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)	Inter-rater reliability of Berg Balance Scale, 30-seconds chair stand test and 6 meters walking test and construct validity of Berg Balance Scale in nursing home residents with mild and moderate dementia
AUTHORS	Telenius, Elisabeth; Engedal, Knut; Bergland, Astrid

VERSION 1 - REVIEW

REVIEWER	Stephen Downs Transitional Aged Care Program Bellingen Hospital Bellingen, NSW Australia
REVIEW RETURNED	22-May-2015

GENERAL COMMENTS	This study is potentially very useful, since only one other study considers the reliability of the BBS in a group with high levels of cognitive impairment. I have three main areas of concern: 1. You have not considered absolute reliability. Absolute reliability is usually described as minimal detectable change with 95% confidence. Bland J, Altman D (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327: 307–310 is a classic paper which demonstrates that relative reliability is only of very limited use, especially to clinicians. Absolute reliability is a more valid form of reliability and far more useful for clinicians. The value of your study would be very much enhanced by calculating the absolute reliability of the measures under consideration. 2. You have not included adequate information of how you have calculated the ICC. This is usually identified by the Shrout and Fliess category. See Shrout P, Fleiss J (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86: 420–428. The particularly high ICC found suggests the possibility that a category 3 calculation was used – if this is the case the calculation will have to be re-done. In any case the Shrout and Fliess category assumptions used must be stated. 3. There are sound reasons to conduct an inter-rater reliability test simultaneously, as has been done by this study, especially considering that some of your subjects might have quickly improving balance. More explanation is required that there is potential of having two people simultaneously recording a measurement overestimating its reliability. Any balance test might potentially be conducted slightly differently by different assessors, even with well defined scripts. Most items of the BBS have score changed by

	whether or not supervision is required (items 2,3,5,6,7,8,9,10,11,12). The tester might easily show by how the test is conducted if they believe supervision is required or not. Finally some mainly minor comments about specific parts of the manuscript: Introduction Page 4: Lines 5-9 – Conradson et al actually found both absolute and relative reliability of the BBS less than other most other authors who were included in a systematic review by this referee Downs S, Marquez J, Chiarelli P. The Berg Balance Scale has high intra- and interrater reliability but absolute reliability varies across the scale: a systematic review.J Physiother. 2013;59:93–99. One possible reason for this discrepancy might be that they studied the BBS within cognitively impaired people. One other factor potentially contributing the lower absolute reliability is that the BBS appears to display poorer absolute reliability toward the middle of the scale. Method: Page 5 Line 14 “by the help” Should read “with the help” Lines 21-22 The sentence “Prior to commencing the study….” Has incorrect grammar Page 6: Lines 33-24. As demonstrated by Bland and Altman’s classic paper, an ICC of 0.9 does not necessarily mean that a test is reliable enough to be considered extremely reliable Discussion: Page 10 Line 46 The high internal consistency of the BBS does not prove that it measures “balance”. It just suggests that they all measure a similar thing. Lines 52-56: Another reason for this discrepancy might be that you excluded people who couldn’t stand up and walk 6 metres, white the previous study appears to have only excluded those unable to stand.

REVIEWER	Vimonwan Hiengkaew Faculty of Physical Therapy Mahidol University Thailand
REVIEW RETURNED	26-May-2015

GENERAL COMMENTS	The study assessed inter-rater reliability of BBS, 30sCST, and 6-mWT. The authors used 2 raters in 33 participants. The study is just a pilot. In real life in clinic, there are many raters. Therefore, it is better if the study investigated inter-rater reliabitly more than 2 raters. The reviewer also provided a marked copy with detailed comments. Please contact the publisher for full information about it.

VERSION 1 – AUTHOR RESPONSE

Replies to reviewer 1

Thank you so much for your comments and input. Here are our replies:

You have not considered absolute reliability. Absolute reliability is usually described as minimal detectable change with 95% confidence. Bland J, Altman D (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327: 307–310 is a classic paper which demonstrates that relative reliability is only of very limited use, especially to clinicians. Absolute reliability is a more valid form of reliability and far more useful for clinicians. The value of your study would be very much enhanced by calculating the absolute reliability of the measures under consideration.
Thank you so much for this comment. We agree with you that absolute reliability is both more valid and more useful for clinicians. We have now calculated SEM and MDC for Berg Balance Scale and 6 meters walking test. This can be found in the methods-chapter

You have not included adequate information of how you have calculated the ICC. This is usually identified by the Shrout and Fliess category. See Shrout P, Fleiss J (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86: 420–428. The particularly high ICC found suggests the possibility that a category 3 calculation was used – if this is the case the calculation will have to be re-done. In any case the Shrout and Fliess category assumptions used must be stated.
We used the category 2/ Case 2 because we consider the evaluators to be a random sample from a population of potential raters.

There are sound reasons to conduct an inter-rater reliability test simultaneously, as has been done by this study, especially considering that some of your subjects might have quickly improving balance. More explanation is required that there is potential of having two people simultaneously recording a measurement overestimating its reliability. Any balance test might potentially be conducted slightly differently by different assessors, even with well defined scripts. Most items of the BBS have score changed by whether or not supervision is required (items 2,3,5,6,7,8,9,10,11,12). The tester might easily show by how the test is conducted if they believe supervision is required or not.
Thank you. This has been addressed in the section “limitations of the study”

Introduction

Page 4, Lines 5-9 – Conradson et al actually found both absolute and relative reliability of the BBS less than other most other authors who were included in a systematic review by this referee Downs S, Marquez J, Chiarelli P. The Berg Balance Scale has high intra- and interrater reliability but absolute reliability varies across the scale: a systematic review.J Physiother. 2013;59:93–99. One possible reason for this discrepancy might be that they studied the BBS within cognitively impaired people. One other factor potentially contributing the lower absolute reliability is that the BBS appears to display poorer absolute reliability toward the middle of the scale.

