Reviewer's report

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

Version: 2 Date: 6 January 2015

Reviewer: Eline H. van Roekel

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 revisions

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)’.

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

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.

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.

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.

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.

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.
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.

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 2460 minutes). Was this according to recommendations? If so, include a reference. Otherwise, clearly state the reason why specific cutoffs were chosen.

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?

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)

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

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?

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?

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.

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.).

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).

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).

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.

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.

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.

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.

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.

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).

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.

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.

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).

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.

Minor essential revisions

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

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?

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.’
4. Background, paragraph 5:
Please include a short rationale on why the test-retest reliability of the questionnaire was investigated.

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.

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.

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

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.

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

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 ‘prevalence’, 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')

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.

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.

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.

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.

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.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, but I do not feel adequately qualified to assess the
statistics.

**Declaration of competing interests:**

I declare that I have no competing interests.