Author's response to reviews

Title: Older adults' reporting of specific sedentary behaviors: validity and reliability

Authors:

Jelle Van Cauwenberg (jvcauwen@vub.ac.be)
Veerle Van Holle (veerle.vanholle@ugent.be)
Ilse De Bourdeaudhuij (ilse.debourdeaudhuij@ugent.be)
Neville Owen (Neville.Owen@bakeridi.edu.au)
Benedicte Deforche (benedicte.deforche@vub.ac.be)

Version: 2
Date: 8 July 2014

Author's response to reviews: see over
Dear Editor and Reviewers

The authors wish to thank the editorial team and referees for their positive feedback, detailed suggestions for improvement, and the opportunity to revise our manuscript.

We have made the revisions according to the comments and have highlighted these modifications in the manuscript. A point-by-point description of the changes is provided below.

Kind regards
On behalf of all co-authors

Jelle Van Cauwenberg

**REVIEWER 1**

**Major compulsory revisions**

1. In the Methods, several references are made to the development of the measure as assessing the range of sedentary behaviours relevant to older adults. How was this established? What formative work was conducted to determine the domains that should be assessed, and the optimal order in which to ask the items to limit duplicate reporting across domains/types?

The authors agree that more information concerning the development of the new questionnaire should be provided. Basically, the selection of specific sedentary behaviors was based on existing questionnaires, including that of Gardiner, for which initial qualitative work was carried out and on recommendations formulated in other previous research articles. To clarify this, we added the following to the Methods section:

To include a wide range of sedentary behaviors, we combined the sedentary behaviors included in previous questionnaires [15, 25, 26], and complemented these with additional sedentary behaviors relevant for older adults. More specifically, we subdivided questions targeting sitting in a car into driving a car, being a car passenger and using public transport. We added one question targeting usual time spent sitting while eating in the last seven days (in minutes/day) as was suggested by Gardiner et al. [15]. Furthermore, we added questions targeting sitting while doing household chores (e.g., ironing, preparing a meal).

To assess older adults’ understanding and relevance of the specific sedentary behaviors we conducted a pilot-study. To prevent duplicate recording, we trained the researchers responsible for data collection to pay particular attention to this issue. This was added to the Methods section as follows:

The new questionnaire was pilot-tested in a convenience sample (n=4) of community-dwelling Flemish older adults to assess older adults’ understanding and completeness of the different items. Researchers involved in data collection were explicitly trained to ensure participants...
reported sedentary behaviors in which they engaged during the last seven days and did not duplicate their reported sedentary times across different sedentary items.

2. Is there any evidence to support your decision to use ‘last seven days’ recall rather than ‘usual week’ (i.e., what evidence did you refer to or collect in formative research for deciding that this format results in easier recall)?

We used ‘last seven days’ because we considered this to be easier to recall than ‘a usual week’. Furthermore, we wanted to use a similar format than the IPAQ, for which the most frequently used time frame is the ‘last seven days’ (Helmerhorst et al., IJBNPA, 2012). The ‘last seven days’-version was also preferred by most study sites in a 12-country validity and reliability study of the IPAQ. We added the following to the Methods section:

Furthermore, the ‘last seven days’ timeframe is the most frequently used time frame of the IPAQ [22] and it was preferred over the ‘usual week’ by most study sites in a 12-country validity and reliability study of the IPAQ [23].

However, we do agree that the chosen time frame might have influenced our results. Especially, it has been raised that older adults might consider a usual rather than the last week when reporting their engagement in physical activity although they were asked to consider only the last week (Heesch et al., IJBNPA, 2010). To acknowledge this, the following has been added to the Methods and Discussion sections:

Furthermore, we used ‘the last seven days’ as the time frame to report sedentary behaviors. Among adults, similar reliability and validity results for self-reported total sitting time have been observed for ‘the last seven days’ and ‘the usual week’ time frame [23]. However, it has been argued that older adults might consider a usual rather than the last week when reporting their engagement in physical activity behaviors although they were asked to consider only the last week [42]. Therefore, in the current study, researchers responsible for data collection were explicitly trained to ensure that participants’ self-reports reflected engagement in sedentary behaviors during the last seven days. To our knowledge, no studies have investigated the influence of administration mode or time frame on the psychometrics of a sedentary behavior questionnaire among older adults. More research is necessary to determine the optimal mode of administration and time frame.

