## PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Reliability, measurement error and minimum detectable change in mobility measures: a cohort study of community-dwelling adults aged 50 years and over in Ireland |
| **AUTHORS** | Donoghue, Orna; Savva, George; Börsch-Supan, Axel; Kenny, RoseAnne |

## VERSION 1 – REVIEW

| | |
|---|---|
| **REVIEWER** | Daniel L. Young<br>University of Nevada, Las Vegas. Las Vegas, USA |
| **REVIEW RETURNED** | 05-Apr-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | This is a well written paper with excellent statistical methods used to evaluate reliability, SEM and MDC for 3 measures of physical function, TUG, RCS, and gait speed (under 3 conditions) among community dwelling seniors. My concern with this paper is that it offers very little to our body of knowledge on these topics and has a serious design flaw. I will here address these 2 concerns in order. The reliability of these measures is well tested as the authors describe in their introduction. A gap for the SEM and MDC of the RCS is highlighted; however, these are simple calculations since the reliability has been previously reported. A gap in dual-task gait speed is highlighted as the reliability is likely dependent on the specific task, and while true this study does not help offer a standard for the task and further contributes to the variety of narrowly applicable results in publication. The one more novel element, that of the potential effect time of day may have on reliability is not enough to overcome the real concern which I present last. The time between assessments of 1-4 months. I was not persuaded by the statement about the stability of performance in these tasks for a relatively healthy sample over this time period (page 14, 1st paragraph). Assessment of the reliability of these measures, and all the calculations thereafter performed, assume that the underlying construct being measured between assessments has not changed. Most of the sample was measured more than 3 months apart (mean of 88 days; page 10, line 17). This introduces too much potential for change in the physical function of the subjects and is a serious threat to the validity of the results. Therefore, I cannot recommend publication of this study. |

| | |
|---|---|
| **REVIEWER** | Julie Richardson<br>McMaster University, Canada |
| **REVIEW RETURNED** | 03-May-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | This is an excellent paper in that it is well executed from a methodological standpoint and well written and clear. It is |

| | |
|---|---|
| | important that these measurement papers are undertaken from these large population based studies as these are not routinely undertaken. It would be useful to note in the background of the large national studies how many have actually published psychometric properties of the outcomes they are using. This is important in terms of understanding population based results and cannot be derived from clinical samples.<br>I am not clear from the description of the sample with the information given about SHARE and TILDA. The sample is from the IRISH sample as part of SHARE but why did this study link with TILDA are these measures that are undertaken as part of TILDA or SHARE. This could be made clearer.<br>There are two issues that I think warrant revision. The first is minor just noting some reference about how you classified the level of reliability. The second is around the training and the standardisation of the protocol used by the raters because of the importance of this issue but also because there was what seems like a systematic difference between Rater1 and Rater2. The authors do state that the raters had 3 years experience. It would also good to have some comment about how this might be addressed in studies going forwards, for example testing 10 patients and looking at systematic differences. The measurement concepts are very well laid out and explained to the reader.<br>The information provided in the tables about the missing data is also very important, however there are no details provided about what the authors did about missing data in terms of examining the effects of it or imputing for it. I presume they just reported it but would be good to clarify. The number of the ethics consent should be noted. |

| | |
|---|---|
| **REVIEWER** | Juliessa Pavon<br>Duke University, United States |
| **REVIEW RETURNED** | 18-Jun-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | This concise and well written article reports on the test-retest reliability of important clinical and research measurements of mobility and lower limb function of older adults. Study is conceptually interesting and provides clinically relevant information about these measures. Protocols for each test were well described as were the reliability measures used for this study.<br><br>Major Issue:<br>-The time-frame of 1-4 months for repeat measurement needs to be defended. Why this time frame vs. 6 months or 1 year? There is also variation that can exist between 1 vs 4 months. Should we expect any difference between repeat performance at 1 month vs. 4 months? was that tested?<br><br>Results Pg 10, line 21-23 recommend including some information about normative values for these tests to help readers interpret if these baseline performance score represent health function/mobility - e.g. TUG score of 8s indicates a functionally robust population of older adults, so would highlight that for readers, but mean RCS of 12 indicates population may be slightly less robust. How would authors describe/classify your population functionally? Clarifying this is important as the authors allude to this in the Discussion, pg 13, Lines 15-17.<br><br>Minor revisions: |

| | |
|---|---|
| | Intro- Pg 5 line 10, suggest adding one more sentence to this paragraph to emphasize what gap in knowledge is begin filled by this study. Is it that the long-term reliability of these measures is unknown? From this paragraph it appears that we already know a lot about TUG - how will this study add new information?<br>Results pg 11, line 9-10, is this supposed to read 24%? (not 27%?)<br>Conclusions- how is the time-frame of 1-4 months applicable to clinical practice or research? Related to above, why was this timeframe selected? Also for consideration in the discussion is clarification on the rationale for why it is important to know the reliability of these measures for this particular timeframe.<br>Tables - try to fit results for each measure; mean (SD) on one line, consider using only 1 decimal.<br>References: suggest adding a few more current references > year 2015. |

## VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Daniel L. Young
Institution and Country: University of Nevada, Las Vegas. Las Vegas, USA Please state any competing interests or state ‘None declared’: none declared

Please leave your comments for the authors below This is a well written paper with excellent statistical methods used to evaluate reliability, SEM and MDC for 3 measures of physical function, TUG, RCS, and gait speed (under 3 conditions) among community dwelling seniors. My concern with this paper is that it offers very little to our body of knowledge on these topics and has a serious design flaw. I will here address these 2 concerns in order. The reliability of these measures is well tested as the authors describe in their introduction. A gap for the SEM and MDC of the RCS is highlighted; however, these are simple calculations since the reliability has been previously reported. A gap in dual-task gait speed is highlighted as the reliability is likely dependent on the specific task, and while true this study does not help offer a standard for the task and further contributes to the variety of narrowly applicable results in publication.
Existing research looks at repeat assessments conducted up to two weeks apart, therefore the novel aspect of this study is the presentation of repeatability data obtained over a longer follow-up period than usual. This has been elaborated on in both the introduction and the discussion. As mentioned here, other gaps in relation to RCS and gait speed measures are already included in the text.

The one more novel element, that of the potential effect time of day may have on reliability is not enough to overcome the real concern which I present last. The time between assessments of 1-4 months. I was not persuaded by the statement about the stability of performance in these tasks for a relatively healthy sample over this time period (page 14, 1st paragraph). Assessment of the reliability of these measures, and all the calculations thereafter performed, assume that the underlying construct being measured between assessments has not changed. Most of the sample was measured more than 3 months apart (mean of 88 days; page 10, line 17). This introduces too much potential for change in the physical function of the subjects and is a serious threat to the validity of the results. Therefore, I cannot recommend publication of this study.
We acknowledge that the follow-up period is longer than that presented in previous papers, however this is the purpose of this analysis and has relevance for both clinical and research purposes. For

example, these estimates should be used when assessing individual changes in mobility performance over similar time-scales e.g. when examining the effects of an intervention or patient progression, when calculating required sample sizes for studies using these outcomes or when applying methods to adjust for measurement error in epidemiological studies. We do acknowledge that acute changes in health and performance can occur even with shorter follow-up, however our relatively healthy sample is unlikely to demonstrate a consistent, genuine change in performance in the time period examined.

Reviewer: 2
Reviewer Name: Julie Richardson
Institution and Country: McMaster University, Canada Please state any competing interests or state 'None declared': None Declared

Please leave your comments for the authors below
This is an excellent paper in that it is well executed from a methodological standpoint and well written and clear. It is important that these measurement papers are undertaken from these large population based studies as these are not routinely undertaken. It would be useful to note in the background of the large national studies how many have actually published psychometric properties of the outcomes they are using. This is important in terms of understanding population based results and cannot be derived from clinical samples.

I am not clear from the description of the sample with the information given about SHARE and TILDA. The sample is from the IRISH sample as part of SHARE but why did this study link with TILDA are these measures that are undertaken as part of TILDA or SHARE. This could be made clearer.
Further details about this have been provided in the Methods. This should clarify that participants were recruited from SHARE-Ireland, the health assessment included the same tests and used the same protocols as the TILDA health assessment; and it was conducted by TILDA research nurses in the TILDA assessment centre.

There are two issues that I think warrant revision. The first is minor just noting some reference about how you classified the level of reliability.

A reference for classifying reliability has now been added to the Methods (Koo & Li, 2016).

The second is around the training and the standardisation of the protocol used by the raters because of the importance of this issue but also because there was what seems like a systematic difference between Rater1 and Rater2. The authors do state that the raters had 3 years experience. It would also good to have some comment about how this might be addressed in studies going forwards, for example testing 10 patients and looking at systematic differences. The measurement concepts are very well laid out and explained to the reader.
We have added some additional detail about the training duration and quality control (in Methods). We have also elaborated on how the difference between nurses could be addressed in future studies (in Discussion).

The information provided in the tables about the missing data is also very important, however there are no details provided about what the authors did about missing data in terms of examining the effects of it or imputing for it. I presume they just reported it but would be good to clarify.
Mixed effects models use all available data, therefore missing data is implicitly assumed to be missing at random. Missing data was not imputed; this is now indicated in the methods.

The number of the ethics consent should be noted.
The ethics committee did not provide a number corresponding to this application.

