PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Agreement Between Electronic and Paper Epworth Sleepiness Scale Responses in Obstructive Sleep Apnoea: secondary analysis of a randomised controlled trial undertaken in a specialised tertiary care clinic
AUTHORS	Chen, Lily; Chapman, Julia; Yee, Brendon; Wong, Keith; Grunstein, Ronald; Marshall, Nathaniel; Miller, Christopher

VERSION 1 – REVIEW

REVIEWER	Xiangdong Tang Sleep Medicine Center, West China Hospital, Sichuan University
REVIEW RETURNED	08-Sep-2017

GENERAL COMMENTS	The aim of this study was to examine the agreement between different evaluation forms of ESS scale. To achieve this aim, 112 OSA patients were administered ESS using a computer at 18:00-19:00, and then the same participants were administered ESS in a paper at 19:30-21:00. Although it is clinical relevance, i have some concerns as bellow. 1. The time gap between measures is too short, impression of the first assessment of participants may impact the results. Further studies should be performed with different time gaps. 2. Because social cognition may be different between females and males, whether there are gender differences of the agreement between pESS and eESS should be explored. 3. In sleep area, other questionnaires, such as PSQI, ISI and STOP, should also be explored for the agreement between e-version and p-version.

REVIEWER	Luc Laberge Cégep de Jonquière, Canada
REVIEW RETURNED	03-Oct-2017

GENERAL COMMENTS	Page 2 of 17, lines 31-32, Design: Please specify this is a randomized clinical trial for which intervention. Page 2 of 17, line 38: Please specify you found no significant difference between what exactly. Page 4 of 17, line 67-67: In the study by Salaffi et al.

	(2013) that is cited by the authors, the Bland-Altman plot is used in conjunction with other techniques to assess the reliability of an electronic version, namely paired Student's t-tests and intraclass correlation coefficients (ICCs) for test-retest administration, smallest detectable difference (SDD) to assess agreement between scores, Spearman correlation coefficients, etc. The authors similarly used the Bland-Altman plot and analysis in conjunction with other methods of measurement (e.g. paired sample t-tests). These other methods could be mentioned along with the Bland-Altman plot in the Introduction. Indeed, Bland-Altman plots have been used in only 10% of the papers in a review on the equivalence of electronic and paper-based patient-reported outcome measures by Campbell et al. (2015) in Qual Life Res and in none of the papers reviewed by Muehlhausen et al. (2015) in Health and Quality of Life Outcomes. Page 5 of 17, Study design and patients, line 91: What is the rationale for "rejection of mechanical treatment within the past 2 years" as an inclusion criterion? Page 5 of 17, pages 106-108: Why the paper-based and electronic questionnaires were not completed in a crossover design? What was the rationale for choosing such a short time interval between completions? Page 7 of 17, Results: A graphic presenting the distribution of the ESS score would be a relevant addition. What is the proportion of overweight and obese patients (BMI of 30 or greater)? What is the proportion of patients with moderate (≥ 15, but < 30 per hour) and severe AHI (≥ 30 per hour)? If the vast majority of patients is severely affected, should it be specified somehow in the title and/or elsewhere in the manuscript? Page 7 of 17, lines 149-150: The sentence "56.3% of patients had an ESS difference within ±1, 80.4% within ±2 and 93.8% within ±4 (inclusive)" could be made clearer. See that of Johns (1992): "The paired scores differed by no more than 1 in 51.7% of students, by no more than 2 in 81.6% and by no more than 4 in 96.6%".

REVIEWER	Asad Khan The University of Queensland Brisbane, Australia
REVIEW RETURNED	23-Nov-2017

GENERAL COMMENTS	It is a pleasure to review the manuscript titled "Agreement Between Electronic and Paper Epworth Sleepiness Scale Responses in Obstructive Sleep Apnoea". The manuscript is a well written where the authors have presented a fair amount of statistics to address their research aims; however, I think there are a few issues with statistical analysis of the data. As mentioned by the authors, ESS has eight items with responses from 0-3. The authors have summed the items to get a total score without examining the unidimensionality of the items. It is inappropriate to sum the items of a scale without examining possible factor structure of the items as well as their internal consistencies. This essentially threats the validity of the total scores computed form the ESS items by the authors. This needs to be addressed before examining agreement between the two measurements.