3. Were the items pilot tested and refined prior to the validity and reliability testing reported here? A brief description of such processes would be useful for highlighting broader issues in assessment of sedentary activity for older adults and how they relate to the strengths of your measure.

The items were indeed pilot tested prior to the actual validity and reliability study. Based on this pilot test, no major modifications to the questionnaire were deemed necessary. We have added the following to the Methods section:

The new questionnaire was pilot-tested in a convenience sample (n=4) of community-dwelling Flemish older adults to assess older adults’ understanding and completeness of the different items.
4. Given the uniqueness of measuring specific types of sedentary activity in this study, it would be of interest to include the descriptive data for the amount of sedentary time reported for each type.

We have added the descriptive data for the specific sedentary behaviors to Table 1.

5. Sub-group analysis for the validity and reliability findings would be a worthwhile addition, and make the findings here more comparable with findings from other studies of validity and reliability of self-reported measures of sedentary behavior. It would also offer some evidence for the assertions made in the Discussion about possible effects of education (for example) on valid reporting. Were there variations in validity for those with higher and lower levels of educational attainment, by gender, etc.?

The authors agree that subgroup analyses would be a worthwhile addition to the manuscript and we thank the reviewer for this useful suggestion. We have now included subgroup analyses for 65- to 74-year-olds vs. +75-year-olds, men vs. women and non-tertiary vs. tertiary educated participants. The results of the validity analyses in the total sample and the subgroups are provided in a new table (Table 2). The other additions per section are provided below:

Introduction section:
Previous studies have reported differences between men and women in validity findings for a sedentary behavior questionnaire, in US overweight adults [17] and in European adolescents [18]. Next to gender, other demographic factors, such as age and educational level, may also influence validity. Except for one study testing the reliability and validity of a question targeting overall sitting time among residents of low versus high socio-economic neighborhoods of Hong Kong [19], differences in validity results of a sedentary behavior questionnaire between demographic subgroups of the older population remain unexplored. Additionally, we examined whether the criterion validity of the questionnaire differed according to age, gender and educational level.

Methods section:
Age was dichotomized as 65-74 years and 75+ years old. Educational level was assessed using a 6-point scale ranging from having completed primary to university education. This was dichotomized as non-tertiary and tertiary (including college and university) education.

To examine the criterion validity in different demographic subgroups, this procedure was repeated for subgroups based on age, gender and educational level.

Results section:
Similar patterns were observed in the subgroups based on age, gender and education; the average of self-reported and accelerometer-derived measurements was significantly positively related to the difference between these two measurements with underestimation at lower and medium averages and overestimation at higher averages (see Figure 2). However, in the 65- to 74-year-old, male and tertiary educated subgroups the correlation between self-reported and accelerometer-derived sedentary time was substantially stronger than in the older, female and non-tertiary educated subgroups. Furthermore, in the 65- to 74-year-old, male and tertiary educated subgroups, the standard deviations of the residuals (and correspondingly the 95% limits of agreement) were smaller.
Discussion section:
Our validity findings differed between demographic subgroups. We observed stronger correlations and narrower limits of agreement for 65- to 74-year-old, male and tertiary educated participants compared to their counterparts. A first explanation for these differences in validity might be that the younger age group, men and those with tertiary education engage more frequently in sedentary behaviors that are easier to recall and report, such as car driving and computer use, compared to their respective counterparts. Furthermore, better cognitive functioning and capacity to recall and report past sedentary behaviors among 65-74 year old compared to 75+ year old participants might explain the better validity results in the younger age group. As mentioned, our validity results for men (rho= 0.35) were better compared to women (rho= 0.24). However, España-Romero et al. [37] reported similar correlations between self-reported sitting time and sedentary time measured by a combined heart rate and movement sensor among men (rho= 0.17) and women (rho= 0.18) in a sample of British 60- to 65-year-olds. The sample of España-Romero et al. [37] was younger than our sample and included many non-retired participants, which possibly means that these men and women were more likely to engage in similar sedentary behaviors (i.e. occupational sitting) and, hence, similar validity results. Our most substantial difference in correlation was found between non-tertiary (rho= 0.25) and tertiary educated participants (rho= 0.39). This is in line with findings by Sabia et al. [38] on the validity of self-reported physical activity among 60- to 83-year-old British participants. They found lower correlations between self-reported and accelerometer-derived PA among those with lower education or occupational position compared to participants with higher education or higher occupational position. Cerin et al. [19] reported better reliability of a question assessing overall sitting time among Hong Kong older adults living in high compared to low socio-economic neighborhoods. However, their validity results did not differ between high and low socio-economic neighborhoods. More research is needed to further examine demographic differences in validity and to increase the specificity of questionnaires for groups in which lower validity has been observed.

