thank you for this suggestion. We’ve deleted all p-values from Table 1 in the revision.

11. You have NA for p-value for race and blank for p-value for weight status. I already suggested removing the p-values in this table, but if you were to include them, you could use Chi-squares and/or Fisher's exact tests for categorical variables.

Done. We’ve deleted p-values in Table 1.

12. In Table 2, you can just show one row for GT3X+ sedentary minutes/day, since the two rows are identical. Also, no need to show the empty VM3 row.

Done.
13. In Table 3, I would suggest listing both Pearson and Spearman correlation coefficients (e.g. show \[.89/.87\] in a particular cell if the Pearson is \(.89\) and the Spearman is \(.87\)).

Thank you for this suggestion. Now, we’ve included both Pearson and Spearman correlations in Table 3 in the revised manuscript.

14. Again in Table 4 we don't need to see the same comparison twice. You could have a single row and call it "GT3X+ - Fitbit" or similar.

Done.

15. In Table 4, for MVPA, I would include all of the pairwise comparisons, including Freedson vs. Troiano, Freedson vs. VM3, and Troiano vs. VM3. Those comparisons are not your main focus, but they useful to report.

Thank you for this suggestion. We’ve included all pairwise comparisons for MVPA estimates in Table 4 in the revision.

16. For the ANOVA results, please specify the exact p-value rather than < 0.01 and < 0.05. Also, I am not sure why the omnibus p-value for SED was < 0.05 but the only pairwise comparison was not significant. Here there are really only two groups, so I don't think a Bonferroni correction would require a lower alpha.

Thank you for this comment. We’ve now specified all of the exact p-values we can possibly retrieve from the tests (both SPSS and SAS did not provide the exact p-values lower than .0001). We also agree with the reviewer’s recommendation on inappropriateness of using Bonferroni adjustment for SED, thus removed it.

Discussion/Conclusions

17. Given the significant overall ANOVA p-value, I think you may need to modify your interpretation of the sedentary results. It appears to me that Fitbit does seem to overestimate sedentary time. At least, there is some evidence to suggest that. It is also possible that you were underpowered to detect a meaningful difference (36 min/day) at \(p < 0.05\).
It was our oversight in the initial submission that in fact the p-value for testing difference in SED is not statistically significant (P = 0.21). As the reviewers suggested, we conducted an additional inferential test, equivalence tests, and we confirmed that the estimate of SED from the Fitbit was significantly equivalent to the SED estimate from the GT3X+ (Please see the results section and Figure 2).

18. First sentence of Conclusions is a bit misleading, because the devices don't provide "equivalent" measures of SED either - just not significant different at the .05 level.

Thank you for this comment. As the other reviewer requested, we performed equivalent testing and presented the result that the estimates of SED were significantly equivalent between two devices (please see the Results section, Table 4, and Figure 2).

19. It would be helpful to directly state that you did not use the low-frequency extension, just to be explicit about that.

Thank you for this suggestion. We now specified that we did not apply the low-frequency extension (page 14, paragraph 1).

Reviewer 2

Comments to Author:

PEER REVIEWER COMMENTS: To view the full report from the academic peer reviewer, please see the attached file. REVIEWER COMMENTS FROM REPORT: In this paper the authors aim to compare the concurrent validity of the Fitbit Flex to the ActiGraph GT3X+ in 67 participants with a wide spectrum of ages (20 - 70 yrs) and BMI's (18.6 - 40) during free-living over a 7 day period. Using three different (and validated) cut-points for the GT3X+ when worn at the waist, minutes spent performing MVPA and sedentary behaviour per day are extracted from the device and compared to outputs from the Fitbit Flex, which was worn at the wrist. The Fitbit estimated significantly higher MVPA compared to the GT3X+ but was comparable with regards to sedentary time. While of some scientific merit (to help advance the public health agenda regarding the objective measurement of physical activity dimensions using wearable devices) this paper has a number of issues that require addressing.