	Furthermore, although the aim was to examine agreement between electronic and paper Epworth Sleepiness Scale Responses, the authors have tested the difference between the two using a paired t-test. In agreement or reliability study, we examine similarities, not the dis-similarities or differences. While the Bland-Altman plot is a graphical method to compare two measurements, it doesn't tell us the extent of agreement between the two measurements. In addition, the BA plot shows some differences that are outside the line of agreements, and we don't know whether these variations are acceptable or not in making a judgement about the interchangeability of the measurements. A desirable measure of reliability that includes both degree of correlation and agreement between measurements is the Intra-class correlation coefficient (ICC). I strongly recommend that the authors compute ICC for their measures so that the readers can understand the magnitude of the agreement between the measurements and make a judgement about the interchangeability between the two questionnaires. I think data analyses issues need to be addressed prior to considering the submission for publication.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Xiangdong Tang
Institution and Country: Sleep Medicine Center, West China Hospital, Sichuan University
Please state any competing interests: none

Please leave your comments for the authors below

The aim of this study was to examine the agreement between different evaluation forms of ESS scale. To achieve this aim, 112 OSA patients were administered ESS using a computer at 18:00-19:00, and then the same participants were administered ESS in a paper at 19:30-21:00. Although it is clinical relevance, i have some concerns as bellow.
1. The time gap between measures is too short, impression of the first assessment of participants may impact the results. Further studies should be performed with different time gaps.

- We agree and further studies should be completed to address this. In response, we have updated the third paragraph of the discussion section.

2. Because social cognition may be different between females and males, whether there are gender differences of the agreement between pESS and eESS should be explored.

- We did not have enough statistical power to look at male and female differences but we agree that this is something that would be good to look at in future larger studies.

3. In sleep area, other questionnaires, such as PSQI, ISI and STOP, should also be explored for the agreement between e-version and p-version.

- We agree with this comment and have updated the third paragraph of the discussion section to reflect this point.

Reviewer: 2
Reviewer Name: Luc Laberge
Institution and Country: Cégep de Jonquière, Canada
Please state any competing interests: None declared

Please leave your comments for the authors below

Page 2 of 17, lines 31-32, Design: Please specify this is a randomized clinical trial for which intervention.

- This is a baseline analysis of a clinical trial that seeks to discover if treating sleepy overweight or obese sleep apnea patients who cannot use standard treatments comparing hypocaloric diet with high protein/ low glycemic index diet for fat mass reduction and simultaneously comparing armodafinil with placebo for improving simulated driving ability and neuro-behavioural functioning. The manuscript reporting the primary outcome results are currently under peer review.

- We were however, unable to fit this information into the abstract given its length and because of the complex patient phenotype described. We have therefore included the clinical trial registration number instead so that readers can find this information. We have now also edited the first paragraph of the methods section in response.

Page 2 of 17, line 38: Please specify you found no significant difference between what exactly.

- This has been updated in the abstract.

Page 4 of 17, line 67-67: In the study by Salaffi et al. (2013) that is cited by the authors, the Bland-Altman plot is used in conjunction with other techniques to assess the reliability of an electronic version, namely paired Student's t-tests and intraclass correlation coefficients (ICCs) for test-retest administration, smallest detectable difference (SDD) to assess agreement between scores, Spearman correlation coefficients, etc. The authors similarly used the Bland-Altman plot and analysis in conjunction with other methods of measurement (e.g. paired sample t-tests). These other methods could be mentioned along with the Bland-Altman plot in the Introduction. Indeed, Bland-Altman plots have been used in only 10% of the papers in a review on the equivalence of electronic and paper-based patient-reported outcome measures by Campbell et al. (2015) in Qual Life Res and in none of the papers reviewed by Muehlhausen et al. (2015) in Health and Quality of Life Outcomes.

- We have inserted the intraclass correlation coefficient as requested in the second paragraph of the results section. We have also updated the fourth paragraph of the methods section, and the third paragraph of the discussion section to reflect this point.

Page 5 of 17, Study design and patients, line 91: What is the rationale for "rejection of mechanical treatment within the past 2 years" as an inclusion criterion?

- These data were fortuitous as we inadvertently collected both electronic and paper versions of the same Epworth sleepiness scale questionnaire, once from a routine clinical sleep study and a second from our clinical trial. In response, we have more clearly specified this information in the strengths and limitations section and have updated the first paragraph of the methods section to better reflect this.