6. Please justify in the Background/Methods why you chose to use accelerometer data as your sole criterion measure for assessing validity.

A justification of the use of accelerometer data has been added to the Methods section, as follows:

The Actigraph GT3X+ accelerometer served as criterion measure of overall sedentary time. Actigraph accelerometers are the most frequently used tools to measure physical activity and sedentary behavior in population-based studies among older adults [27]. These accelerometers register accelerations of the human body; their output (counts/minute) can be used to derive the intensity at which activities were performed. However, this type of accelerometer cannot distinguish between different postures; they cannot distinguish whether registered counts originated from lying, sitting or standing activities [28]. Other types of devices, such as the activPAL, measure thigh inclination from which posture (lying, sitting, or standing) can be inferred [29]. However, their use is less common in population-based studies [9, 27]. Furthermore, Healy et al. [9] have shown that Actigraph accelerometer-derived sedentary time has minimal bias compared to activPAL-derived sedentary time.
7. Please justify why you imposed the exclusion of those limited by their health to walk a couple of 100 metres. It’s not clear why this sub-group would be excluded from a study to measure the validity of sedentary behavior, given they are most likely to be engaged in high levels overall and the attribution of domain-specific sedentary activity to their overall levels of sedentariness could offer insight into potential for decreasing certain types of sedentary behavior. What are the implications of this for the generalizability of your findings to assessment in older adults?

The current validation study was embedded in a larger study on the environmental determinants of physical activity among older adults. Therefore, older adults who were limited by their health to walk a couple of 100 metres were excluded from this study. The authors agree that older adults who suffer from mobility limitations are a potential risk group for high levels of sedentary time. Therefore, we added the following to the limitations component of the Discussion section:

Thirdly, older adults who were limited by their health to walk a couple of 100 meters were excluded from the current study. Therefore, our findings could not be generalized to older adults with such mobility impairment. Since mobility-impaired older adults may be at increased risk for high levels of sedentary time, future research should investigate the measurement properties of sedentary behavior questionnaires in this subgroup.

8. You need to be clear that the validity and reliability reported here is restricted to administration of the questions by interview, and may not be extrapolated to self-completion. Some comment on the likely impact of administration methods on validity and reliability, with reference to other evidence, would be useful for potential users of these measures.

The authors wish to thank the reviewer for this very relevant remark. We agree that our validity and reliability results are only applicable to an interview-based version of our questionnaire. We discussed this referring to previous psychometric studies on questionnaires assessing older adults’ physical activity behavior:

It should be noted that we tested an interview-based version of our sedentary behavior questionnaire and that our results may not be applicable to self-completion of the questionnaire. For a questionnaire assessing older adults’ physical activity, Dinger et al. [40] concluded that their observations of very good test-retest reliability (ICC= 0.91) might have resulted from the use of interviews rather than self-completion. Washburn et al. [41] found better validity results for a telephone-based physical activity questionnaire compared to a self-completion version, but the latter resulted in better test-retest reliability results. Furthermore, we used ‘the last seven days’ as the time frame to report sedentary behaviors. Among adults, similar reliability and validity results for self-reported total sitting time have been observed for ‘the last seven days’ and ‘the usual week’ time frame [23]. However, it has been argued that older adults might consider a usual rather than the last week when reporting their engagement in physical activity behaviors although they were asked to consider only the last week [42]. Therefore, in the current study, researchers responsible for data collection were explicitly trained to ensure that participants’ self-reports reflected engagement in sedentary behaviors during the last seven days. To our knowledge, no studies have investigated the influence of administration mode or time frame on the psychometrics of a sedentary behavior questionnaire among older adults. More research is necessary to determine the optimal mode of administration and time frame.
9. Reference to ‘activPAL’ first appears in the Discussion, with no explanation of what it is. Can you explain this and identify how it differs from accelerometer assessment?

