Author's response to reviews

Title: Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L)

Authors:

Veerle Van Holle (veerle.vanholle@ugent.be)
Ilse De Bourdeaudhuij (ilse.debourdeaudhuij@ugent.be)
Benedicte Deforche (benedicte.deforche@vub.ac.be)
Jelle Van Cauwenberg (jelle.van.cauwenberg@vub.ac.be)
Delfien Van Dyck (delfien.vandyck@ugent.be)

Version: 3  Date: 28 March 2015

Author's response to reviews: see over
Ghent, March 2015

MS: 7078992301485715

Title: Assessment of physical activity in older adults: validity and reliability of the International Physical Activity Questionnaire (IPAQ), long interview version

Dear Editor,

We thank you for considering the publication of our manuscript “Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L)” (MS: 7078992301485715) in BMC Public Health.

We thank the reviewers for their valuable comments, which enabled us to further improve our paper.

In the document below, we responded to each of the comments made by the reviewers and indicated all modifications to the manuscript. According to the Editor’s suggestion, we paid particular attention to Reviewer #1’s comments pertaining to the methodology, and Reviewer #2’s fifth compulsory revision. All modifications have been highlighted in the manuscript itself.

We hope that these changes made our paper acceptable for publication in BMC Public Health.

On behalf of the authors,

Yours sincerely

Veerle Van Holle
REVIEWER #1

REVIEWER’S REPORT:
This is an interesting study on a timely topic: the validity and reliability of an existing physical activity questionnaire – often validated in adult populations of ages 18-65y – in a population of Belgian elderly individuals. The authors have adapted and translated an existing questionnaire for this specific population. Thereafter, they have determined the validity of this questionnaire in determining weekly minutes of MVPA and total PA in relation to hip-worn accelerometer data. They have looked at two definitions of MVPA and total PA cutoffs for accelerometer data, and have also determined the test-retest reliability of the questionnaire. The manuscript is generally well written, but the methods need to be more thoroughly explained, in particular the choices for the methods applied need to be well reasoned, and more explanation is needed on the recruitment and interview procedures. Further, some improvements to the results are necessary, in particular the inclusion of a flowchart of individuals included in the study. In addition, the discussion needs improvement on the discussion of results, strengths and limitations, and implications. Finally, some suggestions for improvements in English writing, and further clarifications in the text are suggested within this report (included as Minor essential revisions, and some as Discretionary Revisions).

MAJOR COMPULSORY REVISION
1. This manuscript reports on a questionnaire that has been adapted for a specific population, i.e. older European (Belgian) adults. Moreover, as is described in the article of Helmerhorst et al. (Int J Behav Nutr Phys Act, 2012), which the authors cite in this manuscript, the validity of a questionnaire is highly dependent on the population of interest. Therefore, the population being investigated, as well as the fact that the questionnaire is adapted and the abbreviation used throughout the manuscript (IPAQ-L), should be clearly stated in the title of the manuscript, e.g. ‘Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L)’.

Response:
We acknowledge the Reviewer’s comment and have adapted the title according to her recommendations. This way, we have stated more clearly that we used an elderly-adapted version of IPAQ. Furthermore, the title now also includes the abbreviation “IPAQ-L”, referring to the long version of the questionnaire.

The original title “Assessment of physical activity in older adults: validity and reliability of the International Physical Activity Questionnaire (IPAQ), long interview version” was changed into

“Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L)”.

Furthermore, the methods section of the abstract was completed with information on the specific modifications made to IPAQ-L:

Page 2, lines 9-12: (modifications underlined, in italics)
“Participants (n=434) completed the last seven days version of IPAQ-L, modified for the Belgian population of community-dwelling older adults. This elderly-adapted version of IPAQ-L combined vigorous and moderate activities, and questions on gait speed and recreational cycling were added. Furthermore, participants wore an ActiGraph GT3X(+) accelerometer for at least five days.”

2. The population of interest should also be clearly stated in the objective of the abstract.

Response:
On page 2, lines 6-7, the abstract was modified to meet the recommendation made by the Reviewer: (modifications underlined, in italics)

“The present study examined criterion validity and test-retest reliability of the long International Physical Activity Questionnaire (IPAQ-L), adapted for Belgian, community-dwelling older adults (65y and older).”

3. The authors indicate the fact that this questionnaire has been specifically adapted for this population as a strength in their discussion. However, they should also clearly state that this also leads to a limited generalizability of this adapted version of this questionnaire to other populations.

Response:
We understand the Reviewer’s concern and have added the following text on the limited generalizability of the sample (modifications underlined, in italics):

Page 22, lines 10-14: “Additionally, as this study used an elderly-adapted version of IPAQ-L, our results are not generalizable to younger populations. Moreover, adaptations also included specific items on recreational cycling, which is more prevalent in Western-European countries, compared to other continents such as North America (Pucher & Dijkstra, 2003). Therefore, the present study findings and utility of IPAQ-L may be less generalizable to non-Western-European populations of older adults.”

However, we still think that the elderly-specific adaptations of the questionnaire could be considered a strength of the present study as well. Therefore, the part of the discussion in lines 18-20 on page 22 was retained. To make this paragraph on the strengths more aligned with the part on the limitations, following text was added on page 22, lines 12-18:

“Despite these limitations, using an elderly-specific measurement method to assess PA levels in Belgian older adults could also be considered a strength of the present study, because the inclusion of recreational cycling and walking pace items may have improved content validity of the questionnaire for this population.”

4. The conclusion section of the abstract indicates the important finding that the IPAQ-L is more valid for measuring total PA than MVPA. However, these results are not included in the results section of the abstract. Furthermore, the next sentence is not an immediate logical consequence of the first sentence of this conclusion (i.e. first sentence indicates that the questionnaire is better in
measuring total PA than MVPA, while the second sentence indicates that this could be improved by including more LPA items, which seems contradictory). It seems that these two sentences represent two separate terms: the first concerns criterion validity (i.e. comparison with accelerometer data), while the second sentence concerns content validity of the questionnaire. Therefore, this should be explained.

Response:
We acknowledge this comment and adapted the abstract accordingly:

In lines 18-21 (page 2), following information on results was added:
“However, plots on agreement between self-reported and accelerometer PA showed a systematic over-reporting of IPAQ-L for MVPA. In contrast, plots indicated that IPAQ-L under-estimated levels of total PA, however, this under-estimation of total PA was substantially lower than the observed over-reporting of MVPA”.

Moreover, we agree with the Reviewer that the first part of the conclusion reports criterion validity, while the second refers to content validity. This was added to the abstract and to the discussion.

Page 3, lines 1-3 of the abstract (modifications underlined, in italics):
Criterion validity results suggest that IPAQ-L is more valid to measure older adults’ weekly minutes of total PA than weekly MVPA minutes. Moreover, results might imply that content validity of IPAQ-L can be improved if specific light-intensity PA items are incorporated into IPAQ-L.

Discussion (page 19, lines 4-19):
[The adaptations made to the discussion are also reported in the response on Comment #3 of the discretionary revisions]

“Based on this, the question arises whether or not it is more convenient to change the focus of PA questionnaires in older adults from MVPA to total PA levels, concurrently accounting for their most prevalent behavior, LPA, as well. This may also enhance content validity of IPAQ-L for estimating older adults’ total PA levels. Although our version of IPAQ-L already contains elderly-specific adaptations, specific questions on LPA (corresponding to activities with a Metabolic Equivalent (MET) value < 3) are still lacking. According to the 2011 Compendium of Physical Activities (Ainsworth et al., 2011), items on LPA could specifically ask for home activities such as “dusting or polishing furniture” (~2.3 METs); “washing dishes, clearing dishes from the table” (~2.5 METs), or “cleaning, sweeping, light effort” (2.3 METs), as these activities are likely to be reported as moderate-intensity instead of light-intensity. Examples of non-home based light-intensity activities could include “food shopping with/without a grocery cart while standing/walking” (~2.3 METs). Hence, the inclusion of specific LPA items might minimize over-reporting bias of MVPA. In addition, if IPAQ-L would be interviewer-administered, the interviewer could preserve possible duplicate over-reporting, by prompting participants when they tend to report the same activity twice (i.e., once in the LPA response, and once in the MPA response). Further research investigating applicability of elderly-specific items on light-intensity activities is needed, however.”
5. Within the methods section of the abstract, the variables used should be specified, e.g. ‘weekly minutes of self-reported moderate-to-vigorous (MVPA) and total PA’, rather than just ‘self-reported moderate-to-vigorous (MVPA) and total PA’. This should also be clearly stated throughout the manuscript.

Response:
We thank the Reviewer for noting this and have adapted the abstract according to this remark. If applicable, “weekly minutes” was added to the description of PA measures throughout the entire manuscript.

6. Methods, Sample and procedures: Please provide more information on the recruitment procedure in the manuscript. Potential participants were visited at home, but were these only visited once or multiple times if not at home? What times were they visited? If people were not at home and therefore classified as non-responders, how could this have affected your results (i.e., selection bias might have occurred as more active people have a higher chance of not being at home)? Please indicate the consequences of this within the Discussion of the paper. Please also include a flowchart concerning the number of individuals at the different stages of recruitment (i.e. potentially eligible, confirmed eligible, participants, etc.) and include in this flowchart also the reasons for non-response. This can be referred to in the Methods and Results section of this paper (see also point 17 of this report). Importantly, the number of included individuals is different in different parts of the manuscript (Methods, Sample and procedures: 508 participants; Methods, self-reported PA, paragraph 3: 438 participants; Results, Sample characteristics and PA levels: 434 participants), and this is not clearly explained. A flowchart would make this more clear.

Response:
We acknowledge that little information was provided on our recruitment procedures. Regarding the Reviewer’s concern on the potential selection bias, we think we made sufficient effort to reduce the chance of missing out the more active older adults, because each interviewer made up to three attempts (one first attempt and 2 additional attempts in case the selected older adult was absent at the time of visit) to recruit the selected older adults at home. Nevertheless, we agree with the Reviewer that this should be stated in the manuscript to avoid confusion. Following lines were added to lines 17-21 on page 7: (modifications underlined, in italics)

“An informative letter was mailed to contact selected older adults, in which the purpose of the study was explained and the visit of a trained interviewer during the subsequent two weeks (between 9 AM and 5 PM) was announced. Approximately one week later, potential participants were visited at home. in case of absence at the moment of visit, two additional attempts were made on different days and different times of the day (AM vs.PM) to reduce potential selection bias.”