Page 5 of 17, pages 106-108:

Why the paper-based and electronic questionnaires were not completed in a crossover design? What was the rationale for choosing such a short time interval between completions?

- Please see the previous point above.

Page 7 of 17, Results: A graphic presenting the distribution of the ESS score would be a relevant addition. What is the proportion of overweight and obese patients (BMI of 30 or greater)? What is the proportion of patients with moderate (≥ 15, but < 30 per hour) and severe AHI (≥ 30 per hour)? If the vast majority of patients is severely affected, should it be specified somehow in the title and/or elsewhere in the manuscript?

- The Bland Altman plots display the ESS distribution so we felt a histogram to be superfluous.
- We have inserted the further requested information into the first paragraph of the results section.

Page 7 of 17, lines 149-150: The sentence "56.3% of patients had an ESS difference within ±1, 80.4% within ±2 and 93.8% within ±4 (inclusive)" could be made clearer. See that of Johns (1992): "The paired scores differed by no more than 1 in 51.7% of students, by no more than 2 in 81.6% and by no more than 4 in 96.6%".

- Thank you – this has now been updated on the second paragraph of the results section.

Reviewer: 3
Reviewer Name: Asad Khan
Institution and Country: The University of Queensland, Brisbane, Australia
Please state any competing interests: Non

Please leave your comments for the authors below
It is a pleasure to review the manuscript titled "Agreement Between Electronic and Paper Epworth Sleepiness Scale Responses in Obstructive Sleep Apnoea". The manuscript is a well written where the authors have presented a fair amount of statistics to address their research aims; however, I think there are a few issues with statistical analysis of the data.

As mentioned by the authors, ESS has eight items with responses from 0-3. The authors have summed the items to get a total score without examining the unidimensionality of the items. It is inappropriate to sum the items of a scale without examining possible factor structure of the items as well as their internal consistencies. This essentially threats the validity of the total scores computed form the ESS items by the authors. This needs to be addressed before examining agreement between the two measurements.

- We would employ your approach if this was a novel instrument. The clinical reality is that this is used worldwide as a single factor questionnaire and is summed up in this way everywhere. We are testing reliability of the way that this instrument is already used clinically.

Furthermore, although the aim was to examine agreement between electronic and paper Epworth Sleepiness Scale Responses, the authors have tested the difference between the two using a paired t-test. In agreement or reliability study, we examine similarities, not the dis-similarities or differences.

While the Bland-Altman plot is a graphical method to compare two measurements, it doesn't tell us the extent of agreement between the two measurements. In addition, the BA plot shows some differences that are outside the line of agreements, and we don't know whether these variations are acceptable or not in making a judgement about the interchangeability of the measurements.

A desirable measure of reliability that includes both degree of correlation and agreement between measurements is the Intra-class correlation coefficient (ICC).
I strongly recommend that the authors compute ICC for their measures so that the readers can understand the magnitude of the agreement between the measurements and make a judgement about the interchangeability between the two questionnaires.

I think data analyses issues need to be addressed prior to considering the submission for publication.

- We have now provided the intraclass correlation coefficient in the manuscript - please see our response to the third point from Reviewer 2.

VERSION 2 – REVIEW

REVIEWER	Xiangdong Tang Sleep Medicine Center, West China Hospital, Sichuan University, China
REVIEW RETURNED	24-Dec-2017

GENERAL COMMENTS	I have no further comments.

REVIEWER	Luc Laberge ÉCOBES - Recherche et transfert, Cégep de Jonquière
REVIEW RETURNED	12-Dec-2017

GENERAL COMMENTS	I agree with Reviewer 1 that the time interval between is too short (computer 18:00-19:00 and paper 19:30-21:00; maximum 3 hours). Also, we do not know the proportion of patients for whom the time interval was only 30 minutes (i.e., computer at 19:00 followed by paper at 19:30). In such instances, there is a really high likelihood that patients have remembered their responses from the first version when completing the second version of the ESS. If this occurred, it would have significantly contributed to the consistency found across online and paper versions. This limit could be emphasized a little more in the Study limitations section. The authors compare their study to those of Olajos-Clow et al. (2010) and Koho et al. (2014) in which patients respectively waited 2 hours between completions and had morning and afternoon completions. On the other hand, the study by Bishop et al. (2010) only specified that patients were instructed to complete the instruments twice on the same day. In fact, it may reveal quite hard to find studies in which the time interval was as short as 30 minutes. In all, I suggest emphasizing this limit more clearly and underlining the necessity to replicate these results in this population.