Reviewer: 3

Reviewer Name: Juliessa Pavon
Institution and Country: Duke University, United States Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below
This concise and well written article reports on the test-retest reliability of important clinical and research measurements of mobility and lower limb function of older adults. Study is conceptually interesting and provides clinically relevant information about these measures. Protocols for each test were well described as were the reliability measures used for this study.

Major Issue:
-The time-frame of 1-4 months for repeat measurement needs to be defended. Why this time frame vs. 6 months or 1 year? There is also variation that can exist between 1 vs 4 months. Should we expect any difference between repeat performance at 1 month vs. 4 months? was that tested?
We acknowledge that the follow-up period is longer than that presented in previous papers, however this is the purpose of this analysis and has relevance for both clinical and research purposes. For example, these estimates should be used when assessing individual changes in mobility performance over similar time-scales e.g. when examining the effects of an intervention or patient progression, when calculating required sample sizes for studies using these outcomes or when applying methods to adjust for measurement error in epidemiological studies. We do acknowledge that acute changes in health and performance can occur even with shorter follow-up, however our relatively healthy sample is unlikely to demonstrate a consistent, genuine change in performance in the time period examined.

Results Pg 10, line 21-23 recommend including some information about normative values for these tests to help readers interpret if these baseline performance score represent health function/mobility - e.g. TUG score of 8s indicates a functionally robust population of older adults, so would highlight that for readers, but mean RCS of 12 indicates population may be slightly less robust. How would authors describe/classify your population functionally? Clarifying this is important as the authors allude to this in the Discussion, pg 13, Lines 15-17.
Level of function relative to published norms has now been addressed in the results.

Minor revisions:
Intro- Pg 5 line 10, suggest adding one more sentence to this paragraph to emphasize what gap in knowledge is begin filled by this study. Is it that the long-term reliability of these measures is unknown? From this paragraph it appears that we already know a lot about TUG - how will this study add new information?
We agree that there is quite a lot of information available for TUG, however there is less data available for RCS and gait speed as outlined in this section. The novel aspect of this study is the measurement of longer-term reliability for all measures. This has been included in the final paragraph of this section, along with an extended rationale for the selection of the follow-up period.

Results pg 11, line 9-10, is this supposed to read 24%? (not 27%?)
27% is the correct figure here (the inverse logarithm of 0.24 is 1.27). This has been indicated in the text for clarity.

Conclusions- how is the time-frame of 1-4 months applicable to clinical practice or research? Related to above, why was this timeframe selected? Also for consideration in the discussion is clarification on the rationale for why it is important to know the reliability of these measures for this particular timeframe.
See response to first point above. The rationale for inclusion of this follow-up period has been included at the end of the Introduction. The potential applications of this work have been further elaborated on in the Discussion.

Tables - try to fit results for each measure; mean (SD) on one line, consider using only 1 decimal.

We have changed the orientation to landscape so that all information can be captured on one line.

References: suggest adding a few more current references > year 2015.
We have searched for more recent references, however these papers refer to younger adults, patient populations (e.g. chronic stroke) and/or or use different methods of data collection (e.g. gait speed measured on a treadmill rather than on GAITRite mat). As these are not directly comparable to our data, they have not been referenced.

## VERSION 2 – REVIEW

| | |
|---|---|
| **REVIEWER** | Daniel Young<br>University of Nevada, Las Vegas USA |
| **REVIEW RETURNED** | 31-Jul-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | I appreciate the authors efforts to address the comments from myself and the other reviewers; however, I may not have been clear about my major concern. As previously stated, my major concern is with the time between test-retest assessments of 1-4 months being used to calculate reliability. Reliability calculations assume that the underlying construct being measured between assessments has not changed. In this study 3 measures of physical function are being evaluated for their reliability. The reliability calculation requires that physical function (the underlying construct) cannot change between the 2 time points at which the measures were taken. As long as this assumption is true then the reliability calculation tells us about the error in the measure itself and/or the variability in the raters as an inherent part of measurement error. If the underlying construct changes then the reliability calculation includes that change and is artificially lower, no longer reflecting measurement error alone (what we want) but the combination of measurement error and changed function. Having stability of physical function (for the 3 measures used) over the time period between assessments is essential to having a valid reliability estimate. I do not think it is sufficient for the authors to say that their “relatively healthy sample is unlikely to demonstrate a consistent, genuine change in performance in the time period examined.” Even the data in table 1 shows that the mean value for RCS was not the same across their sample between the 2 time periods. A more helpful and compelling piece of evidence that function hadn’t changed would be a scatter plot of each individual’s score with the 2 different time periods as X and Y axis. Subjects with stable measurements would lie close to the diagonal and subjects whose function had changed would not. The authors could also provide citations from studies of similar people whose physical function had been shown stable over 1-4 months. The authors response to the review states that “these estimates should be used when assessing individual changes in mobility performance over similar time-scales.” I disagree, reliability estimates are not an appropriate way to account for change in the underlying construct. |

| | |
|---|---|
| **REVIEWER** | Dr Julie Richardson<br>McMaster University, Canada |

|  |  |
|---|---|
| **REVIEW RETURNED** | 06-Aug-2019 |

|  |  |
|---|---|
| **GENERAL COMMENTS** | Excellent contribution to the literature. No further revisions required |

|  |  |
|---|---|
| **REVIEWER** | Juliessa Pavon<br>Assistant Professor, Geriatrics<br>Duke University, USA |
| **REVIEW RETURNED** | 04-Aug-2019 |

|  |  |
|---|---|
| **GENERAL COMMENTS** | Reviewer comments have been sufficiently addressed. |

## VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Daniel Young

Institution and Country: University of Nevada, Las Vegas USA

Please state any competing interests or state 'None declared': none declared

Please leave your comments for the authors below

I appreciate the authors efforts to address the comments from myself and the other reviewers; however, I may not have been clear about my major concern. As previously stated, my major concern is with the time between test-retest assessments of 1-4 months being used to calculate reliability. Reliability calculations assume that the underlying construct being measured between assessments has not changed. In this study 3 measures of physical function are being evaluated for their reliability. The reliability calculation requires that physical function (the underlying construct) cannot change between the 2 time points at which the measures were taken. As long as this assumption is true then the reliability calculation tells us about the error in the measure itself and/or the variability in the raters as an inherent part of measurement error. If the underlying construct changes then the reliability calculation includes that change and is artificially lower, no longer reflecting measurement error alone (what we want) but the combination of measurement error and changed function. Having stability of physical function (for the 3 measures used) over the time period between assessments is essential to having a valid reliability estimate. I do not think it is sufficient for the authors to say that their "relatively healthy sample is unlikely to demonstrate a consistent, genuine change in performance in the time period examined." Even the data in table 1 shows that the mean value for RCS was not the same across their sample between the 2 time periods. A more helpful and compelling piece of evidence that function hadn't changed would be a scatter plot of each individual's score with the 2 different time periods as X and Y axis. Subjects with stable measurements would lie close to the diagonal and subjects whose function had changed would not. The authors could also provide citations from studies of similar people whose physical function had been shown stable over 1-4 months. The authors response to the review states that "these estimates should be used when assessing individual changes in mobility performance over similar time-scales." I disagree, reliability estimates are not an appropriate way to account for change in the underlying construct.

Thank you for elaborating on your concern as this has led us to improve our manuscript. We understand your concern, and appreciate that we did not do enough to explain to readers our intention with using this lag period, or to reassure that there was no significant change in the underlying mobility of our participants during the time between assessments. To address this, we have made several specific additions to the manuscript.

First, we have strengthened the discussion to highlight the importance of understanding variation over this time period both for epidemiologic research, for interpreting changes in individuals and for planning intervention studies.  As you rightly say, it is not pure measurement error of the instruments themselves that we are measuring but this error combined with day-to-day variation in participant function that is attributable to the patient; this is not of interest when we are testing for true underlying changes in mobility.

Second, to address the concern that mobility genuinely changes over the relatively long period between assessments, which would lead to an inflated within-person error, we have plotted within-person errors against delay between assessments and showed that overall there is no effect of the length of the delay on the observed differences. (While there is a small increase for RCS, there is a corresponding decrease in TUG and no change at all for gait speed measures.) This strongly suggests that the differences we are seeing are not driven by underlying changes in state, but are due to natural variation in performance of these tests. Hopefully this mitigates concerns about the length of time between the tests being a significant factor in the variance that we observed.

As requested, we have also included scatter plots of the initial vs repeat performances. These show a distribution of variation with some observations close to the diagonals while others are not. This is expected as some observations may have larger measurement error associated with them than others, so these simply reflect the between- and within-participant standard deviations that we reported elsewhere.

Finally, with respect to the final statement regarding interpretation of individual changes, it is appropriate to compare observed changes to the expected within-participant standard error of measurement, as this the established basis for calculating minimum detectable change.

Reviewer: 3

Reviewer Name: Juliessa Pavon

Institution and Country: Assistant Professor, Geriatrics, Duke University, USA

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Reviewer comments have been sufficiently addressed.

The authors would like to thank the reviewer for the constructive comments and suggestions received.

Reviewer: 2

Reviewer Name: Dr Julie Richardson

Institution and Country: McMaster University, Canada

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Excellent contribution to the literature. No further revisions required

The authors would like to thank the reviewer for the constructive comments and suggestions received.

## VERSION 3 – REVIEW

| | |
|---|---|
| **REVIEWER** | Daniel Young<br>University of Nevada, Las Vegas USA |
| **REVIEW RETURNED** | 14-Oct-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have done an excellent job of responding to my concerns. I have no further criticisms of their work. |