In addition, we acknowledge that more clarity on the recruitment procedures, exclusion criteria and the size of the total analytic sample is needed. Therefore, a flowchart was created as suggested by the Reviewer (“see Figure 1”), providing an overview of each step in the recruitment and data cleaning process. We would also like to thank the Reviewer for noting the typing error of “438” instead of “434” participants (page 14, line 8) of the revised manuscript). This error was rectified.
Next to adding the flowchart in Figure 1, following parts of the methods were adapted to elucidate the recruitment procedures to the readers of the manuscript:

Validity sample (analytic sample size: n=434)
Page 12, lines 15-19:
“Based on the data, 74 participants of the validity study were excluded before analyses were conducted (see Figure 1). Twenty-five participants were excluded because their total reported levels of PA exceeded 6720 weekly minutes (~16 hours/day) (Sahlqvist, Song, & Ogilvie, 2012). Twenty-eight participants were excluded because of accelerometer failure and 21 participants had fewer than five valid days of accelerometer data. The total analytic sample for the validity study consisted of 434 participants.”

Reliability sample: (analytic sample size: n=29)
Page 8, lines 13-16: (modifications underlined, in italics)
In addition, during the second home visit, IPAQ-L was reassessed in a random sample (stratified on gender) of 30 participants, in order to collect data for examining test-retest reliability of the questionnaire. *This subsample’s socio-demographic characteristics were similar to those of the validity sample.*

Page 14, lines 3-5:
“The total analytic sample of the reliability analysis consisted of 29 participants, as one participant was excluded based on incomplete retest questionnaire data (see Figure 1).”

7. Methods, Measures, Self-reported PA: adapted IPAQ-L interview version, first paragraph:
Was there only one interviewer, or multiple interviewers? When the questionnaire was administered twice (for the reliability study), was it done by the same interviewer? This has important implications for the reliability results, as there is an important difference between inter-rater and intra-rater reliability. Please describe this in the methods.
In addition, the way the interviews were performed might also have influenced results (and the potential occurrence of information bias, see also point 28 of this report). Did the interviewers receive some sort of training for administering this questionnaire? Did they follow a standardized procedure? Please include information on this in the methods.

**Response:**
Multiple interviewers assessed the questionnaires, but all test-retest procedures within the same participant were performed by the same interviewer. We acknowledge that this should be elucidated in the manuscript and added following sentence to lines 16-17 on page 8:
“For the reliability study, both assessments of IPAQ-L were conducted by the same interviewer and the mean time interval between both visits was 9.6 ± 1.7 days.”

We were not able to statistically assess inter-rater differences because each participant was only visited by a single interviewer. However, several precautions were taken to minimize potential inter-rater bias.
Firstly, the interviewer questionnaire itself contains interviewer clarifications. Examples of the interviewer version can be found at [https://sites.google.com/site/theipaq/questionnaire_links](https://sites.google.com/site/theipaq/questionnaire_links) (online English telephone version of IPAQ-L). We followed IPAQ forward- and backward translation guidelines for translating our elderly-adapted version of IPAQ-L into Dutch (see page 10, lines 6-8) and hence, we think these translated interviewer prompts and clarifications were sufficiently appropriate to be used by each of our interviewers in this study.

Secondly, before initiating data collection, the first and fourth author (VVH and JVC) conducted some (n=5) interviews together, to ensure that the questionnaire was assessed similarly by different raters. All other interviewers (n=10) received a standardized interviewer training for assessing IPAQ-L, led by VVH or JVC (referred to as the “trainer” in the next part of this response):

The training consisted of three consecutive steps: 1) Firstly, trainees attended a session where the trainer contacted a participant at his/her home residence (going door by door), explained the procedures of the home visit, instructed participants how to wear the accelerometer, and assessed the interview. 2) During a second session, trainees contacted a new potential participant at his/her home residence and assessed an interview themselves, assisted by the trainer. 3) Lastly, a third participant was contacted and a third interview was conducted by the trainee. This last session was attended by the trainer, who corrected the trainee when this was needed.

We acknowledge that more elucidation on the training of the interviewers should be provided. Therefore, following text was added on page 8, lines 8-12):

“To minimize inter-rater bias, all interviewers received a standardized training to adequately conduct the complete home visit procedure (i.e., contacting selected older adults; explaining procedures concerning the interview and accelerometer data collection; assessment of the interviewer-administered questionnaire).”

8. Methods, Measures, Self-reported PA: adapted IPAQ-L interview version, second paragraph:

Please state the reason why items on walking were completed with an item assessing gait speed.

Response:

Following reason for including gait speed items was added to lines 20-24 on page 9: (modifications underlined, in italics)

“Secondly, since walking is the most prevalent type of PA in older adults (Lim & Taylor, 2005), items on walking were completed with an item assessing gait speed (i.e., low; moderate; high pace). Low-pace walking in older adults is classified as “light PA” (Ainsworth et al., 2011). Adding a gait speed item to IPAQ-L enables researchers to make a distinction between light- and moderate intensity walking.”

9. Methods, Measures, Self-reported PA: adapted IPAQ-L interview version, third paragraph:

Please state the reason why variables were truncated at specific maximum durations (e.g. total MVPA at 2520 minutes). Was this according to recommendations? If so, include a reference. Otherwise, clearly state the reason why specific cutoffs were chosen.
Response:
We agree that more clarification on the choice of truncation rules was lacking in the manuscript. Initially, it was intended to follow IPAQ-L scoring protocol recommendations (http://www.ipaq.ki.se/scoring.pdf; November 2005 version) to truncate the summary scores. However, these truncation rules only describe truncation for the “total walking, total moderate and total vigorous PA variables”. As our elderly-adapted questionnaire combines activities of moderate and vigorous PA, it was difficult to adopt these truncation rules on the data in the present study. Hence, it was considered appropriate to apply modified truncation rules.

A second reason for not applying the IPAQ-L truncation rules is that the IPAQ scoring protocol does not describe any truncation rules for the four domains of PA. In the present study, domain-specific truncation rules were applied on the summary variables for transport, domestic, leisure-time and work-related PA, respectively. These truncation rules were adopted from the truncation rules used in the study of Dubuy et al. (2013).

To make the truncation process more transparent, following text was added to the manuscript in lines 14-24 on page 10 and in lines 1-5 on page 11: (modifications underlined, in italics)

“Weekly minutes of PA were calculated for each specific reported PA behavior. Weekly minutes of total MVPA were calculated by summing weekly minutes of all reported PA behaviors, excluding walking at low pace for work-related, transportation and recreational walking, respectively. This new variable “Total MVPA” was truncated at a maximum of 1680 weekly minutes (~4 hours.day⁻¹). In addition, this variable was dichotomized according to the Public Health Recommendations (150 min.week⁻¹). Next, weekly minutes of time spent doing work-related PA, transportation PA, domestic PA, and recreational PA were calculated and subsequently truncated to a maximum of respectively 1800 (~6 hours.working day⁻¹), 1260 (~3 hours.day⁻¹), 1680 (~4 hours.day⁻¹), and 1680 (~4 hours.day⁻¹) weekly minutes. Truncated domain-specific variables were summed to create “Total PA”, which was truncated at a maximum of 2520 weekly minutes (~6 hours.day⁻¹). Data were truncated according to the truncation rules described by Dubuy et al.(2013), previously applied in Flemish adults. Data truncation is a commonly used method and is also described in the IPAQ scoring protocol guidelines (http://www.ipaq.ki.se/scoring.pdf). However, this IPAQ scoring protocol only provides separate data truncation rules for total walking, total moderate and total vigorous PA. In the present study, it was decided to apply domain-specific data truncation rules on the summary variables for transport, domestic, leisure-time and work-related PA, respectively (Dubuy et al., 2013), as our adapted version of IPAQ-L does not distinguish moderate from vigorous PA.”

[Reference Dubuy et al. (2013):]


10. Methods, Measures, Self-reported PA: adapted IPAQ-L interview version, third paragraph: Please explain what was done in case of missing values on one or multiple questions. Was this reduced by the interview assessment method?

Response:
Indeed, missing cases were reduced by the interview assessment method and there were no missing cases for the validity study. For the reliability study, one participant provided incomplete data on the retest questionnaire. As shown in Figure 1 (flowchart), this participant was excluded from the analytic sample.

11. Methods, Statistical analyses: Please provide some information on the power of this study, especially concerning the reliability study which was performed in only 29 individuals, which seems a rather low number. Was this sufficient for this type of analyses? (see also point 26 of this report)

Response:
We agree that the reliability study was conducted in a small sample of 29 older adults and that having a larger sample would be more beneficial. However, many other studies on the reliability of self-report data also reported results based on samples with fewer than 50 participants (e.g., Matton et al., 2007; Van Cauwenberg et al., 2014; Washburn et al., 1999). Furthermore, we would like to point out that although the power may be limited, some efforts were made to make our data transparent to the readers of the paper. Specifically, confidence intervals on the ICC measures were provided to show that there is indeed some uncertainty regarding our data (Table 2) and this observed variability in the data was discussed (e.g., page 20, line 16).

Nevertheless, as we agree with the Reviewer that this should be more clearly pointed out in the manuscript, we added the small sample size as a limitation of our study on page 21 lines 14-19:

“Secondly, the reliability analyses were performed in a small sample of 29 participants, which may have reduced the power of the reliability study. Nevertheless, other studies on reliability of self-report data also reported results based on samples with fewer than 50 participants (e.g., (Matton et al., 2007; Van Cauwenberg et al., 2014; Washburn & Ficker, 1999). Moreover, some efforts were made to make this limitation transparent to other researchers (i.e., ICC 95% confidence intervals showed that indeed, there is large inter-individual variability regarding our data).”

