Author's response to reviews

Title: Which activity monitor to use? Validity, reproducibility and user friendliness of three activity monitors

Authors:

Brenda AJ Berendsen (brenda.berendsen@maastrichtuniversity.nl)
Marike RC Hendriks (marike.hendriks@maastrichtuniversity.nl)
Kenneth Meijer (kenneth.meijer@maastrichtuniversity.nl)
Guy Plasqui (g.plasqui@maastrichtuniversity.nl)
Nicolaas C Schaper (n.schaper@mumc.nl)
Hans HCM Savelberg (hans.savelberg@maastrichtuniversity.nl)

Version: 2 Date: 16 May 2014

Author's response to reviews: see over
Dear editor,

Thank you for taking our manuscript ‘Which activity monitor to use? Validity, reproducibility and user friendliness of three activity monitors’ (MS: 2889009671204315) into consideration. We would like to thank the referees for the helpful reviews of our manuscript. Please find the point-by-point revisions below and amendments in the revised manuscript in red font. We hope that our revisions meet your expectations.

Sincerely on behalf of all authors,

Brenda Berendsen MSc
Dept. of Human Movement Science, Maastricht University
P.O. box 616
6200 MD Maastricht, the Netherlands
(+31)433884219/ fax (+31)433670972
E-mail: brenda.berendsen@maastrichtuniversity.nl

Referee 1

1. I had some concerns regarding the description of the methods and the methods used.
It was unclear why the authors chose to evaluate the ActiGraph 100 cpm cut-point in the free-living and not in the laboratory. It would be much clearer if the same outcomes were used in both studies. I recommend adding the AG100 to the lab rather than removing it from the free-living, as it’s well established the inclinometer function is poor and more people use the 100 count cut-point.

We are aware of the limitations of the inclinometer function and our data confirmed the findings of earlier studies. We had decided post hoc that analysis of the 100 count cut-point might be of added value and therefore applied it to the data where possible.

Output by ActiLife (ActiGraph software) is extractable on minute level, while the activity periods of our protocol generally lasted less than one minute. So we were not able to obtain 100 counts cut-point results for the exact activity periods in the protocol. Although it is possible to retrieve the time spent in inactivity or activity for the entire protocol, we have no information of the activities that occurred in between the protocol activities because we did not control the transition periods. Therefore, we cannot compare the 100 counts with the actual events.

For the analyses of the free living data we could apply the 100 counts cut-point because we used total time per day spent in inactivity or activity.

To increase clarity, we added to the manuscript: ‘In the laboratory measurement, we assessed the intensity measure and the inclinometer function to discriminate postures of the three activity monitors.’ and ‘We evaluated four methods in the free living experiment (AP, AG inclinometer, AG counts and CAM).’ (line 139-140 and 164-165)
Five subjects for the laboratory seems like a very small sample. Are power calculations or other rationale available for this number?

We understand your question regarding our sample size. We analyzed the protocol execution of five participants twice, each containing 330 data samples (see line 156-158). The protocol validation analyses were thus based on 3300 samples. We haven’t calculated sample size. However, despite the small sample size in some of our analyses, we did find significant agreement between the activity monitors and the comparators. To account for the sample size, we added confidence intervals of the agreement statistics (ICC’s) to the results section (line 238-239 and 252-262).

More information on how the data were derived. How was the ActiGraph initialized? Should clearly define the parameters from each monitor in the text, before the statistical analysis.

The ActiGraph was initialized to measure 1 second epochs, as shown in table 1, and measured continuously during the free living measurement. Wake time data was extracted based on the diary prior to analyses. The outcome parameters were described in the analysis section, but I moved it forward to increase clarity (line 126-136).

P9, line 165: How were the cronbach’s alphas determined?

We used the Cronbach’s alpha to assess the reliability of the three scales of our self-constructed user friendliness questionnaire. The alpha was determined with the three questions constituting each scale. Reliabilities of >0.6 per scale were deemed acceptable, so all scales could be used as shown in table 2. These alphas are included in the methods section rather than in the results section, because they were determined prior to the actual analyses answering the research question.

2. It appears the data are sound, however I found the results and figures very confusing.

It seems odd that subject 4 had 98.9% of sitting time classified as non-wear while all other subjects had <5% of sitting time classified as non-wear. Do you have any insight to this?

Thank you for pointing out this mistake. During drawing up the manuscript, we mixed up sitting time percentage with non-wear time percentage. After discovering this mistake, we checked all other analyses, and this mistake was only present in table 4.

a. Figures 2: Is the difference the difference in that exact second between the two protocols?

This question pointed out an issue we hadn’t considered. The analysis you are referring to, is the reproducibility of the intensity measure during walking on the treadmill. Comparison between the exact seconds of the two protocols is not meaningful. Therefore, we re-analysed the reproducibility by comparing mean value of intensity for each walking speed between the two measurements instead of correlating each value/sec (line 238-240).

b. Figures 3 and 4: It appears that the standing/active time are simply the inverse of one another (figure 3a)? (e.g., at ~2000 min active is +800 and standing is ~800min), and a number of other look similar., including on the CAM monitor (Figure 2b) Could this be due to a miscoding or misinterpretation of the diary?
Has the diary been validated?