Thank you. The paragraph now reads: “To the authors´ knowledge, only one other study has investigated the reliability of BBS in a population of nursing home residents [15]. In that study 67% had dementia. They demonstrated a high ICC-value but a relatively low absolute reliability (minimal detectable change) of 7.7 points. However, inter-rater reliability was not tested”

Method:

Page 5 Line 14 “by the help” Should read “with the help”
Thank you. This has been corrected

Lines 21-22 The sentence “Prior to commencing the study….” Has incorrect grammar
Thank you. We have rewritten this section. It now reads: “The study was carried out by two physiotherapists. The examiners were trained in the standardised instructions of the tests and had experience from testing 120 patients in a study three months earlier.”

Page 6: Lines 33-24. As demonstrated by Bland and Altman’s classic paper, an ICC of 0.9 does not necessarily mean that a test is reliable enough to be considered extremely reliable.
Thank you for this important and relevant reference. This has been corrected. It now reads: An ICC of 0.8 or higher reflects high relative reliability, between 0.6 and 0.8 moderate reliability and less than 0.6 indicates poor reliability [44].

Discussion:

Page 10, Line 46: The high internal consistency of the BBS does not prove that it measures “balance”. It just suggests that they all measure a similar thing.
We agree with you. This has been corrected.

Lines 52-56: Another reason for this discrepancy might be that you excluded people who couldn’t stand up and walk 6 metres, white the previous study appears to have only excluded those unable to stand.
Thank you. It now reads: “Reasons for this discrepancy may be that our participants took part in an exercise study and therefore were more fit than the general nursing home population, and that we had somewhat stricter inclusion criteria regarding physical function.”

Replies to Reviewer: 2

Thank you for your valuable input. Here are our replies:

Reviewer Name Assoc. Prof. Dr. Vimonwan Hiengkaew
Institution and Country Faculty of Physical Therapy, Mahidol University, Thailand

Please leave your comments for the authors below

The study assessed inter-rater reliability of BBS, 30sCST, and 6-mWT. The authors used 2 raters in 33 participants. The study is just a pilot. In real life in clinic, there are many raters. Therefore, it is better if the study investigated inter-rater reliability more than 2 raters.
We agree that it would have strengthened the study, but regretfully this was not done.

General comment
The study assessed inter-rater reliability and concerned variation between 2 raters. However, in clinic or multicenter research projects, there are several raters. Therefore, it is better if the study concerns inter-rater reliability more than 2 raters
Abstract - Objective: Please provide full word of BBS
Introduction - fine
Methods
- See Table 1

- Please describe how participants did 30s CST. Do they need any helps? How to score? Execution and scoring of 30 seconds CST are described under “instruments”

- Please give number of participants who walk with and without devices for 6m
This has been included in table 1.

WT Results:
- Demographic characteristic is repeated
Thank you. This has been corrected.
- Distribution of scores and inter-rater reliability: it repeats with the data in table 2
Thank you. This has been corrected.
- Table 2: some data need to be added and corrected. See Table 2
Thank you, this has been corrected.
- Please give data of walking with and without devices for 6m WT
We have added information about use of walking-aid during 6 meter walking test in table 1. In total, 16 participants (50%) walked independently during the walking test.

Discussion: The study use 2 raters. However, in clinic there are many raters. Therefore, this is a limitation in the study for clinical application.
Thank you for this comment. This has been added under limitation of study.

Table 1:
1. In the table it presented only women participants, not all participants so I am not sure about the age, length of stay in nursing home, and others that are data of all participants or only female participants.
The results are for all 33 participants, however 25 (75.8%) were women.

2. Please specify neurological, heart, and musculoskeletal disease in paragraph of participants in method section
The most common diseases within each category have been included in the results section

3. MMSE: is it an inclusion criteria? How many minimum scores that the authors accept in this study?
MMSE was not used in inclusion of participants. Inclusion criteria regarding cognition was the score of 1 or 2 on Clinical dementia rating scale, CDR

4. Barthel index: Please give range of BI
Range have been included regarding BI, age, length of stay, MMSE and number of diagnosis and medications

5. Walking: Please give how many of participants walk independently with and without devices, and how many walk dependently with and without devices?

By walking independently we mean without any device. In table 1 we give the number of how many participants who walk independently in daily life (n=10, 30%), and how many who walked independently during the walking speed test (n=16, 50%).

6. What is number of diagnosis, and medication? Table 2:
It is the number of diagnoses and medications that are listed in the participants' medical cardex

7. For BBS o Please give SD - For 30s chair stand. Please show data of each tester o What does it mean about number in parenthesis (3.2)?
Standard deviation (SD) of BBS and CST is given in table 1. It is the number in the parenthesis. The mean score of both testers can be found in table 2: 6 number of rises (SD=3.2). The two evaluators scored the same, and therefore there is only one score in the table.

- For 6m WT o What does it mean about number in parenthesis (0.16),(0.18)? o Why does the minimum range of tester 1 and 2 much difference? Table 3: Change "evaluator (E)" to "tester (T)"
Thank you for noticing that there was a lack of information in the table: The number in paranthesis is the standard deviation (SD). The difference in minimum range between evaluators reflect the difference/ reliability in scoring. Thank you for your advice regarding changing "evaluator" to "tester". To make it easier to read, we prefer to call the testers evaluators to separate it from the word 'test'.