REVIEWER	Asad Khan The University of Queensland
REVIEW RETURNED	07-Dec-2017

GENERAL COMMENTS	Thank you once again for giving me the opportunity to review the authors' responses. I appreciate that the authors have added ICC in

	their manuscript, although it isn't clear how did they deal with some of the potential outliers, demonstrated by the BA plot, in the calculation of ICC. I understand that reliability was the focus of the paper; however, I think we need to know whether total score from the ESS scale is a valid one before we look at its repeatability. Given that the authors have the relevant data, why can't they examine unidimensionality of the items before taking the total scores. If everybody is using ESS as a single factor questionnaire, it doesn't provide a guarantee that the scale would work the same way in a particular population. If it does offer a guarantee, why do we even need a reliability study for ESS?

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Xiangdong Tang
Institution and Country: Sleep Medicine Center, West China Hospital, Sichuan University, China
Please state any competing interests: none

Please leave your comments for the authors below I have no further comments.

Reviewer: 2
Reviewer Name: Luc Laberge
Institution and Country: ÉCOBES - Recherche et transfert, Cégep de Jonquière Please state any competing interests: None

Please leave your comments for the authors below

I agree with Reviewer 1 that the time interval between is too short (computer 18:00-19:00 and paper 19:30-21:00; maximum 3 hours). Also, we do not know the proportion of patients for whom the time interval was only 30 minutes (i.e., computer at 19:00 followed by paper at 19:30). In such instances, there is a really high likelihood that patients have remembered their responses from the first version when completing the second version of the ESS. If this occurred, it would have significantly contributed to the consistency found across online and paper versions.

This limit could be emphasized a little more in the Study limitations section. The authors compare their study to those of Olajos-Clow et al. (2010) and Koho et al. (2014) in which patients respectively waited 2 hours between completions and had morning and afternoon completions. On the other hand, the study by Bishop et al. (2010) only specified that patients were instructed to complete the instruments twice on the same day. In fact, it may reveal quite hard to find studies in which the time interval was as short as 30 minutes. In all, I suggest emphasizing this limit more clearly and underlining the necessity to replicate these results in this population.

- We have now emphasized this point more clearly in the study limitations section of the discussion.

Reviewer: 3
Reviewer Name: Asad Khan
Institution and Country: The University of Queensland Please state any competing interests: None

Please leave your comments for the authors below

Thank you once again for giving me the opportunity to review the authors' responses. I appreciate that the authors have added ICC in their manuscript, although it isn't clear how did they deal with some of the potential outliers, demonstrated by the BA plot, in the calculation of ICC.

- All data were included in the ICC calculation and we have included this in the second paragraph of the results section.

I understand that reliability was the focus of the paper; however, I think we need to know whether total score from the ESS scale is a valid one before we look at its repeatability. Given that the authors have the relevant data, why can't they examine unidimensionality of the items before taking the total scores.

If everybody is using ESS as a single factor questionnaire, it doesn't provide a guarantee that the scale would work the same way in a particular population. If it does offer a guarantee, why do we even need a reliability study for ESS?

- We do understand R3's point about the ESS potentially having more than one factor being an important consideration. Respectfully, however, and as the Reviewer suggests, this is not our research question and the ESS questionnaire is currently only used everywhere as a unidimensional scale - please see this topic explored in Kendzerska, T. B., Smith, P. M., Brignardello-Petersen, R., Leung, R. S., & Tomlinson, G. A. (2014). Evaluation of the measurement properties of the Epworth sleepiness scale: a systematic review. Sleep medicine reviews, 18(4), 321-331. DOI: 10.1016/j.smrv.2013.08.002. We agree with the Reviewer that further studies, with the main focus of exploring the dimensionality of the ESS, should be undertaken.

VERSION 3 – REVIEW

REVIEWER	Luc Laberge ECOBES - Recherche et transfert, Cégep de Jonquière, Québec, CANADA
REVIEW RETURNED	19-Jan-2018

GENERAL COMMENTS	No further comments.