In response to a previous comment (comment 6), an explanation of the activPAL-device and how it differs from Actigraph accelerometers was added to the Methods section:

The Actigraph GT3X+ accelerometer served as criterion measure of overall sedentary time. Actigraph accelerometers are the most frequently used tools to measure physical activity and sedentary behavior in population-based studies among older adults [27]. These accelerometers register accelerations of the human body; their output (counts/minute) can be used to derive the intensity at which activities were performed. However, this type of accelerometer cannot distinguish between different postures; they cannot distinguish whether registered counts originated from lying, sitting or standing activities [28]. Other types of devices, such as the activPAL, measure thigh inclination from which posture (lying, sitting, or standing) can be inferred [29]. However, their use is less common in population-based studies [9, 27]. Furthermore, Healy et al. [9] have shown that Actigraph accelerometer-derived sedentary time has minimal bias compared to activPAL-derived sedentary time.

10. Is there likely to be a practice effect of having previously (and so recently) completing the survey, which may affect your test-retest reliability assessment? Is there evidence for the amount of difference attributable to practice effects alone (and over different test-retest intervals)?

During both test and retest, participants completed an interview-version of the International Physical Activity Questionnaire (IPAQ) prior to the new sedentary behavior questionnaire. Therefore, since the new sedentary behavior questionnaire had a similar format as the IPAQ, the participants already practiced this format during the first assessment. This reduces the chance that there still was a substantial practice effect during the retest. Second, as described in the discussion section, we also performed a validity analysis in the subsample used for the reliability analyses (n= 28). In this subsample, the recall period overlapped with the accelerometer assessment period, but the validity results were very similar to the result in the total sample. This might also be considered as evidence that the participants did not respond more accurately to the questionnaire during the second compared to the first assessment.

11. Why was criterion validity assessed against the initial self-report assessment rather than at the second assessment (when the recall period would have overlapped with the accelerometer assessment period)? Although the protocol cannot be changed, some justification for this (over the alternative possible protocol) is needed.

The authors agree that it would have been better if the self-report assessment overlapped with the accelerometer assessment. As described in the Discussion section, this was the case for the retest assessment of the questionnaire in the subsample in which the reliability analyses were performed (n= 28). These are the results for the validity analyses within this subsample:
Although $b_1$ was smaller than for the validity analyses in the large sample, the Spearman’s rho, standard deviation of the residuals, $D$ and LOA were very similar. Therefore, we chose to use the validity analyses in the large sample since our conclusions would not be altered. This also allowed us to perform the subgroup analyses requested by the reviewer.

We considered presenting the above results in the manuscript, but decided it would complicate things unnecessarily and would add many extra words to what is now already a much-longer manuscript. Of course, we could add these results if deemed appropriate by the reviewer.

**Minor editorial comments**

12. Specify your definition of ‘older adults’ (i.e., >65 years old) in the abstract.

We have added this definition to the abstract.

13. Findings of validity of self-report measures in older populations from other studies are reported in hours/day in your Background. Can you please convert to mins/day for easier comparability between these findings and yours?

We have converted these figures to minutes/day.
14. Some of the written language could be improved, e.g., "with 81.88 minutes/day for a mean average" could be re-written as "BY 81.88 minutes/day RELATIVE TO the mean average of the two measurements". Similar examples are throughout and should be re-worded for clarity.

We have modified these sentences and had a native speaker checking our manuscript on language mistakes.

REVIEWER 2

Major compulsory revisions

1. The validity study: According to figure 1, the instrument seems to be “quite random” in its precision. Was absolute values used for mean and deviation? The instrument is both overestimating by 6-7 h AND underestimating by 6-7 h, (and everything in between). Is the statistics based on the deviations? You conclude that there is an underestimation of sedentary time, yet many measures clearly overestimate compared to accelerometer values. It seems that since the deviations go both ways – this might contribute in giving a better/more positive result than the data really allows for. Please clarify.

Absolute values are indeed used for means and deviations and the instrument indeed under- and overestimates with 6-7 hours for some individuals. However, the regression line shows that for low and medium averages of self-reported and accelerometer-derived sedentary time, self-reported total sitting time generally underestimated the accelerometer-derived measurement. On the other hand, for higher averages, self-reported total sitting time generally overestimated the accelerometer-derived measurement. This is now more explicitly described in the Results section:

For lower and medium averages of self-reported and accelerometer-derived sedentary time, self-reported total sitting time underestimated the accelerometer-derived measurement. For averages higher than 640 minutes/day, self-reported total sitting time overestimated the accelerometer-derived measurement.