12. Methods, Statistical analyses: Please provide the reason why Spearman correlation coefficients were calculated, rather than Pearson correlation coefficients.

Response:
To explain why Spearman, and not Pearson, correlations were calculated, following sentence was added to page 12, lines 24, and page 13, line 1:

“Nonparametric Spearman rank order correlation coefficients were calculated because the data were non-normally distributed (Field, 2005).”

13. Methods, Statistical analyses: The last part of this paragraph on the reliability is unclear. Please indicate clearly which measures were calculated, for which variables. Specifically, provide information on the type of ICC being calculated (consistency or agreement, probably agreement?). Why was a Spearman correlation coefficient calculated, rather than the ICC for consistency, which also indicates the reliability in terms of ranking individuals?

Response:
We acknowledge that little information was given on the ICC calculation methods. For the original manuscript, ICC’s were calculated using the two-way mixed method with measures of consistency. However, given that we want to make inferences on the stability of participants’ responses from test to retest and (as the Reviewer suggested) that we are less interested in the stability of their relative rank order over time, it is better to include ICC’s of absolute agreement, rather than consistency. The Reviewer’s suggestion was also discussed with the statistical expert of our Department, who advised us to follow this suggestion for including absolute agreement ICC’s. Therefore, Table 2, results and discussion were adapted accordingly.

In addition, we acknowledge the Reviewer’s comment that reporting ICC’s for consistency would be superior to reporting Spearman’s rank order coefficients and that it may have been better to replace reports of Spearman’s rho coefficients by ICC’s for consistency. However, as explained in the first paragraph of this response, we consulted our statistical expert and based on her advice, we consider it more relevant to only report ICC’s for absolute agreement. Hence, Spearman’s rho’s were deleted from Table 2 and results/discussion.

Furthermore, the text in lines 22-23 on page 13 of the methods section was completed with following information on the ICC calculation method: (modifications underlined, in italics)

“Single measures Intra-class correlation coefficients (ICC’s; two-way mixed methods, absolute agreement) with 95% confidence intervals were calculated, comparing participants’ test-retest self-reported weekly minutes spent in four PA domains (work-related PA, transportation PA, domestic PA, and recreational PA) and summary measures of self-reported MVPA and total PA.”

14. Methods, Statistical analyses: Please provide the a priori hypotheses concerning the correlation coefficients and ICC’s calculated for comparing the questionnaire with accelerometer data, and for the test-retest reliability analyses. What did the authors expect to find?

Response:
We included a priori hypotheses in the Background section on page 6, lines 19-24 and on page 7, lines 1-11:

“In their review, Kowalski et al. (2012) described an average correlation coefficient of r=0.38 between older adults’ self-reported PA and direct measures (e.g., accelerometers). Moreover, the review of Helmerhorst et al. (2012) observed median Spearman correlations of r=0.41 between older adults’ self-reported and objectively-measured PA measures. Hence, in the present study, it was hypothesized that similar validity results would be obtained, with correlation coefficients ranging between 0.35 and 0.45. Weekly minutes of total PA and MVPA were defined using the cut points of Freedson et al. (1998), as well as those of Copeland & Esliger (2009). Since the Copeland & Esliger cut point for MVPA is elderly-specific, a second hypothesis concerning validity results of this study was that we hypothesized to find higher agreement between self-reported MVPA and accelerometer-derived MVPA defined by the Copeland & Esliger cut point, compared to validity results comparing self-reports with accelerometer MVPA defined by the commonly used adults-specific Freedson cut point. In addition to assessments of an instrument’s criterion
validity, it is also important to examine the stability of this instrument over time and whether it is capable of measuring a variable (e.g., PA) with consistency (Portney & Watkins, 2009). Hence, a second aim of this study was to assess test-retest reliability of the elderly-adapted IPAQ-L questionnaire. Previous studies on test-retest reliability of PA questionnaires in older adults reported median Intraclass Coefficient estimates of 0.65 (Helmerhorst et al., 2012) and similar estimates were hypothesized to be found in this Belgian study in older adults.”

Furthermore, these a priori hypotheses were addressed in the discussion (See also the response to comment #2 of the discretionary revisions).

- Discussion of the results in relation to the a priori hypothesis on reliability: (modifications underlined, in italics)

Page 20, lines 10-13:
“Concordant with our hypothesis, coefficients found in our study (i.e., ICC range 0.43-0.81) are comparable to those reported in other reliability studies on PA questionnaires for older adults. Specifically, a literature review observed median reliability ICC estimates of 0.65 in other studies conducted in older adult samples (Helmerhorst et al., 2012).”

- Discussion of the results in relation to the a priori hypothesis on validity: (modifications underlined, in italics)

Page 17, lines 6-10:
“As hypothesized, similar correlation coefficients for MVPA as observed in the present study were found in other validation studies using Spearman correlations between accelerometers and PA questionnaires in older adult populations (ρ=0.43 (Dinger et al., 2004); ρ=0.37(Hurtig-Wennlöf, et al., 2010); ρ=0.31(Kolbe-Alexander et al., 2006)). Moreover, a review on measurement properties of PA questionnaires described that validity of questionnaires in older adult samples showed a median Spearman correlation of ρ=0.41 (Helmerhorst et al., 2012).”

Page 17, lines 17-18:
“This finding is in concordance with our hypothesis that better validity of the elderly-adapted IPAQ-L would be observed regarding Copeland MVPA, which seems a promising finding.”

15. Methods, Statistical analyses:

The authors indicate that they had to apply a non-parametric approach for generating Bland-Altman plots. The resulting plots do not include any limits of agreement, for which the reason is not adequately explained in the text. In addition, a percentage is being calculated, but it is not clear whether this percentage is relative to self-reported or accelerometer data. Furthermore, no reference is included when explaining the non-parametric method, which makes it impossible to check the adequacy of this method. Please include the appropriate reference where this non-parametric approach is described. Also include information whether this method has been applied previously in other validity studies, as it is not familiar to this reviewer.

Response:
We acknowledge that limits of agreement were missing in the original manuscript and therefore, these were added to the plots in figures 1-3 and supplementary file 3. Limits of agreement were calculated based on the guidelines provided by Bland and Altman (1999), for which the reference was already provided in the methods section of the original manuscript (e.g., see line 10, page 13; reference [37]; and see the reference provided at the bottom of this response).

According to the recommendations of Bland and Altman (1999) about the “nonparametric approach to comparing methods” (Bland & Altman (1999), chapter 6 on page 157), and by mutual agreement with our Department’s statistical expert, nonparametric 90% limits of agreement can be calculated as the values outside which 10 percent of the observations fell.

Therefore, for each of the plots in Figures 2, 3, 4 and supplementary file 3, the 5th and 95th percentile were calculated to obtain nonparametric limits of agreement, which were subsequently superimposed on each scatter diagram, together with the median value. Furthermore, 95% confidence intervals for these limits were calculated as the standard errors of both 5th and 95th percentiles, using bootstrapping (these confidence intervals around each limit of agreement are reported in the legend below each Figure).

Additionally, since the data were non-normally distributed, the median (P50) was reported. Consequently, the initially reported mean values (for each type of accelerometer PA measure) were substituted by these medians throughout the results section (e.g., page 15, line 11).

To our knowledge, few studies have applied nonparametric approaches to report validity results. However, we could not apply parametric Bland-Altman plots because of following issue with the data (this reasoning is also described in the methods in lines 5-10 on page 13):

Generally, in a normal Bland-Altman plot, the difference between both measurement methods (i.e., self-report and accelerometer) is regressed on the average of both measurement methods. As in the present study, there was a significant relationship between differences and averages, a linear regression approach was proposed to construct Bland-Altman plots and to calculate limits of agreement. However, for constructing parametric Bland-Altman plots and calculating limits of agreement, it is required that the absolute values of the residuals of the linear regression are normally distributed. This was not the case in the present study, and therefore, a nonparametric approach is suggested (Bland & Altman, 1999).

Based on this and after discussing this with our statistical expert, we consider the nonparametric method the appropriate method to apply to our data. We hope the Reviewer agrees with our decision and that we made it more transparent to potential readers of the paper.

To elucidate the applied nonparametric approach and to explain the calculation of the nonparametric limits of agreement, following text was added to the description of the statistical analyses in lines 17-20 on page 13 of the methods:

“Specifically, for each of the 16 above-mentioned plots, 5th and 95th percentiles of the percentage-of-difference measures were calculated and subsequently superimposed on the scatter diagram, reflecting 90% limits of agreement. 95% confidence intervals around these limits were calculated by bootstrapping methods, using the standard errors of these 5th and 95th percentiles (Bland & Altman, 1999).”
Moreover, all relevant information on the scatter diagrams with 90% limits of agreement (and these limits’ confidence intervals) is now provided in the legend of each figure.

With regard to the lack of clarity on whether the percentages were relative to self-reported or accelerometer data, following clarification was added to lines 12-13 on page 13: (modifications underlined, in italics)

“According to the non-parametric method, the difference between self-reported PA (using IPAQ-L) and accelerometer-derived PA was calculated and expressed as a percentage of difference [self-report data as a percentage of accelerometer data].”

[Reference Bland & Altman (1999)]

16. Methods, Statistical analyses: Please include all cutoffs used for classifying Spearman correlation coefficients, as well as ICC’s (e.g. what is defined as ‘fair’, what is ‘good’, etc.).

Response:
According to the suggestion of the Reviewer, following clarification was added to the methods section.

For validity: (page 13, lines 1-3)
“Correlation coefficients ≥0.40 were considered as good validity, coefficients between 0.30 and 0.40 were classified as moderate, and coefficients <0.30 were categorized as poor validity (Portney & Watkins, 2009).”

For reliability: (page 14, lines 1-3)
“In the present study, ICC estimates >0.75 were considered as good reliability scores, ICC’s between 0.50 and 0.75 were classified as moderate, and ICC’s <0.50 indicated poor reliability (Portney & Watkins, 2009).”

Based on these clarifications, we acknowledge that the term “fair” in the original manuscript actually reflects “moderate” validity and/or reliability, and that the use of different terms may have caused confusion. Therefore, “fair” was replaced with “moderate” throughout the paper.