Based on the figures, it might look like standing/active time are simply the inverse of one another. After looking at the data in more detail we can conclude that this is not the case. To illustrate, the example you gave concerned data points of two different participants. In one participant standing time was overestimated by 836 minutes, and in the other participant active time was underestimated by 848 minutes. We recognize the disadvantage of using diary data and this is also mentioned as a limitation in the discussion section. Ethical issues refrained us from using direct observation. We strived to reach the most valid and reliable diary data, by asking the participants to distribute each 15 minutes into four easy to understand categories (i.e. sitting, walking, standing and other activities). We believed reporting each 15 minutes was feasible to recall events and estimate their duration within those 15 minutes. When activities occurred in only one category for longer than 15 minutes, participants were allowed to report them after the first transition. We were not able to control reporting accuracy, but we observed in the diary data that people made an effort to report in detail (i.e. in minutes precise).

I expanded the information regarding diary reporting in the methods section: “When activities occurred in only one category for longer than 15 minutes, participants were allowed to report them after the first transition.” (Line 169-171).

c. Figures 3 and 4: Is the same line fit for all three classifications (i.e., how is the mean difference 0 for all of them?) A more useful approach might be to look at separate bland-altmans by intensity category.

In response to this question we made separate Bland Altman plots for all provided categories for each activity monitor (figures 3 - 6). Now, the systematic differences within sitting/lying and standing are more evident for the reader.

d. Figures 3 and 4: It’s also not clear why the averages are so different along the x-axis for the different tools.

The x-axis values of the Bland Altman plots are different between the four methods. This is due to the different methods of measurement and comparison. The AP was worn during sleeping, so sleeping data is also included in the analyses and is recognizable in the high mean sitting/lying values of more than 3000 minutes. In contrast, CAM and AG were worn during waking hours only, explaining the lower mean values of sitting/lying time. Between the two AG classification methods (inclinometer and counts cut-point) the difference is caused by the respective categories that were used. The two classification methods use a different approach and therefore time in the categories inactive/active and sitting/lying/upright are not comparable or interrelated.

The authors clearly show the non-wear function is invalid in the AG. What happens if the non-wear time was classified as lying time? Perhaps mentioning this in the detailed examination of AG would be helpful.

We added some details in the results section to expand on the non-wear classification: “Overall 98.1% of classified non-wear occurred during lying, 1.7% occurred during sitting and 0.2% occurred during upright time.” (line 232-233)

The limitations should be that a criterion measure was not used during free-living measurements (direct observation), not that an objective measure was not used. The monitors are objective.
I agree and replaced ‘objective measure’ with ‘direct observation’ (line 351).

4. **Is the writing acceptable?**
   I think the writing could be improved. The methods and results were difficult to follow and there were some inconsistencies with terminology (See below).
   
   **Minor Discretionary Revisions**
   P1, line 53: add “prolonged” before sitting
   P1, line 53: remove “very” before variable.
   P2, line 76: change “is” to “was”
   P2, line 83: Should that say AP3? Otherwise, I’m sure the distinction the authors are trying to make.
   P11, line 222: suggest changing “in AG in detail” to “by AG in detail”.

   **Table 4:** comma’s should be periods.
   P7, line 128: Suggest revising to read “To determine test-retest reproducibility, a 19.5 minute protocol was completed twice by all participants…”
   P14, line 291: The authors suggest placement by elastic belt is not optimal, what is optimal?

   The authors refer to the inactive time as static in text (p12, lines 240-249), this should be consistent with the figures

   *Thank you for your detailed remarks; based on these, we adapted the writing to improve readability.*
Referee 2

Major Compulsory Revisions
(1) The primary concern this reviewer has is with the small sample size. As is, it is very difficult to make any strong conclusions with this sample. This is especially the case with the free-living data where comparisons were at the day level. Also, in this case the sparse missing data appears to have really impacted the data.

We understand your question regarding our sample size. We analyzed the protocol execution of five participants twice, each containing 330 data samples (see line 156-158). The protocol validation analyses were thus based on 3300 samples. Despite the small sample size in some of our analyses, we did find significant agreement between the activity monitors and the comparators. To account for the sample size, we added confidence intervals of the agreement statistics (ICC’s) to the results section (line 238-239 and 252-262).

After re-reading the text regarding missing data, we felt that it might not be clear that information with ‘n=..’ did not always mean that this data was missing. We clarified in the text how many data was available from the free-living measurements for each activity monitor: “All activity monitors were worn for an average of four days, ranging from two to six days per participant (eight participants). On average per person, three days of the AP-data were usable (seven participants). An average of four days of the AG could be used for the 100 counts cut off point (eight participants) and an average of two days could be used for the inclinometer analyses (seven participants). An average of three days of the CAM measurements could be used in analyses (eight participants).” (Line 244-249)

(2) The reliance of diary data as a gold standard is problematic for free-living. This is especially concerning with 15min windows, although this reviewer understands that finer windows is likely not feasible. A much stronger approach would have been some type of direct observation. It is very difficult to have any confidence in the free-living results given the number of transitions that are likely to occur in any given 15min period and the reliance on reporting. Furthermore, there was not reporting of quality assurance to ensure that subjects were reporting after the 15min period and not all at the end of the day.