Furthermore, we have added the standard deviations of the residuals to Table 2. These standard deviations are used to calculate 95% limits of agreement and indeed illustrate that there is a lot of variability around the general trend (the regression line). The limits of agreement are also presented in Table 2 (this table was added in response to a comment of reviewer 1) and their formula is provided in the legend. We have added the following to the Results section to acknowledge and clarify the meaning of the wide limits of agreement:

Corresponding 95% limits of agreement were wide (-364.16; 200.41 minutes/day), implying strong variability surrounding these general trends.

Our finding that for higher averages of self-reported and accelerometer-derived sedentary time, self-reported total sitting time overestimated the accelerometer-derived measurement is now also more extensively addressed in the Discussion section:
For higher levels of average self-reported and accelerometer-derived sedentary time, self-reported total sitting time overestimated the accelerometer-derived measurement. This might be explained by those with high levels of sitting time also engaging in longer bouts of sitting time for which durations might be more difficult to estimate and more likely to be rounded up. Furthermore, these longer bouts might have been interrupted by non-sedentary activities which are registered by the accelerometers as non-sedentary, but which are included in the sedentary time reported by the participants.

We have nuanced the reporting of our findings in the Discussion section as follows:

Overall, our findings on validity were not ideal; the correlation coefficients did not reach 0.50 (which was defined as good validity for physical activity questionnaires) [34] and the 95% limits of agreement were wide.

2. The conclusions should reflect that this does not seem to be a very useful questionnaire, at least when it comes to validity based on your results. But since there is few in the field, can you comment on if this is a better option using your instrument than developing a new questionnaire if someone should do a new study on sedentary behavior among elderly? Would they be better off trying to develop a new instrument and testing it than using this? What does your wide 95% limit of agreement imply? A discussion on the usefulness/practical implications of this new instrument is lacking. Thereby also including if the aim of the study is accomplished.

We have added the following paragraph describing the practical implications of our findings to the Discussion section:

Overall, our findings on validity were not ideal; the correlation coefficients did not reach 0.50 (which was defined as good validity for physical activity questionnaires) [34] and the 95% limits of agreement were wide. In addition to the reasons described above, another possible explanation for this absence of strong validity is that the target period of our sedentary behavior questionnaire did not overlap with the period the accelerometer was worn. However, in our subsample for the reliability analysis both periods did overlap, but a correlation analyses between their self-reported total sitting time and accelerometer-derived sedentary time did result in a similar correlation ($\rho= 0.32$) and similar width of the 95% limits of agreement. Additionally, although we followed standard procedures for accelerometer initialization and processing, there is not yet a consensus about many of these procedures [27]. Although our findings showed only modest validity, they were no worse than those reported for previous questionnaires targeting older adults’ (specific) sedentary behavior(s) [15, 16]. Moreover, our questionnaire targeted a wide range of specific sedentary behaviors which may have resulted in a lower level of underestimation of total sitting time compared to previous questionnaires [15, 16]. Additionally, our questionnaire’s format is similar to the IPAQ, which might facilitate administration in studies that assess both physical activity and sedentary behaviors. Given the high prevalence of sedentary behaviors among older adults and the associated health risks, researchers should not delay studies on the health risks, prevalence and correlates of sedentary behaviors and could use the new questionnaire to assess older adults’ sedentary behaviors. Objective measures of sedentary time should be preferred, complemented with the questionnaire to get context-specific information. Our questionnaire might be especially useful for the specific sedentary behaviors for which we found acceptable reliability; TV viewing time, computer use, and car driving. In the meantime, more research is needed to develop questionnaires and objective criterion
measures that measure older adults’ engagement in sedentary behaviors more accurately. These validity studies should use different criterion measures (e.g. Actigraph accelerometers and activPALs) and could include log books to examine the validity of specific sedentary behaviors.