17. Sample characteristics and PA levels: Refer here to the flowchart presented in the methods when discussing recruitment procedures. If possible, please include some data on a comparison of certain characteristics (e.g. age, gender) between responders and non-responders, as this might shed some light on the potential occurrence of selection bias (see earlier comments in point 6 of this review report).

Response:
As suggested by the Reviewer in comment #6, a flowchart was added to elucidate the recruitment procedures. However, as the composition of the total analytic sample was already discussed in lines
15-19 on page 12 and in lines 3-5 on page 14 (reliability sample) of the methods section, we considered it redundant to repeat this information in the results section.

Regarding the second part of the Reviewer’s comment on including certain characteristics of responders vs. non-responders, we would also like to refer to the response on Comment #6 of the Reviewer’s report. There, we explained that the researchers aimed to minimize selection bias (i.e., interviewers made three attempts to reach the selected older adults at their home residence). Additionally, age and gender were stratification criteria to select the sample, so it was not considered useful to make inferences on responders vs. non-responders based on these demographic criteria.

Another reason for not including comparisons between responders and non-responders is that additional information on these groups was not available. Specifically, we were not able to verify whether all older adults that were not at home when the interviewers visited them, also still lived at the address provided by the public service of Ghent. Moreover, if they did live at the correct address, it would still not be possible to make inferences on whether or not the selected person would be willing to participate in the study or whether or not this person would be physically capable for participation in the study (i.e., able to walk 100 meters without severe physical difficulties).

Nevertheless, to come towards the suggestion of the Reviewer on the representativeness of the sample/possible selection bias, a comparison between both validity and reliability samples with the general Belgian population of older adults was included in the results section (see also our response on comment #24)

Following text was added to the results on page 14, lines 14-16:

“Fifty-four percent were women, which is similar to the gender distribution in Belgium (54% women; (National Institute of Statistics - Belgium, 2008). In contrast, compared to the Belgian population of older adults, a higher percentage of participants lived with a partner (65.8% versus 56.2% for Belgium; (National Institute of Statistics - Belgium, 2008).”

Representativeness was also discussed in the limitations of the manuscript (page 22, lines 5-14).

“Lastly, representativeness of both validity and reliability samples should be taken into account and results should be interpreted carefully. In general, both samples were comparable with the Belgian population of older adults, but both samples had higher percentages of participants living with a partner than the general population and the reliability sample (n=29) was higher educated. This may jeopardize generalizability to the lower-educated and those living alone. Additionally, as this study used an elderly-adapted version of IPAQ-L, our results are not generalizable to younger populations. Moreover, adaptations also included specific items on recreational cycling, which is more prevalent in Western-European countries, compared to other continents such as North America (Pucher & Dijkstra, 2003). Therefore, the present study findings and utility of IPAQ-L may be less generalizable to non-Western-European populations of older adults.”

18. Results, Table 2: P-values of the Spearman correlation coefficients tests are included. However, these are not informative as these do not represent the strength of association, but merely the chance that these findings were observed if the Spearman correlation coefficient was zero, which seems not relevant in this case as a correlation of 0 is highly unlikely. It might be more informative to include ICC for consistency with 95% confidence intervals, as a measure of the reliability of
ranking individuals (that is, if the currently reported ICC in this table is indeed an ICC for agreement, see also point 13 of this review report).

Response:
We refer to the response on Comment #13, in which we describe why the Spearman’s rank order correlation coefficients were deleted from Table 2 and why we chose not to include ICC’s for consistency.

19. Results, Figure 3: It is not clear why the authors have chosen to show in Figure 3 the results of Freedson total PA, as the Copeland cutoffs seem more relevant for this population, as these were validated in elderly individuals, and this study is also performed in a population of older adults. Therefore, it would be more logical to include the Copeland results in the figure and the Freedson results as supplementary material.

Response:
We thank the Reviewer for this suggestion and agree with this rationale for including the Copeland results in the figure and Freedson results as supplementary material. Therefore, Figure 4. Was changed into Supplementary File 3, and vice versa.

20. Discussion, Criterion validity: The results of this study show that especially for min/week of MVPA there is substantial over-reporting by participants. This should be clearly included in the discussion, and be discussed in relation to the Spearman correlation coefficients found. Maybe these questionnaires are quite good at ranking individuals in terms of their level of PA, but far worse in predicting actual values of min/week of PA. This has implications for the applicability of these questionnaires.

Response:
We thank the Reviewer for bringing this to our attention and for making an appropriate suggestion on the interpretation of the Spearman correlations regarding validity.
We would like to refer to our response on the next comment (see Comment#21), where we listed the changes made to the discussion section.

21. Discussion, Criterion validity:

The authors speculate on the applicability of the different cutoffs (Freedson vs Copeland). They should include a recommendation on further research, which can shine more light on the appropriateness of these cutoffs, as the results from this study are still speculative on which cutoff would be superior.

The authors indicate that they think that the Freedson cutoffs seem to be superior, but the Copeland cutoffs seem a priori more relevant for this population, as these were also validated in a population of elderly individuals. The authors indicate that the mean BMI within this sample was rather high, but do not give any information on the mean BMI of the populations in which these cutoffs were validated, nor on the relation between BMI and adherence to physical activity guidelines the authors implicitly suggest. It can be questioned whether their reasoning is plausible.
Further, the authors only describe that the ‘Copeland cutoffs were validated in a specific group of older adults’, please provide some more detail on the specifics of this group. Figure 2 seems to suggest that the Copeland cutoff results in less variability as the authors indicate themselves, which could suggest that this cutoff could be better than Freedson. However, further research is needed to adequately determine which cutoff would be adequate for this population.

Response:

The “specific population of older adults” in which the Copeland cut point for MVPA was defined, was referred to as “specific” because we wanted to stress the specificity for elderly. However, we assume that the Reviewer interpreted this “specificity” in terms of sample characteristics of this older adults sample. To avoid confusion, we modified the sentence in lines 1-3 on page 17, now reading:

“The second accelerometer cut point set was defined by Copeland & Esliger(Copeland & Esliger, 2009) in a sample of older adults and has a lower threshold value for MVPA (1,041 counts.min⁻¹).”

Although we agree with the Reviewer that there is a great need for adequate and applicable cut points to define MVPA in the elderly, the main aim of the present study was not to make inferences on the most applicable cut point (Freedson or Copeland). We apologize if this might have been understood this way by the Reviewer. In reality, our intention was to validate the IPAQ-L questionnaire against two different cut points in order to make a stronger conclusion towards its applicability in large populations. We think it is not appropriate to formulate recommendations on which cut point to use in future research on older adults, as our data did not allow us to check actual energy expenditure indicators (e.g. oxygen consumption) and compare these with the accelerometer counts.

Because the comparison between both cut points was both speculative and confusing, this part of the text was softened in the discussion. Given that the reasoning on bmi prevalence was questioned by both Reviewer#1 and Reviewer #2, this was removed from the discussion. Moreover, a substantial part of the discussion was modified in order to elucidate the most important finding of our study, i.e., that our elderly-adapted version of IPAQ-L might be improved by assessing light-intensity PA as well and by including more items that specifically assess this type of activity intensity.

Page 17, lines 10-24 and page 18, lines 1-13 were adapted as follows: (modifications underlined, in italics)

“These moderate-to-good Spearman coefficients should be interpreted with care, since they can only suggest that IPAQ-L was relatively good at ranking individuals by their reported MVPA levels, when compared to their objectively-measured MVPA levels. In order to evaluate validity of the IPAQ-L, it is recommended that these Spearman rank order correlations are interpreted in combination with the results of the agreement plots. Plots for MVPA showed a general pattern of over-reporting of IPAQ-L and this over-reporting was higher when self-reported MVPA was compared with the Freedson MVPA, than when it was compared to the Copeland MVPA. This finding is in concordance with our hypothesis that better validity of the elderly-adapted IPAQ-L would be observed regarding Copeland MVPA, which seems a promising finding. On the other hand, it is also logical to find less over-reporting when self-reports are compared to Copeland MVPA, because the cut point has a lower threshold for defining MVPA than the Freedson MVPA cut point. However, also the Copeland MVPA
cut point might not be the ideal solution for defining accelerometer MVPA in older adults. Percentages of participants in the current study reaching the Public Health Recommendations for MVPA were 27.4% and 71.2% for Freedson and Copeland, respectively. Given that worldwide, only 30-40% of older adults reach the Public Health Recommendations for PA (Centers for Disease Control and Prevention, 2013; Eurobarometer, 2010; Tafforeau, 2008), the percentage found in the present study using the Copeland MVPA threshold seems rather high and Freedson MVPA levels could be a more realistic estimate of participants’ actual MVPA levels. Although it is beyond the purpose of the present study to make inferences on the applicability of cut points for defining MVPA in older adults, our findings do indicate that researchers should be cautious when selecting an appropriate accelerometer threshold for defining MVPA in older adults. Moreover, it is suggested that more research is conducted on the feasibility of different MVPA cut points in older populations.

Irrespective of the cut point applied, however, the plots for validity of IPAQ-L regarding MVPA showed that the highest over-reporting was found in older adults with lower objectively-measured activity levels (i.e., the lower quartiles shown in the plots). As PA levels show a systematic decline with increasing age (Schrack et al., 2014), this may have important consequences for the utility of IPAQ-L in longitudinal studies, and in the oldest old.”

Page 19, lines 4-19: (modifications underlined, in italics)

“Based on this, the question arises whether or not it is more convenient to change the focus of PA questionnaires in older adults from MVPA to total PA levels, concurrently accounting for their most prevalent behavior, LPA, as well. This may also enhance content validity of IPAQ-L for estimating older adults’ total PA levels. Although our version of IPAQ-L already contains elderly-specific adaptations, specific questions on LPA (corresponding to activities with a Metabolic Equivalent (MET) value < 3) are still lacking. According to the 2011 Compendium of Physical Activities (Ainsworth et al., 2011), items on LPA could specifically ask for home activities such as “dusting or polishing furniture” (~2.3 METs); “washing dishes, clearing dishes from the table” (~2.5 METs), or “cleaning, sweeping, light effort” (2.3 METs), as these activities are likely to be reported as moderate-intensity instead of light-intensity. Examples of non-home based light-intensity activities could include “food shopping with/without a grocery cart while standing/walking” (~2.3 METs). Hence, the inclusion of specific LPA items might minimize over-reporting bias of MVPA. In addition, if IPAQ-L would be interviewer-administered, the interviewer could preserve possible duplicate over-reporting, by prompting participants when they tend to report the same activity twice (i.e., once in the LPA response, and once in the MPA response). Further research investigating applicability of elderly-specific items on light-intensity activities is needed, however.”