We appreciate the reviewer’s thoughtful review. We have carefully read all of the reviewer’s comments and have responded to each recommendation. We believe our manuscript is stronger as a result of the modifications outlined below. Please note that the authors’ responses are provided in blue font color throughout this document.
REQUESTED REVISIONS:

Major points

1. As the devices are worn at different anatomical locations and utilise different on-board processing steps what explains the potential differences between the two devices? Is this study a fair comparison? I understand consumer-based devices are more of a black box with proprietary algorithms etc. but is it possible to provide the dynamic range of the Fitbit Flex and sampling frequency so the reader can make a more informed comparison between the devices?

Thank you for this comment. As the reviewer mentioned, the resolution of data from the Fitbit in generic form is lower than that from the GT3X+ (e.g., minute-by-minute vs. hour-by-hour). In order to address this issue, we processed the data from the Fitbit by downloading the data via Fitabase, which is a research application of the Fitbit devices and allows exporting the data at higher resolution (at 60-sec epochs) than Fitbit’s generic data form (hourly).

We’ve explained procedures of Fitbit data processing more thoroughly in the revised manuscript (please see ‘Data reduction section’, (page 8, line 19 - 22).

In terms of the reviewer’s comment about the location of device placement, please see our response to the comment below.

2. Why did the researchers choose the non-dominant wrist to wear the Fitbit Flex device? Can you please provide rationale or support for this decision? Similar to my point above, dominance is not standardised between devices as the Actigraph is worn on the dominant hip. It seems like this comparison is between apples and oranges.

Thank you for this comment. We chose the non-dominant wrist for the Fitbit Flex because it is default/recommended placement by the manufacturer.

For the ActiGraph GT3X+, we chose the dominant hip as a placement site for the GT3X+ because it has been suggested as a standard placement site of the GT3X accelerometer for adults. Between the wrist and hip location, previous studies have consistently reported that the hip-worn GT3X+ accelerometers provide more accurate classification of PA intensity (also PAEE, step counts) (Hildebrand et al. 2014, Ellis et al. 2014, Stec et al., 2012, Tudor-Locke et al., 2014) and superior reliability (Ozemek et al, 2014) than the wrist-worn accelerometer in adults.

In the revised manuscript, we’ve now added sentences with relevant references to better explain our decision on the placement of devices (Please see page 7, line 21 – 23)
3. Take a closer look at your references throughout. For example, in the data reduction section of the manuscript you present reference years rather than number, and in some instances both.

All references are cited consistently in the revised manuscript.

4. The fact you see no differences in mean bias between devices is merely an artefact of under and over estimations cancelling each other out. To truly understand the differences between the two devices the authors should consider calculating mean absolute error (Please see example: [http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171720](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171720)). You could also consider adding equivalence testing also reported in this paper.

Thank you for this suggestion. As the reviewer suggested, we calculated the mean absolute percent errors and performed equivalence tests, and included results from these additional analyses in the revised manuscript (please see the Results section, Table 4, and Figure 2).

5. As neither of these devices are considered a representative criterion (gold-standard) isn't the discussion regarding under or over predicting somewhat arbitrary? For example, without a criterion measurement, if the Fitbit Flex over predicts MVPA compared to the GT3X+ this could also be that the GT3X+ under-predicts relative to the Fitbit Flex with no indication of which device is more accurate. From a public health perspective, what is an acceptable level of agreement between different devices if we are to make recommendations regarding optimal levels of MVPA for various health outcomes? A difference of 60 - 79 min/day translates to a 420 - 553 min/week differences that would have significant implications for observing whether someone is meeting the MVPA guidelines of 150 min/week. The authors should comment on these issues.