In addition, we rephrased the conclusion as follows, with the first sentence being the answer to the research aim:

To conclude, we examined criterion validity and test-retest reliability of a newly-developed sedentary behavior questionnaire, but our findings did not exhibit ideal validity for self-reported total sitting and test-retest reliability for most of the specific sedentary behaviors. However, our findings were comparable to what has been reported by previous studies. Furthermore, our questionnaire tended to result in a lower level of underestimation of sedentary time compared to other questionnaires, possibly explained by the inclusion of additional specific sedentary behaviors. We also observed better validity results for 65- to 74-year-old, male and tertiary educated participants compared to their counterparts. Further research is needed to develop self-report tools and objective criterion measures that accurately measure older adults’ engagement in specific sedentary behaviors and total sitting time.

3. How was the randomization done for the test-retest study? Seems odd that the subsample is different in so many aspects from the total sample. How many declined to participate in the test-retest?

Older adults that participated in the test-retest study were randomly selected by the researcher (stratified by gender) and asked if they were willing to answer an additional questionnaire during the second home visit. If they agreed, the sedentary behavior questionnaire was assessed a second time during the second home visit. Unfortunately, we did not keep record of the response rate for our reliability study. To clarify this, the following was added to the Methods section:

Additionally, participants were randomly selected by the researcher (stratified by gender) and asked whether they were willing to answer an additional questionnaire during the second home visit. During the second home visit a structured interview assessed demographic factors, anthropometric measures (weight and height) were performed and the accelerometer was collected. In a subsample of 28 participants who agreed to answer the additional questions (response rate not recorded), the same questionnaire targeting engagement in different sedentary behaviors was administered for the second time to assess its test-retest reliability.

4. Please include a reference on why only n=28 are enough for valid test retest analyses.

There are no clear guidelines on the necessary sample size for valid test-retest analyses. A sample size calculation could be performed but then appropriate 95% confidence intervals should be defined for which no clear guidelines are available (Giraudeau, Statist. Med., 2001). As recommended by Terwee et al. (2010) we presented 95% confidence intervals of the ICCs to provide the precision of our estimates. However, it is a limitation that we could not perform the requested subgroup analyses for the test-retest reliability (which were requested by reviewer 1). We added this to the limitations:
Fourthly, the size of the subsample for our reliability analysis was not sufficient to perform reliability analyses in different demographic subgroups.

5. There seems to be one question missing: How many days do you sit and eat? Why was this question left out?

Since eating can be expected to occur on a daily basis, time spent sitting while eating was assessed with only one question. This question targeted the usual time spent sitting while eating in the last seven days. To clarify this, the following was added to the methods section:

Except for usual time spent sitting while eating, all specific sedentary behaviors were assessed with two open-ended questions. Similar to the IPAQ, a first question assessed on how many days the behavior was performed in the last seven days, while the second question prompted how long, on average, the participant engaged in that sedentary behavior on such a day. Since eating can be expected to occur on a daily basis, sitting while eating was assessed with one question targeting the usual time spent sitting while eating in the last seven days.

6. How was the questionnaire developed? Theory lacks. Since PA and sedentary behavior is considered as separate constructs, why did you use IPAQ?

In response to comments 1 and 3 raised by reviewer 1, the rationale behind the development of the questionnaire has been extended:

To include a wide range of sedentary behaviors, we combined the sedentary behaviors included in previous questionnaires [15, 25, 26], and complemented these with additional sedentary behaviors relevant for older adults. More specifically, we subdivided questions targeting sitting in a car into driving a car, being a car passenger and using public transport. We added one question targeting usual time spent sitting while eating in the last seven days (in minutes/day) as was suggested by Gardiner et al. [15]. Furthermore, we added questions targeting sitting while doing household chores (e.g., ironing, preparing a meal).

The new questionnaire was pilot-tested in a convenience sample (n=4) of community-dwelling Flemish older adults to assess older adults’ understanding and completeness of the different items. Researchers involved in data collection were explicitly trained to ensure participants reported sedentary behaviors in which they engaged during the last seven days and did not duplicate their reported sedentary times across different sedentary items.

The authors agree that physical activity and sedentary behaviors are separate constructs, but this does not imply they should be measured by different formats. The use of comparable question formats throughout a questionnaire can facilitate comprehension and ease of administration. As described in the Methods section, we used a similar format as the IPAQ because studies examining sedentary behavior will most likely also include measures of physical activity. Since, the IPAQ is the most frequently used physical activity questionnaire, we chose to develop our questionnaire in a similar format. In the current study, the IPAQ was administered prior to the sedentary behavior questionnaire. We experienced that after completing the IPAQ, participants were really acquainted with the questionnaire format which facilitated responding to the sedentary behavior questionnaire. To clarify this, we added the following to the Methods section:
In the current study, a version of the IPAQ, specifically adapted for administration among Flemish older adults, was completed prior to the sedentary behavior questionnaire. This version of the IPAQ only included questions targeting physical activity behaviors and did not include questions targeting sitting time.