22. Discussion, Test-retest reliability: The authors indicate that the Spearman’s rank order coefficients are larger than ICC’s, but do not give a reason for this nor explain the implications of this finding. Please include this. It can be questioned whether it wouldn’t be better to include both types of ICCs (consistency and agreement), and maybe exclude the Spearman correlation coefficient.

Response:
We refer to our response on the Reviewer’s previous comment #13. Given that Spearman’s rank order correlations are less useful for reporting test-retest reliability indications, the Spearman’s rho’s were deleted from Table 2, results and discussion.

23. Discussion, Limitations and strengths, first paragraph:
The part of the sentence ‘Accelerometer-derived levels of total PA and MVPA may have differed from participants’ actual levels,...’ is really vague. Please explain why this is the case. An explanation might be that a hip-worn accelerometer may not accurately measure cycling. Explain how this might have influenced results, as therefore the accelerometer data is no real golden standard for determining criterion validity. Also specify variables measured (e.g. ‘time spend in total PA and MVPA’), rather than ‘levels of total PA and MVPA’.
The second part of that sentence:
‘which could only be determined by the current gold standard; the doubly-labeled water method [35].’ is not true. Double labelled water is the gold standard for measuring average energy expenditure over a long period of time. However, the aim of this questionnaire is not to measures energy expenditure, but to measure minutes in activities with a certain energy expenditure, which can never be measured with double labelled water. Please adapt this part of the discussion appropriately.

Response:
Based on the Reviewer’s suggestion, we specified why accelerometer-derived measured may have differed from self-report data. Following text was added to the manuscript on page 21, lines 20-23 “Specifically, accelerometers were worn at the hip and may not have appropriately captured upper-body movements, cycling and water-based activities (Swartz et al., 2000), which may have consequently caused an under-estimation of time spent doing PA.”

Concerning the remark on Doubly Labeled water, we agree with the Reviewer on this poor interpretation and adapted the limitations according to her suggestion. The comparison of accelerometer and DLW measurements was deleted from the text and replaced with following explanation in lines 23-24 on page 21 and lines 1-2 on page 22: (modifications underlined, in italics)
“However, accelerometers are able to accurately discriminate between frequency and intensity of PA and are considered to be an acceptable criterion to validate self-reported total PA and MVPA (Pruitt et al., 2008). Besides, several previous studies on the validity of PA questionnaires for older adults have utilized Actigraph accelerometers as the criterion measure (Cerin et al., 2012; Dinger et al., 2004; Grimm, Swartz, Hart, Miller, & Strath, 2012; Hurtig-Wennlöf et al., 2010; Kolbe-Alexander et al., 2006).”

24. Discussion, Limitations and strengths, first paragraph: The limited generalizability of this questionnaire to populations outside of the target populations should be clearly discussed. This should also be included in the Conclusions. In addition, the authors should discuss the potential selection bias resulting from their recruitment procedure resulting in a substantial number of non-responders, which might differ from the individuals included in this study and might further influence the generalizability of the findings of this study. (see also points 6 and 17 of this report).

Response:
We refer to our responses on Comment #6 (and #17), in which we explained that we added information on how we tried to deal with possible response bias, and that we added some text regarding the limited generalizability of our findings in the limitations on page 22, lines 6-14.

Furthermore, to come towards the Reviewer’s suggestion on including some comparison data, the first part of the results section (sample characteristics) was completed with a comparison between both validity and reliability samples and the general Belgian population of older adults, from which the sample was drawn.

Following text was added to the results on page 14, lines 14-16:
“Fifty-four percent were women, which is similar to the gender distribution in Belgium (54% women; (“National Institute of Statistics - Belgium, 2008 ). In contrast, compared to the Belgian population of older adults, a higher percentage of participants lived with a partner (65.8% versus 56.2% for Belgium; (“National Institute of Statistics - Belgium, 2008 ).

In addition and as mentioned before in the response on Comment #17, we discussed the representativeness of the validity and reliability samples in the limitations of the manuscript (page 22, lines 5-14).

“Lastly, representativeness of both validity and reliability samples should be taken into account and results should be interpreted carefully. In general, both samples were comparable with the Belgian population of older adults, but both samples had higher percentages of participants living with a partner than the general population and the reliability sample (n=29) was higher educated. This may jeopardize generalizability to the lower-educated and those living alone. Additionally, as this study used an elderly-adapted version of IPAQ-L, our results are not generalizable to younger populations. Moreover, adaptations also included specific items on recreational cycling, which is more prevalent in Western-European countries, compared to other continents such as North America (Pucher & Dijkstra, 2003). Therefore, the present study findings and utility of IPAQ-L may be less generalizable to non-Western-European populations of older adults.”

On page 24, lines 8-10, following sentence was added to the conclusion: 
"More research on the feasibility of IPAQ-L in populations of older adults is recommended, as the results of the present study are probably only applicable to Western-European settings and may not account for all population subgroups.

25. Discussion, Limitations and strengths, first paragraph:
A further limitation that needs to be discussed is the limited power of the reliability study (see also point 11 of this review report). Was the number of 29 individuals sufficient? The confidence intervals are rather wide, which indicates that power was limited.

Response:
We refer to the Response on comment #11. In that response, we acknowledged that the sample size was indeed limited, but that we tried to make our results transparent to the readers by reporting the variability in results.
Moreover, we referred to the fact that other reliability studies also made inferences on questionnaires’ reliability based on smaller samples (e.g., Matton et al., 2007; Van Cauwenberg et al., 2014; Washburn et al., 1999).

As stated in the response on comment #11 and to come towards the Reviewer’s suggestion, the small sample size for the reliability study was added as a limitation of our study on page 21, lines 14-19:

“Secondly, the reliability analyses were performed in a small sample of 29 participants. Although other studies on reliability of self-report data also reported results based on samples with fewer than 50 participants (e.g., (Matton et al., 2007; Van Cauwenberg et al., 2014; Washburn & Ficker, 1999), the small sample size may have reduced the power of the reliability study. Nevertheless, some efforts were made to make this limitation transparent to other researchers (i.e., ICC 95% confidence intervals showed that indeed, there is large inter-individual variability regarding our data).”

26. Discussion, Limitations and strengths, seconds paragraph: The authors indicate that some LPA activities might have been reported as MVPA, which could have contributed to over-reporting of MVPA levels. They also state that combining moderate with vigorous activities might have reduced this bias. This is not logical as currently stated; please include a thorough explanation why combining moderate and vigorous activities could have led to less LPA activities being recorded as MVPA.

Response:
The Reviewer made a pertinent remark on the ambiguity of this sentence. We believe that combining moderate and vigorous items into a single item could have reduced reporting bias, since it may have contributed to a minimization of cumulative over-reporting. Specifically, it could occur that participants over-reported their time spent in vigorous activities, and also in moderate activities. Combining items on vigorous and moderate activities may have minimized this cumulative bias on the reported time spent doing MVPA. In addition, combining items on moderate- and vigorous-intensity items could also minimize the chance for duplicate reports of moderate- and vigorous-intensity activities that are very similar (i.e., an event that only occurred once in reality could be reported twice by the participant: once as a response on the moderate-intensity item, and once as a response on the vigorous-intensity item).

However, we do not completely agree with the last part of the Reviewer’s comment, because we think that combining moderate and vigorous items cannot be directly linked to the minimization of LPA being recorded as MVPA. We aimed to state that despite combining moderate- and vigorous-intensity items may have been a good step towards minimization of over-reporting, still some over-reporting may have occurred because older adults may have reported LPA as being MVPA (e.g., light-intensity household activities may have been reported as moderate-intensity activities).

Nevertheless, we agree that this was ambiguously stated in the description of the strengths, and that this should have been more thoroughly explained. Therefore, we modified this part of the strengths:

Page 22, lines 18-24 and page 23, lines 1-3:
“Moreover, as suggested by Cerin et al. (2012), our elderly-specific version of IPAQ-L put less emphasis on items on vigorous-intensity PA by combining vigorous and moderate questions into one item. Combining these items could have reduced over-reporting bias, since it may have contributed to a minimization of cumulative over-reporting (i.e., accumulation of over-reporting on the MPA item and over-reporting on the VPA item). In addition, combining moderate- and vigorous-intensity items may have reduced the chance for duplicate reports (i.e., the same activity being reported twice: once as a response on the item regarding moderate-intensity, once on the item for vigorous-intensity). However, apart from that, researchers should be aware that despite combining moderate- and vigorous-intensity items may have been a good step towards minimization of over-reporting, still some over-reporting of MVPA may occur because older adults may have reported activities of light intensity as MVPA.”

27. Discussion, Limitations and strengths, seconds paragraph: The authors discuss that interviewer guidance was expected to be beneficial for an elderly population. Please indicate more clearly how this might have influenced results (e.g. more precise reporting because the interviewer can ask questions to trigger the memory of participants on activities that could possibly be forgotten otherwise, on the other hand it could also lead to more over-reporting, because participants might want to ‘please’ the interviewer and therefore indicate higher time in certain activities, which could lead to information bias). In this part, the authors can refer back to the methods described on the interview procedures (see point 7 of this report).

Response:
The text on page 23, lines 4-12 was completed with following clarification: (modifications underlined, in italics)

“A Belgian study in adults (18-65y) observed less over-reporting bias when the interviewer-administered version versus the self-administered version of IPAQ-L was assessed (Van Dyck, Cardon, Deforche, & De Bourdeaudhuij, 2014). Additionally, older adults may experience more cognitive difficulties when responding to a questionnaire (Matthews, 2002), and therefore, the guidance by trained interviewers is likely to be beneficial for this older adults population, because more accurate responses can be obtained. For instance, interviewers could prompt participants to trigger their memory on activities that might have been forgotten otherwise. On the other hand, interviewers could also point out possible overlap in reports of certain activities (e.g., walking for leisure and walking for transport).”