Thanks for the comments. We totally agree with you that there was no gold-standard method used in the present study. So, as the title of this paper indicates, we examined the “concurrent validity” (rather than “criterion validity”) of the Fitbit against the ActiGraph. Although neither of the two devices was a true gold standard method, we used the ActiGraph as the reference method in this research where the goal was evaluating the concurrent validity of the Fitbit, given the more established validity of the ActiGraph; numerous studies have been performed to validate the ActiGraph method against true gold standard methods (Van Remoortel et al., 2012; Welk et al., 2000; Rothney et al., 2010; Kozev et al., 2010; Chomistek et al., 2017). However, evidence on the validity of the Fitbit is currently scarce, so evaluating the concurrent validity of the Fitbit (albeit not criterion validity) against an established research-grade monitor (e.g., ActiGraph) is an important methodological priority. Nevertheless, we acknowledge that we didn’t indicate the lack of a true gold standard method as one of the limitations of the study. So, we have now added
it as one of the limitations as follows: “Another limitation was that there was no true gold-standard method used to evaluate the validity of the Fitbit, thus no inference can be made about the criterion validity of the Fitbit” (page 16, lines 13 – 14).

As far as we know, there is no such accepted level of agreement between different devices. We have specifically indicated the issue with large differences observed for MVPA comparisons in the Discussion section, so that potential Fitbit users are aware of the large measurement error in assessing MVPA time.

6. Given how quickly the physical activity monitoring field is progressing with technological developments and analytical improvements consumer-based wearable technology can become obsolete very quickly. The Flex has already been replaced by the next generation Flex 2. Therefore are the findings from this study now irrelevant given researchers will be attracted to the latest technology or generations of a wearable device. What is the specification differences between the Flex original and Flex 2 and are your findings translatable? I'm afraid we do not currently know this, which drastically limits the generalizability of your findings. This is a major limitation of this study.

In a separate study, we tested the inter-monitor reliability between Fitbit Flex 1 and 2. There was high inter-monitor reliability as evidenced by a high intra-correlation coefficient (ICC) value of 0.91 (data not presented in detail herein).

Fitbit, inc. released an updated version of the original Fitbit Flex, the Fitbit Flex 2, in 2016. The Fitbit Flex 2 and original Fitbit provide PA estimates using a tri-axial accelerometer. There is no indication that the accelerometer specifications (e.g. sampling rate) differ between the Fitbit Flex 2 and Fitbit Flex other than the size of the device encasing the accelerometer and other technologies (i.e. Bluetooth transceiver, vibration motor, and lithium-ion battery) included in both devices. Similarly, on-board memory for both devices is seven days of minute-by-minutes activity data and 30 days of daily totals. The Fitbit Flex 2 is water-resistant up to 50 meters versus 10 meters with the Fitbit Flex. Fitbit Flex 2 includes a slimmer wrist-band and device profile than the Fitbit Flex and additional user feedback features due to changes in the Fitbit app. Though Fitbit discontinued the Fitbit Flex, the device still functions with the Fitbit mobile and desktop applications and is still utilized in published PA monitoring research (Fitbit, Inc. (2018). Fitbit Flex 2 User Manual version 1.2. https://help.fitbit.com/?p=flex_2, Fitbit, Inc. (2018). Fitbit Flex Product Manual. https://help.fitbit.com/?p=flex)
ADDITIONAL REQUESTS/SUGGESTIONS:

Minor points

Abstract:

7. Line 14: You've abbreviated sedentary behavior (SED) but not PA, please correct.

Thank you for pointing this out. We have defined the abbreviation in the abstract.

8. Line 26: Should be written as 'time spent performing sedentary (SED) and moderate-to-
vigorous PA (MVPA) was estimated…'

Thank you for this suggestion. The sentence now reads as recommended (page 2, lines 9 – 10)

9. Line 34: Pearson correlations do not evaluate agreement, merely associations. Please revise this sentence.

We revised the sentence to indicate association as the outcome of the Pearson correlation (page 2, line 20).

10. Line 41: I'm not sure why you are presenting a range for MVPA (MD = -59 - 77 min/day) without specifying so. Also you state MVPA is significantly higher yet the values are proceeded with a minus symbol. Please also state the P value for significance.