We believe that using a similar format as the IPAQ is an advantage of our questionnaire and have added the following to the Discussion section:

Secondly, our questionnaire had a similar format as the IPAQ, which we used to increase the ease of administration (since the participants were acquainted with the format by previously completing the IPAQ).

7. Was mental status assessed? This may be of importance to people >85-90yrs.

The authors agree that cognitive functioning may cause problems to answer the questions among the oldest old. However, we did not include a specific question targeting mental status. The issue is partially addressed by the mere inclusion of home-dwelling older adults which excludes the possibility that our participants suffered from serious cognitive impairment. Furthermore, as was requested by reviewer 1, we now performed separate validity analyses in the younger and older participants. Comparison of their results revealed that the validity results for the younger were better than those for the older participants. This might indeed be attributable to decreased cognitive functioning among the oldest age group. Therefore, the following has been added to the Discussion section:

Furthermore, better cognitive functioning and capacity to recall and report past sedentary behaviors among 65-74 year old compared to 75+ year old participants might explain the better validity results in the younger age group.

8. You state that 508 is the total sample, and 442 wore accelerometer. There is 7 elderly missing according to your explanation? Please clarify.

As mentioned in the description of the self-reported sedentary behaviors, seven participants were excluded because they reported total sitting times higher than 18h/day. To further clarify this, we added the following:

This resulted in the inclusion of 442 participants with complete questionnaire and accelerometer data with a mean of 15.0 ± 1.4 valid hours/valid day.

We also added the following to the description of the statistical analyses:

For the criterion validity analysis, the total analytic sample included 442 participants.

9. The title of table 1 is total sample, but yet you use 442. Is this your total sample then, and not 508?

Our total sample is indeed 508, but we had complete questionnaire and accelerometer data from 442 participants. We clarified this by adding ‘analytic’ before ‘sample’ in the title of Table 1 and in Table 1 itself.
10. It seems strange that the target period for the questionnaire and the accelerometer was not the same. Why was this chosen? Is this a major limitation to the study?

This issue was also raised by reviewer 1 (comment 11). We provided the following response:

The authors agree that it would have been better if the self-report assessment overlapped with the accelerometer assessment. As described in the discussion section, this was the case for the retest assessment of the questionnaire in the subsample in which the reliability analyses were performed (n= 28). These are the results for the validity analyses within this subsample:

<table>
<thead>
<tr>
<th>Spearman’s rho</th>
<th>Regression equation:</th>
<th>Bland-Altman procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D = b₀ + (b₁ x A)</td>
<td>Standard deviation of the residuals</td>
</tr>
<tr>
<td>0.32</td>
<td>-345.63 + (0.39 x A)</td>
<td>149.58</td>
</tr>
</tbody>
</table>

Although $b₁$ was smaller than for the validity analyses in the large sample, the Spearman’s rho, standard deviation of the residuals, D and LOA were very similar. Therefore, we chose to use the validity analyses in the large sample since our conclusions would not be altered. This also allowed us to perform the subgroup analyses requested by the reviewer.

We considered to present the above results in the manuscript, but decided it would complicate things unnecessarily. Of course, we could add these results if deemed appropriate by the reviewer.

11. Your results were “better” than ref. 16. In what way?

We agree that this should be described more precisely. To address this comment, we added the following to the Discussion section:

Validity was not strong, with a Spearman correlation of 0.30 in the total sample and wide limits of agreement. However, these relationships were stronger than those reported by Hekler et al. [16], who found a correlation of 0.12 among US older adults, and are comparable to the findings of Gardiner et al. [15], who reported a correlation of 0.30 among Australian older adults. Furthermore, with 82 minutes/day for a mean average of self-
reported and accelerometer-derived sedentary time, our self-report measure of total sitting time underestimated accelerometer-derived sedentary time substantially less than the self-report measures used in the previous studies [15, 16].

12. Other authors have stated that ICC is ‘excellent’ (# .81), and ‘good’ (.61 -.80). Why do you think that your 0.7 is a “conservative” measure of acceptable values?