28. Discussion, Conclusions:

Please indicate in this part the differences between the validity findings in terms of ranking individuals (looking at Spearman correlation coefficient) and the actual measurement of weekly minutes of MVPA and total PA (looking at plots with differences between values of questionnaire and accelerometer). The correlation coefficients show that this adapted questionnaire has a fair-to-moderate criterion validity for MVPA and total PA in terms of ranking individuals according to their level of PA. However, the validity in terms of absolute measurement of weekly time spent in MVPA and total PA seemed limited, in particular for MVPA where there was substantial over-reporting.
Response:
We would like to refer to our response to comment #21, in which we listed the changes made to the discussion on the Spearman rank order correlations and the over-reporting of MVPA that was observed in the plots.

MINOR ESSENTIAL REVISIONS
1. Background, third paragraph, first sentence: please include an explanation of adult populations being investigated in ref 8-10.

Response:
Adult populations were specified by their age range on page 5, line2: (modification underlined, in italics)
“IPAQ-L has been shown to have fair validity and acceptable reliability in a wide range of adult populations (18-65y) (Craig et al., 2003; D. Macfarlane, Chan, & Cerin, 2011; D. J. Macfarlane, Lee, Ho, Chan, & Chan, 2007), but research investigating its applicability to elderly populations remains scarce (Helmerhorst et al., 2012).”

2. Background, third paragraph: please include for the second study in Hong Kong elderly, with which measure the IPAQ-L was compared for determining validity. Was that an accelerometer?

Response:
Indeed, accelerometers were used as the criterion measure in the Hong Kong study of Cerin et al. (2012). This was added to the text on page 5, lines 7-11: (modifications underlined, in italics)
“The second study on measurement properties of IPAQ-L, which was conducted in Hong Kong urban-dwelling older adults, compared both validity (using accelerometers as criterion measure) and reliability (2-week interval) of the Chinese interviewer-administered IPAQ-L across different neighborhood types (varying on socio-economic status and walkability).”

3. Background, fourth paragraph, second sentence suggests that all accelerometers are hip-worn, which is not true. Suggested adaptation to that sentence and the previous one: ‘PA questionnaires are generally validated through examining their agreement with an objective criterion measure of PA, mostly assessed through hip-worn activity monitors or accelerometers [11]. These devices capture human-body accelerations and translate them into activity counts, which in turn give an indication of an individual’s activity degree.’

Response:
We acknowledge that this sentence was poorly formulated and adapted it according to the Reviewer’s suggestion. Following text on page 5, lines 21-23, was modified: (modifications underlined, in italics)
“These devices, mostly worn at hip bone level, capture human-body accelerations and translate them into activity counts, which in turn give an indication of someone’s activity degree.”
4. Background, paragraph 5: Please include a short rationale on why the test-retest reliability of the questionnaire was investigated.

Response:
Following sentence was added to the Background section on page 7, lines 6-8:
“In addition to assessments of an instrument’s criterion validity, it is also important to examine the stability of this instrument over time and whether it is capable of measuring a variable (e.g., PA) with consistency (Portney & Watkins, 2009).”

5. Results, Sample characteristics and PA levels: It would be more informative to include some actual numbers on average time spent on MVPA and total PA in the results. In addition, it would be more informative to report results on MVPA and PA levels in hours/week, rather than min/week. This applies to results section and tables.

Response:
According to the Reviewer’s suggestion, actual numbers on time spent in MVPA and total PA were included in the text. Variables were reported in minutes per week and their equivalent in hours per day was reported as well. Moreover, Table 1 (sample characteristics) was completed with information on the median and interquartile ranges of all PA variables, because they provide a more accurate view on the non-normally distributed data than the mean does.

Adaptations can be specified as follows:
Page 14, lines 17-23: (modifications underlined, in italics)
“Table 1 further shows that for the total sample, the highest levels of weekly minutes of total PA were found for Copeland total PA \(\text{median}=2308.2 \text{ min.week}^{-1}, \text{ equivalent to 5.5 hours.day}^{-1}\)\), whereas the lowest number of total PA minutes were self-reported \(\text{median}=577.5 \text{ min.week}^{-1}, \text{ or 1.4 hours.day}^{-1}; p<0.001\). However, for weekly minutes of MVPA, self-reported activity levels were highest \(\text{median}=540.0 \text{ min.week}^{-1}, \text{ or 1.3 hours.day}^{-1}\), followed by Copeland MVPA \(\text{median}=283.5 \text{ min.week}^{-1}, \text{ or 0.7 hours.day}^{-1}\), and Freedson MVPA \(\text{median}=111.5 \text{ min.week}^{-1}, \text{ or 0.3 hours.day}^{-1}\), respectively \((p<0.001)\).”

Page 15, lines 1-5: (modifications underlined, in italics)
“Regarding domain-specific PA, participants reported to have spent most of their time doing domestic PA \(\text{median}=240.0 \text{ min.week}^{-1}, \text{ or 0.6 hours.day}^{-1}\). Less time was spent doing recreational PA \(\text{median}=68.0 \text{ min.week}^{-1}, \text{ or 0.2 hours.day}^{-1}\), transportation PA \(\text{median}=60.0 \text{ min.week}^{-1}, \text{ or 0.1 hours.day}^{-1}\) and work-related activities \(\text{median}=0.0 \text{ min.week}^{-1}\).”

6. Results, Sample characteristics and PA levels, second paragraph: The authors have added items on recreational cycling to the questionnaire. To assess the appropriateness of this, it is important to know the prevalence of cycling. Please include these in the text/tables.

Response:
Following information was added to page 10, lines 2-4 of the methods to assess the appropriateness of including recreational cycling items:
“Besides, since a study in 48,879 Flemish older adults showed that 53.8% of the participants reported to walk or cycle for recreation at least once a week (Van Cauwenberg et al., 2012), it was considered useful to add an item on recreational cycling to the elderly-adapted IPAQ.”

7. Results, Sample characteristics and PA levels: Please quantify what the authors mean by ‘Characteristics were similar across both samples’, e.g. differences < ....

Response:
The text in lines 9-12 on page 14 was modified in order to clarify similarities and differences between sample sizes of the validity sample (n=434) and the reliability subsample (n=29): (modifications underlined, in italics)

“Mean age, BMI, gender distribution, proportion of former white collar workers and proportion of older adults living with a partner were similar across both samples. In contrast, the samples differed in educational level, with more highly educated people being represented in the reliability subsample.”

8. Results, Criterion validity: It would be informative to include scatterplots on the measurements of the questionnaire and accelerometer data, separately for total PA and MVPA, and for the different cutoffs (Freedson and Copeland). These can be included as supplementary material.

Response:
If the Reviewer agrees, we prefer not to include additional scatterplots on the measurements of the questionnaire and accelerometer data. We think the information now provided in the results of the revised manuscript (descriptive statistics, including median values and interquartile ranges in Table 1; medians and percentiles of difference between questionnaire and accelerometer data in the text) is sufficient to provide an appropriate overview of the data.
In addition, as we already included three supplementary files in the original manuscript, adding more scatterplots may obscure the essential findings of the study.

9. Results, Test-retest reliability: Please the include the mean time between the two interviews for the reliability study.

Response:
Following clarification was added to lines 15-17 on page 8 of the methods: (modifications underlined, in italics)

“For the reliability study, both assessments of IPAQ-L were conducted by the same interviewer and the mean time interval between both visits was 9.6 ± 1.7 days.”

10. Some improvement in writing can be made:
a. Abstract, Conclusions:
‘Validity results suggest that IPAQ-L is more valid to measure older adults’ total PA than MVPA.’ (delete ‘instead of’, add ‘than’)
b. Background, first paragraph, last sentence:
Nevertheless, to make judgments on the prevalence of older adults’ PA, and to identify its most important determinants, adequate monitoring of PA in this age group is essential. (delete ‘prevalance’, add ‘and’)

c. Background, third paragraph, first sentence:
‘...but research investigating its applicability to elderly populations remains scarce [11].’ (add ‘to elderly populations’)

d. Results, Validity for measuring total PA:
‘Because plots for each quartile of Copeland total PA were very similar to these of Freedson total PA, these were not included as a separate figure, but they are shown in Additional File 3. For both Freedson and Copeland total PA, plots show that a decrease in under-reporting is observed between self-reports and accelerometer-derived total PA with increasing magnitude of accelerometer-derived total PA. (delete ‘can be retrieved’)’

e. Discussion, Criterion validity, paragraph 2:
‘On the other hand, plots showed a general pattern of over-reporting, especially for older adults with lower objectively-measured activity levels. As PA levels show a systematic decline with increasing age [30], this may have important consequences for the utility of IPAQ-L in longitudinal studies, and in the oldest old. (add ‘pattern of’, delete ‘and’, and ‘or’; add ‘and’)’

Response:
We corrected all errors pointed out by the Reviewer.

DISCRETIONARY REVISIONS

1. Methods, statistical analyses: For the reliability analyses, it might also be useful to report the Standard Error of Measurement (SEM), as a measure of reliability of individual scores obtained from the questionnaire.

Response:
We thank the Reviewer for this suggestion. However, based on the fact that for each participant, test and retest of the questionnaire were assessed by only and the same interviewer, we think it is sufficient to only include the 95% confidence intervals reported in Table 2. Nevertheless, if the Reviewer insists on including SEM values as well, we are happy to include them after all.

2. Discussion, Test-retest reliability:
In the sentence:
‘Coefficients found in our study are comparable to those reported in other reliability studies on PA questionnaires for older adults [11].’
Please include some examples on coefficients observed, as this might be more easy to compare with the results of this study.