We’ve corrected the negative values, specified the values in range are from the estimates using three different GT3X+ cut-points, and included P-values in the revised manuscript.

Background:

11. Line 9: I think it would be useful to specify the health outcomes, please provide examples for the reader.

Thank you for this suggestion. We have specified a few of the leading health concerns related to PA epidemiology (page 4, line 3).
12. Line 21: What do you mean by feasibility in this context? Please explain, are you referring to wearability, cost etc?

We have changed the word ‘feasibility’ to ‘reliability’ in the revised manuscript (page 4, line 9).

13. Line 29 -30: consistency with abbreviations at the first time of use; IC (indirect calorimetry?), MET (metabolic equivalents)

Thank you for these recommendations. IC and MET have been defined upon first introduction.

14. Line 38: do you mean minutes of 'moderate to' vigorous PA which is the most common and clinically relevant category captured by wearable devices.

Thank you for the comment. The sentence has been revised to read “…several activity count cut-points have been developed to estimate the amount of time spent in sedentary behavior (SED) as well as different intensities of PA, namely moderate-to-vigorous intensity PA (MVPA).” (page 3, lines 15 – 17).

15. Line 46: insert 'intensity' after PA. Given the preceding sentence you are discussing the intensities of PA here specifically, not for example PAL or PAEE. Clarity around the terminology you use is very important and seemingly jumps around throughout the manuscript. The end of this sentence doesn't make sense. What do you mean by 'duration of activities used to calibrate the equation'. What equation are you referring to here (do you mean cut-point) and what is the calibration step? Perhaps you should also discuss that these cut-points are not validated for specific population groups (over-weight obese men and women or children) for example: https://www.jsams.org/article/S1440-2440(11)00475-0/pdf https://symbiosisonlinepublishing.com/exercise-sports-orthopedics/exercise-sports-orthopedics02.pdf

Thank you for the comment. The sentence has been updated to clarify that PA intensity classification is the intended outcome. The word duration was removed and the word cut-points replaced “equation” (page 4, line 22). We are not referring to a specific calibration step here. Rather, the sentence refers to the inconsistent methodologies used to establish cut-points for classifying PA intensity from accelerometer counts. The intention was to be succinct with this reference. However, we are happy to elaborate further should the reviewers request it.
16. Pg 3. Line 12. Perhaps define the popularity of these consumer-based wearable devices. For context, how much of the general population own one of these devices? Moreover, what do you mean by acceptability? Do you mean that these devices are commonly worn in everyday life? Providing evidence and a reference for my first question would help support this assertion.

Thank you for this comment. We’ve deleted the word ‘acceptability’ in the revision. Also, we added a sentence “The Fitbit has achieved the largest market share among numerous wearable activity trackers available on the market” with relevant references in the revision (page 5, lines 6 – 7).

Further, we provide some evidence for Fitbit’s popularity here. Two of Fitbit, Inc.’s recent yearly financial statement press releases convey the scale of which Americans purchase and use (‘active users’) Fitbit products. As of 2017, there were 50.2 million ‘active [Fitbit] users’ who make-up what Fitbit claims is the largest social fitness network. Information about previous earning and specific sales of models like the Fitbit Flex are less accessible because the company was privately owned until 2015. Right before its public filing, Fitbit had experienced substantial year-over-year growth in units sold and customer acquisition (Dolan, 2015).


17. Pg 3. Line 17. Models? It is advisable to be more specific when referring to wearable devices used in each study rather than simply stating the manufacturer name. What Fitbit models are you referring to exactly; i.e. Fitbit charge 1,2,3, Versa, Ionic?? Please define throughout the manuscript. It is also pertinent to address whether any of these utilise multi-sensor signals (i.e. HR) as this would explain an improved estimate of EE.
Thank you for the comment. We clarified specific Fitbit models throughout the manuscript, including specifying the use of the Fitbit Flex throughout our study. (e.g. page 6, line 1).