We used the term ‘conservative’ because many authors consider 0.60 as acceptable test-retest reliability while we could not find any reference that justifies this. We used 0.70 as the threshold for acceptable reliability since this was the recommended threshold to evaluate physical activity questionnaires proposed by Terwee et al. (Sports Med, 2010). To eliminate ambiguity, we have deleted the sentence stating that 0.70 is ‘conservative’.

13. Please state what an “acceptable value” of Spearman’s coefficient is (with reference). This will make it easier for the reader to assess your value of 0.3.

We agree that providing a definition of an acceptable value would facilitate interpretation of the findings. Therefore, we added the following to the Analyses section:

For physical activity questionnaires, Terwee et al. [34] proposed to use a threshold of 0.50 to define a measure of self-reported total physical activity as valid against accelerometer counts.

14. Please make the limitation and strength of the study more clear. There is a suggestion for the use of log books, yet I assume this is not a limitation?

We have reformulated and re-organized the limitations and strengths section as follows:

A first strength of the current study is the examination of a sedentary behavior questionnaire that included an extensive list of specific sedentary behaviors. Secondly, our questionnaire had a similar format as the IPAQ, which we used to increase the ease of administration (since the participants were acquainted with the format by previously completing the IPAQ). Thirdly, we investigated differences in validity according to age, gender and education.

Our study has limitations, however. First is the use of accelerometers as criterion measure to assess sedentary behavior. Secondly, we only examined the validity of self-reported total sitting time and not the validity of self-reported specific sedentary behaviors. Future studies could include sedentary behavior log books to assess the validity of self-reported specific sedentary behaviors. Thirdly, older adults who were limited by their health to walk a couple of 100 meters were excluded from the current study. Therefore, our findings could not be generalized to older adults with such mobility impairment. Since mobility-impaired older adults may be at increased risk for high levels of sedentary time, future research should investigate the measurement properties of sedentary behavior questionnaires in this subgroup. Fourthly, the size of the subsample for our reliability analysis was not sufficient to perform reliability analyses in different demographic subgroups. The subsample for our reliability analysis was also rather highly educated which might have led to better reliability results.
15. What do you mean by tertiary education? University? How many years?

We have added detail to our measure of educational level as follows:

*Educational level was assessed using a 6-point scale ranging from having completed primary to university education. This was dichotomized as non-tertiary and tertiary (including college and university) education.*

16. Couple of hundred meters is rather imprecise. Do you mean 200? Or is it 1 km as in table 1?

This inclusion criterion was derived from a question included in the SF-36, the most frequently used questionnaire to assess health status and quality of life. In the physical functioning subscale of this questionnaire, participants are asked whether and to which degree their health limits several daily activities. Response categories are: (1) yes, seriously limited, (2) yes, somewhat limited, and (3) no, not limited. First, the questionnaire asks whether and to which degree the participant is limited to walk more than 1 kilometer. Consecutively, it targets being limited to walk a couple of hundred meters. Older adults reporting to be seriously or somewhat limited to walk a couple of hundred meters were excluded from participation. The following has been added to provide additional information about this inclusion criterion:

*The latter criterion was derived from an item included in the SF-36, the most frequently used questionnaire to assess health status and quality of life [20, 21]. Participants were asked whether and to what degree they were limited by their health to walk a couple of 100 meters. Response categories are: (1) yes, seriously limited, (2) yes, somewhat limited, and (3) no, not limited. Those who reported being seriously or somewhat limited were excluded from participation.*

**Minor Compulsory Revisions**

1. Please include the word “new” before questionnaire in the aim of the study.

“New” has been added to the study aim.

2. **Please be consequent in using numbers, not letters. (i.e. 508)**

We have modified this sentence as follows:

*This resulted in 508 participating in the study, a response rate of 44.8% (508/1,135 eligible participants found at home).*

3. **Key words: please include self-report and accelerometer**

These have been added to the Key words.
4. Please use the same heading of table 1 as in the text when referring to it (results).

This has been modified, the heading and text are now the same.

5. There is a spelling mistake in ref. 22. Please check all references.

The authors thank the reviewer for noticing this. Apparently, this is a spelling mistake in Web of Science from where the reference was imported into Endnote. We have checked all references.

6. Please have someone "native speakers" read for corrections and better language.

A native speaker has read and corrected our manuscript to improve the language.