Response:
To improve interpretation of this statement, the text in lines 10-13 (page 20) was adapted: (modifications underlined, in italics)
“Concordant with our hypothesis, coefficients found in our study (i.e., ICC range 0.43-0.81) are comparable to those reported in other reliability studies on PA questionnaires for older adults. Specifically, a literature review observed median reliability ICC estimates of 0.65 in other studies conducted in older adult samples (Helmerhorst et al., 2012).”

In addition, to enhance interpretation of the validity results in the discussion, we also adapted the text in lines 6-10 on page 17: (modifications underlined, in italics)

“As hypothesized, similar correlation coefficients for MVPA as observed in the present study were found in other validation studies using Spearman correlations between accelerometers and PA questionnaires in older adult populations (ρ=0.43 (Dinger et al., 2004); ρ=0.37 (Hurtig-Wennlöf et al., 2010); ρ=0.31 (Kolbe-Alexander et al., 2006)). Moreover, a review on measurement properties of PA questionnaires described that validity of questionnaires in older adult samples showed a median Spearman correlation of ρ=0.41 (Helmerhorst et al., 2012).”

3. Discussion, Criterion validity, paragraph three: in the sentence ‘This under-reporting of specific LPA is due to the main focus of IPAQ-L, i.e., measuring population levels of MVPA, with few items specifically assessing LPA,’ please specify which LPA items are included. In addition, in the sentence ‘Although our version of IPAQ-L already contains elderly-specific adaptations, specific questions on LPA are still lacking.’ Include which items the authors think are lacking. The reader can then understand which items are already included, and which the authors think should be included.

Response:
Because the sentence “This under-reporting of specific LPA is due to the main focus of IPAQ-L, i.e., measuring population levels of MVPA, with few items specifically assessing LPA” has been removed from the discussion, the first part of Comment#3 is no longer applicable.

Regarding the second part of the comment, we agree with the Reviewer that the paper would benefit from including a description on which LPA items could be lacking in IPAQ. Hence, examples of LPA items were included on page 19, lines 7-17. The content of these items were based on the 2011 Compendium of physical activities (Ainsworth et al., 2011); https://sites.google.com/site/compendiumofphysicalactivities/Activity-Categories. Moreover, we discussed the benefits of including specific LPA items next to items on MPA.

Following text was added: (modifications underlined, in italics)

“Although our version of IPAQ-L already contains elderly-specific adaptations, specific questions on LPA (corresponding to activities with a Metabolic Equivalent (MET) value < 3) are still lacking. According to the 2011 Compendium of Physical Activities (Ainsworth et al., 2011), items on LPA could specifically ask for home activities such as “dusting or polishing furniture” (~2.3 METs), “washing dishes, clearing dishes from the table” (~2.5 METs), or “cleaning, sweeping, light effort” (2.3 METs), as these activities are likely to be reported as moderate-intensity instead of light-intensity. Examples of non-home based light-intensity activities could include “food shopping with/without a grocery cart while standing/walking” (~2.3 METs). If IPAQ-L would be interviewer-administered, the interviewer
could preserve possible duplicate over-reporting, by prompting participants when they tend to report the same activity twice (i.e., once in the LPA response, and once in the MPA response)."

Moreover, following examples were added to lines 2-6 on page 20: (modifications underlined, in italics)

“This under-reporting of specific LPA could be due to the fact that accelerometer total PA also comprises measures of the very low-intensity activities (e.g., standing upright; doing light work in the kitchen or home office; getting dressed; grooming), while IPAQ-L does not. Including items on such daily activities of very light intensity might further enhance validity of IPAQ-L towards total PA.”

4. Methods, Measures, Socio-demographics and physical measures: Please include examples of ‘blue collar’ and ‘white collar’ occupations, as these terms might not be familiar to all readers.

Response:
According to the Reviewer’s suggestion, we added some clarifications on the “blue collar” and “white collar” occupation categories on page 8, lines 21-24 and on page 9, line 1: (adaptations underlined, in italics)

“Participants self-reported their age, current living situation (responses dichotomized into “having a partner” and “having no partner”), educational level (responses dichotomized into “tertiary education” and “non-tertiary education”) and former occupational status (responses categorized into “household”, “blue collar” including workman and self-employed, and “white collar” including education/teaching; employee; executive staff member and profession).

5. Methods, Measures, Self-reported PA: adapted IPAQ-L interview version: it is not clear what the authors mean by ‘real employment’, paid employment? Please replace by an appropriate term.

Response:
On page 9, line 7, “real employment” was replaced with “paid employment”, according to the Reviewer’s remark.
REVIEWER #2

GENERAL COMMENT TO THE REVIEWER:
Based on the comments of Reviewer #1, parts of the background, results and discussion have been modified. Specifically, a priori hypotheses were added to the Background, median values were added to the Results and ICC’s on absolute agreement (rather than the originally reported ICC’s on consistency) were reported. Moreover, a substantial part of the discussion was adapted. Now, we tried to state more clearly that there is a difference between validity findings in terms of ranking individuals (Spearman correlation coefficients) and the actual measurement of weekly minutes of MVPA/total PA (plots with differences between values of questionnaires and accelerometers). The correlation coefficients show that this adapted questionnaire has moderate-to-good criterion validity for MVPA and total PA in terms of ranking individuals according to their level of PA. However, the validity in terms of absolute measurement of weekly time spent in MVPA and total PA seemed more limited, in particular for MVPA where there was substantial over-reporting. This reasoning was not very clear in the original manuscript, but now it has been integrated in the discussion. In addition, as both Reviewer #1 and #2 seemed to question the reasoning on bmi (see also our reponse on comment #6), this was deleted from the manuscript.

We would like to ask the Reviewer to check the modifications made to the manuscript (the discussion in particular) and hope these adaptations are acceptable.

REVIEWER’S REPORT:

This study addresses an interesting subject area the validity of a widely used physical activity questionnaire among a population of older adults. Given that older adults have the lowest physical activity levels of any age group and that it has proved hard to increase physical activity in this age group, studies which aid us to understand how to accurately quantify activity levels are a very useful for situations where objective measurements of physical activity are not possible. This paper therefore has the potential to make a very useful contribution to monitoring physical activity in this highly inactive sub-population.

MAJOR COMPULSORY REVISIONS
1. The paper is well written on the whole. I think it would be useful to clarify in the title, abstract and introduction that the authors are testing a modified version of the IPAQ which combines vigorous activity with moderate, adds in gait speed and adds in cycling, - this is essential when considering the wider applicability of these results.

Response:
In concordance with the Reviewer’s suggestion, the title of the paper was modified in order to clarify that we assessed an elderly-adapted version of IPAQ-L.
The original title:
“Assessment of physical activity in older adults: validity and reliability of the International Physical Activity Questionnaire (IPAQ), long interview version”

was changed into:
“Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L)”.

Furthermore, more information on the specific adaptations was added to the abstract on page 2, lines 9-12: (modifications underlined, in italics)

“Participants (n=434) completed the last seven days version of IPAQ-L, modified for the Belgian population of community-dwelling older adults. This elderly-adapted version of IPAQ-L combined vigorous and moderate activities, and questions on gait speed and recreational cycling were added. Furthermore, participants wore an ActiGraph GT3X(+) accelerometer for at least five days.”

2. Page 6 is the response rate really 508/1135, surely it should be 508/1750 as these were the people who were originally selected for the study.

Response:
We agree with the Reviewer’s concern and also took the other response rate calculation under consideration.
However, we could not verify whether all older adults that were not at home when the interviewers visited them, also still lived at the address provided by the public service of Ghent. Moreover, if they did live at the correct address, it would still not be possible to make inferences on whether or not the selected person would be willing to participate in the study or whether or not this person would be physically capable for participation in the study (i.e., able to walk 100 meters without severe physical difficulties). Therefore, we chose to calculate response rates based on those who could be contacted in person by the interviewer, and not only through postal mail.

3. Page nine line 5, did the authors test different windows of continuous zeros? Is 60 minutes long enough for this population? Would it be appropriate to consider longer sedentary bouts eg a TV program might last one hour, and it is plausible that people in this age group might sit for more than one hour at a time.

Response:
We thank the Reviewer for noting this and acknowledge that we made a typing error. In reality, periods of 90 minutes of consecutive zeros (instead of 60 minutes) were classified as “non-wearing”. Classifying ≥90 minutes of consecutive zeros as “non-wearing” in older adults is recommended in the study of Choi et al. (2012).

We rectified the typing error in the manuscript and also provided the reference of Choi et al. (2012) in the text on page 11, lines 22-23: (modifications underlined, in italics)

“Periods covering ≥ 90 minutes of consecutive zeros were defined as “non-wearing”, as recommended by Choi et al. (2012)”
4. Page nine line 9-10, when making the dichotomy, did the authors insist on >=150 mins/week in bouts of >=10 minutes or just >=150 mins/week in total? Please clarify.

Response:
When the dichotomy was made to define Public Health Recommendations for accelerometer MVPA measures, we did not only include ≥150 weekly minutes in bouts of ≥10 minutes, but all recorded MVPA ≥150 weekly minutes.

To clarify this, following text was added to lines 2-4 on page 12: (modifications underlined, in italics)

“In addition, both accelerometer-derived MVPA variables were dichotomized according to the Public Health Recommendations for PA (all MVPA records <150 min.wk⁻¹ = 0; all MVPA records ≥150 min.wk⁻¹ = 1).”

Furthermore, we would like to refer to our response on comment #6, in which we explain the reasons why we calculated accelerometer minutes/week, rather than number of MVPA bouts lasting ≥10minutes.

5. Page nine line 12, the two definition of MVPA based on Copeland and Freedson is entirely appropriate as it is explained that Freedson was developed for middle age and is widely used but probably not as appropriate as Copeland which is developed for older adults. However I do not understand why the cut point for SB would vary according to whether Copeland or Freedson is used to define MVPA- I have not seen this approach taken anywhere else. Surely we are trying to make the best possible estimate of the behaviour, I cannot see why SB criteria should be dependent on the MVPA criteria. I would favour doing some work to evaluate whether 50 or 100 is the better cut point in this dataset and then using one or other. 100 is commonly used in studies of older adults.

Response:
We understand the Reviewer’s concern that the SB cutpoint cannot be dependent of the Freedson and Copeland cut points defining MVPA.