We recognize the disadvantage of using diary data and this is also mentioned as a limitation in the discussion section. Ethical issues refrained us from using direct observation. We strived to reach the most valid and reliable diary data, by asking the participants to distribute each 15 minutes into four easy to understand categories (i.e. sitting, walking, standing and other activities). We believed reporting each 15 minutes was feasible to recall events and estimate their duration within those 15 minutes. When activities occurred in only one category for longer than 15 minutes, participants were allowed to report them after the first transition. We were not able to control reporting accuracy, but we observed in the diary data that people made an effort to report in detail (i.e. in minutes precise).

I expanded the information regarding diary reporting in the methods section: “When activities occurred in only one category for longer than 15 minutes, participants were allowed to report them after the subsequent transition.” (Line 169-171).

(3) The user-friendliness data may be the most interesting aspect of the manuscript. This is an important consideration that most investigations overlook.
That said, this reviewer finds it difficult to understand the meaningfulness of these data from subjects that have used the sensors for only 40 minutes (laboratory sample) or 3 days (free-living) when most subjects in research settings will be asked to wear the sensor for at least 7 days.

The activity monitors were worn for at least 3 days in free living. Of nine participants, five wore the devices for six days and four wore the devices for three days. Two participants did not wear one or more activity monitors. Although data was excluded due to measurement errors of the devices, user friendliness data was still based on all days the device was worn.

Even in our data, including short measurements, differences between the user friendliness of the activity monitors were revealed. Although we cannot extrapolate these findings to longer measurements, such as seven days, we wished to emphasize the relevance of the user friendliness of the devices, especially in free living measurements. We would also suggest evaluation of user friendliness in studies relying on accelerometry data, to get insight into probability of non-wear and/or adaptations in behavior.

With regards to your question about the meaningfulness of the user friendliness data of the 40 minute measurements, analyses with only free living data yielded the same significant differences between the activity monitors as the analyses reported in the manuscript.

(4) What measures were undertaken to assure that the data were aligned across devices at the 1sec epoch level? Were clocks synced in any way prior to initialization? Related, it was noted that the only the ‘middle’ part of the data was retained for analysis. Please define what this means. Was is only the interquartile range retained?

The devices were synchronized with the protocol and each other by means of jumping at the start of the measurement (CAM) or their internal clocks (AG en AP). To account for influence of possible small misalignment, we retained the middle part of the data for analyses. This means that we excluded the first and last part of the events, and thereby ensured that all included seconds were spend in the category that was intended.

To clarify, I’ve added some information to the methods section: “The devices were synchronized with the protocol and each other by means of jumping at the start of the measurement (CAM) or their internal clocks (AG en AP).” (line 152-153) And adapted a sentence to: “To prevent inclusion of transition phases, the first and last ten seconds of the data of each condition were excluded; if the condition duration was 30 seconds or less, the first and last five seconds were excluded.” (line 153-156).

(5) There is no justification for the use of the low frequency extension for the AG. How might this have affected intensity counts across the intensity spectrum?

The low frequency extension (LFE) was developed to make the AG GT3x more sensitive to movements of lower intensity, and thereby attempt to increase comparability with the older AG versions. Studies assessing the difference between the AG GT3x with and without the LFE have revealed that applying the LFE leads to less classification in sedentary time and more physical activity time. In addition, using the LFE increased comparability with the uni-axial AG. Therefore, we have decided to assess the AG with the LFE and found a significant and good correlation between the AG counts and walking speed.
(6) For test-retest analyses, it was stated that data were non-normal so a spearman rho was used. Was any attempt made to normalize with log, square root, or other type of transformation. The superior statistical test here would be the intraclass correlation coefficient, which is relatively robust to non-normal data and a better overall assessment of test-retest. Also, a simple mean comparison should be included to ensure absolute comparison of value in addition to relative comparisons that you would get with a correlation coefficient.

Like you advised, we used ICC instead of the Spearman rho. In addition, we also adapted the analyses of the agreement between diary and activity monitors in free living to use ICC. We acknowledge that a significant correlation does not ensure absolute agreement between the two measurements. For this reason, we plotted the absolute differences in Bland Altman plots (figure 2 – 6).

(7) Line 280-281 concludes that the AP was user-friendly. Notwithstanding the limitation noted above in (3), the data presented does not suggest this is so. Skin irritation while or after wearing the AP was indeed reported by participants in both the protocol and the free living measurements. Nevertheless, the AP showed highest scores in three of the five user friendliness categories, indicating its high user friendliness. We added “(despite reported skin irritation in four participants)” in the conclusion of the AP (line 295).

(8) Related to (8), descriptive data are not presented for the user-friendliness data. We can only rely upon statistical reporting. It would be good to see the absolute levels of these questions. Were they all rated high despite statistical difference between AG and AP/CAM?

We recognize your request for descriptive data and have added boxplots of the user friendliness scores as figure 7.