18. Pg 3. Line 29. What sort of activities does the Fitbit under and over-predict for specifically (i.e. cycling, household chores etc). This paper demonstrates that the Fitbit Charge HR over-predicts walking and loaded walking (which presumably would be the mainstay of EE in free-living for the general population) and under-predicts for cycling: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171720

As requested, we listed some specific activities that hip-worn and/or wrist-worn Fitbit models over- or underestimate PA. (page 5, paragraph 1, continued on page 6).

Methods

19. Pg 5. Line 38. Can you please provide more information pertaining to the definition of non-wear time? It is frustrating as a reader to be directed to another source.

The description of the non-wear time validation criteria by Choi et al. can be described in short as below.

“1-minute time intervals with consecutive zero counts for a time window of at least 90-minutes window 1), allowing intervals with non-zero counts lasting up to 2-minutes (allowance interval) if no counts are detected during both the 30-minutes upstream and downstream from that interval (window 2); any nonzero counts except the allowed short interval are considered as wear time. Derived Epoch Length is 60-second.”

Although we agree with the reviewer’s comment, given the length of the description and the space limit of the manuscript, we decided not to include the detailed description of the non-wear time validation criteria by Choi et al. in the revised manuscript. However, we would be more than welcome to include it in the revision if both reviewers and/or the editor request it.

20. Pg 6. Line 11. Did you check the normality of your data before proceeding with a parametric analysis (Pearson correlation coefficients)?

Thank you for this comment. We performed normality tests (i.e., Kolmogorov-Smirnov and Shapiro-Wilk tests) and confirmed that the majority of variables met the normality assumption and the rest of variables were near normal distributions. That said, as the other reviewer also suggested to calculate Spearman’s correlations, we’ve now included both Pearson’s and Spearman’s correlations in the revised manuscript (please see Table 3).
21. Pg 6. Line 29. Re-write as significant overall ANOVA 'effects' were followed by…

Done.

22. Pg 6. Line 34. BA also identifies random bias (LoA) as well as systematic. Consider also explaining that this approach is identifying the agreement between devices.

Thank you for this suggestion. We’ve now explained it with revised BA plots in the revised manuscript (Please see ‘Statistical analysis’ and ‘Results’ sections and Figure 1).

Results

23. Pg 6. Line 43. Consider changing subjects to participants throughout the manuscript: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1115535/

Thank you for the recommendation. The word “subjects” was replaced with the word “participants” throughout the manuscript as requested (e.g. page 2, line 7).

24. Pg 6. Line 53. I would not say that 14.9 hour/day demonstrates high compliance when you've asked participants to wear the devices during all waking and sleep hours (Pg 4. Line 20). Is it that these 14.9 hours represents those that you've removed as 'sleep'? Perhaps you should make this clearer in this sentence.

We have reworded the sentence accordingly by removing the words ‘high compliance’ and clearly specified that the reported wear time was calculated after removing sleep time in the revised manuscript (page 9, paragraph 1).

Discussion

25. Pg 7. Line 32: change to 'suggesting that the Fitbit Flex may overestimate MVPA in active individuals compared to the GT3X+'.

Thank you for this suggestion. Now, we’ve changed the sentence as the reviewer suggested in the revision (please see page 12, lines 18 – 19).
26. Pg 8. Paragraph from Line 8. I admire you trying to explain your results relative to the wider literature and identifying that sex might play a part (and your sample is mostly made up of females). However, a clearer description of how this is relevant needs to be provided. Is it that males are more active than females? Is this important because the range of values in a comparison can strengthen the magnitude of associations (i.e. larger range in males, 5 - 500 min/day vs. 3 - 200 min/day in females may explain the differences across studies)?

Thank you for this comment. We agree with the reviewer’s comment that our attempt to interpret the observed lower correlations in MVPA estimates in relation to sex differences in MVPA levels is deemed irrelevant, thus we deleted the paragraph in the revised manuscript.

27. Author contributions: please define.

Done.