However, these “Freedson vs. Copeland & Esliger cut points” on page 9, line 12 of the original manuscript (see Reviewer’s comment) we referred to are not the MVPA cut points described in the paragraph preceding it, but they are the original cut points defining SB, defined in the research papers of Freedson et al. and Copeland & Esliger, respectively. What we actually wanted to report was that, similar to the difference in Freedson MVPA and Copeland MVPA cut points, sedentary behaviors are also defined by different cut points according to Freedson et al. (1998) and according to Copeland & Esliger (2009). Specifically, Freedson et al. define sedentary behavior as all accelerometer data ≤ 100 counts.min⁻¹. Copeland & Esliger define sedentary behavior as all accelerometer data ≤ 50 counts.min⁻¹.

We acknowledge that defining “Freedson MVPA” and “Copeland MVPA” in the first paragraph on page 11 (lines 1-4), may have caused confusion reading the “Freedson vs. Copeland & Esliger cut points” on page 9, line 12 of the original manuscript (see Reviewer’s comment).
Therefore, to avoid any confusion for the readers, we modified the text on page 12, lines 6-12 and the original references to elucidate our SB (and total PA) cut point definition: (modifications underlined, in italics)

“Similar to the difference in Freedson MVPA and Copeland MVPA cut points, sedentary behaviors are also defined by different cut points according to Freedson et al. (Freedson et al., 1998) and according to Copeland & Esliger (Copeland & Esliger, 2009). Depending on whether Freedson vs. Copeland & Esliger cut points were used, sedentary behaviors were defined as all accelerometer data of ≤ 100 counts.min\(^{-1}\) or ≤ 50 counts.min\(^{-1}\), respectively. Total accelerometer PA was calculated as the sum of all counts.min\(^{-1}\) above these thresholds and will be referred to as “Freedson total PA” (>100 counts.min\(^{-1}\)) and “Copeland total PA” (>50 counts.min\(^{-1}\)) in the further sections of this paper.

6. Page 13, line 15, why is it questionable given the prevalence of overweight? Isn’t it likely that the authors have not calculate meeting 150 min/week in bouts of >=10 minutes - if this was done (the questionnaire specifies that only activities lasting >10 minutes are reported), then I suspect that the estimated of meeting guidelines based on accelerometer data would be vastly different and the conclusions in this section might also change.

Response:
We agree with the Reviewer’s rationale on including bouts, but we did not use the number of accelerometer bouts as an additional criterion measure because of following reason:
Cerin et al. (2012) did calculate estimates of accelerometer activity bouts (next to total weekly minutes of accelerometer PA) in their validity study, but the prevalence of such bouts was low in older adults. We checked the average bout number in the present study, and prevalence was also very low for both Copeland MVPA and Freedson MVPA.
This might be probably due to the fact that activities lasting ≥10 consecutive minutes are very likely to be reported when the questionnaire is assessed, but in reality, such activities may be interrupted by small breaks (e.g., waiting at a crosswalk during a bout of walking). Consequently the number of activity bouts may be reduced substantially and the accumulation of bouts only may underestimate actual PA levels of older adults.
In addition, the majority of studies using estimates of accelerometer MVPA in older adults do not use the number of bouts as an outcome measure for objective MVPA, but generally report weekly minutes (e.g., Carlson et al., 2012; Cerin et al., 2011; King et al., 2011), or daily minutes (e.g., Ding et al., 2014).

Nevertheless, we acknowledge the Reviewer’s comment that when accelerometer MVPA bouts would be estimated instead of MVPA minutes/week, a substantially lower number of participants would be classified by the Copeland MVPA cut point as reaching the Public Health Recommendations for PA. To come towards this remark, we added the lack of including bouts to the limitations of the present study:

Page 21, lines 3-6:
“Another issue related to the accelerometer measure could be that we used weekly minutes of accelerometer MVPA to assess validity, instead of MVPA bouts lasting ≥10 consecutive minutes. This might have biased our estimate of objective MVPA and may have affected validity results.”
Moreover, we also acknowledge that the reasoning on the prevalence of overweight was a questionable and speculative part of the discussion. Since both Reviewer #1 and Reviewer #2 made this remark, this part of the discussion was modified and the part on BMI prevalence was removed. In addition, as it was beyond the scope of the present study to make inferences on the applicability of the cut points of Freedson vs. Copeland, the comparison between both was softened in the discussion.

Specifically, lines 10-24 on page 17 and lines 1-13 on page 18 were adapted as follows: (modifications underlined, in italics)

“These moderate-to-good Spearman coefficients should be interpreted with care, since they can only suggest that IPAQ-L was relatively good at ranking individuals by their reported MVPA levels, when compared to their objectively-measured MVPA levels. In order to evaluate validity of the IPAQ-L, it is recommended that these Spearman rank order correlations are interpreted in combination with the results of the agreement plots. Plots for MVPA showed a general pattern of over-reporting of IPAQ-L and this over-reporting was higher when self-reported MVPA was compared with the Freedson MVPA, than when it was compared to the Copeland MVPA. This finding is in concordance with our hypothesis that better validity of the elderly-adapted IPAQ-L would be observed regarding Copeland MVPA, which seems a promising finding. On the other hand, it is also logical to find less over-reporting when self-reports are compared to Copeland MVPA, because the cut point has a lower threshold for defining MVPA than the Freedson MVPA cut point. However, also the Copeland MVPA cut point might not be the ideal solution for defining accelerometer MVPA in older adults. Percentages of participants in the current study reaching the Public Health Recommendations for MVPA were 27.4% and 71.2% for Freedson and Copeland, respectively. Given that worldwide, only 30-40% of older adults reach the Public Health Recommendations for PA (Centers for Disease Control and Prevention, 2013; Eurobarometer, 2010; Tafforeau, 2008), the percentage found in the present study using the Copeland MVPA threshold seems rather high and Freedson MVPA levels could be a more realistic estimate of participants’ actual MVPA levels. Although it is beyond the purpose of the present study to make inferences on the applicability of cut points for defining MVPA in older adults, our findings do indicate that researchers should be cautious when selecting an appropriate accelerometer threshold for defining MVPA in older adults. Moreover, it is suggested that more research is conducted on the feasibility of different MVPA cut points in older populations.

Irrespective of the cut point applied, however, the plots for validity of IPAQ-L regarding MVPA showed that the highest over-reporting was found in older adults with lower objectively-measured activity levels (i.e., the lower quartiles shown in the plots). As PA levels show a systematic decline with increasing age (Schrack et al., 2014), this may have important consequences for the utility of IPAQ-L in longitudinal studies, and in the oldest old.”

7. P 16 line 21- did you test for interviewer effects- did some interviewers prompt the participants more than others and consistently end up with higher self-reported PA levels than other interviewers? If so, this should be controlled in the analyses.

Response:
We were not able to statistically assess inter-rater differences because each participant was only visited by a single interviewer. However, several precautions were taken to minimize potential inter-rater bias. These attempts are explained below:

Firstly, the interviewer questionnaire itself contains interviewer clarifications. Examples of the interviewer version can be found at https://sites.google.com/site/theipaq/questionnaire_links (online English telephone version of IPAQ-L). We followed IPAQ forward- and backward translation guidelines for translating our elderly-adapted version of IPAQ-L into Dutch (see page 10, lines 6-8) and hence, we think these translated interviewer prompts and clarifications were sufficiently appropriate to be used by each of our interviewers in this study.

Secondly, before initiating data collection, the first and fourth author (VVH and JVC) conducted some (n=5) interviews together, to ensure that the questionnaire was assessed similarly by different raters. All other interviewers (n=10) received a standardized interviewer training for assessing IPAQ-L, led by VVH or JVC (referred to as the “trainer” in the next part of this response):

The training consisted of three consecutive steps: 1) Firstly, trainees attended a session where the trainer contacted a participant at his/her home residence (going door by door), explained the procedures of the home visit, instructed participants how to wear the accelerometer, and assessed the interview. 2) During a second session, trainees contacted a new potential participant at his/her home residence and assessed an interview themselves, assisted by the trainer. 3) Lastly, a third participant was contacted and a third interview was conducted by the trainee. This last session was attended by the trainer, who corrected the trainee when this was needed.

We acknowledge that more elucidation on the training of the interviewers should be provided. Therefore, following text was added on page 8, lines 8-12:

“To minimize inter-rater bias, all interviewers received a standardized training to adequately conduct the complete home visit procedure (i.e., contacting selected older adults; explaining procedures concerning the interview and accelerometer data collection; assessment of the interviewer-administered questionnaire).”

MINOR REVISIONS
8. Page 6 line 11, please state if the monitor was worn during waking hours or for 24 hours?

Response:
Following text was added to lines 6-7 on page 8: (added text underlined, in italics)

“Furthermore, participants were instructed how to wear an Actigraph GT3X(+) accelerometer for the next consecutive seven days, during waking hours excluding contact sports, bathing or swimming activities.”
9. Page eight line 20, 60" epochs were used, presumably this means 60 seconds. Please use the SI unit s for seconds to avoid confusion.

Response:
According to the Reviewer’s suggestion, we changed “ into s on page 11, line 13.

10. Page eight line 21, the reference to the actigraph website is very vague, please either provide a specific url with a link to the exact page, or better still, a link to a published paper.

Response: The reference was substituted with following reference to the web page of the online manual for accelerometer screening and scoring manual (first author: Kelli Cain):


11. Footnote for table 1 should clarify that total PA is eg 100-1952 or 50-1040 cpm

Response:
We thank the Reviewer for noting this and added following lines to the footnote for Table 1 on page 29, clarifying accelerometer count ranges defining Freedson and Copeland total PA, respectively:

“d Accelerometer-derived total PA > 100 counts.min⁻¹”
“e Accelerometer-derived total PA > 50 counts.min⁻¹”

12. Page 12, line 19- is psychometrics really the right term here?

Response:
The term “psychometrics” was replaced by “measurement properties” throughout.

13. Page 13, line 13 spell out PHR

Response:
“PHR’s” was modified to “the Public Health Recommendations”.
