PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Interrater and Test-retest Reliability of Quality Assessments by Novice Student Raters Using the Jadad and Newcastle-Ottawa Scales
AUTHORS	Oremus, Mark ; Oremus, Carolina; Hall, Geoffrey; McKinnon, Margaret; Systematic Review Team, ECT & Cognition

VERSION 1 - REVIEW

REVIEWER	Tatyana A Shamliyan. I do not have COI.
REVIEW RETURNED	30-May-2012

THE STUDY	The abstract does not describe a selection of the articles (13-20) and design distribution. The abstract does provide sample size calculation (10 students) and definitions of poor, fair, or excellent reliability. Conclusions about low reliability rating for specific questions are not supported by result section.
RESULTS & CONCLUSIONS	Limitations of the study do not address limitations of the used scales (Jadad scale and Newcastle-Ottawa Scale). The scales with judgmental and interpretive questions could have poor reliability despite several phases of training. The training to clarify “ambiguity” in the methodological and reporting quality of the evaluated studies may not improve reliability testing that was due to scale content and structure. The recent publication by the Cochrane Bias Methods Group and the Cochrane Statistical Methods Group recommended :” Do not use quality scales. Quality scales and resulting scores are not an appropriate way to appraise clinical trials. They tend to combine assessments of aspects of the quality of reporting with aspects of trial conduct, and to assign weights to different items in ways that are difficult to justify. Both theoretical considerations and empirical evidence suggest that associations of different scales with intervention effect estimates are inconsistent and unpredictable.” BMJ 2011;343:d5928 This publication should be mentioned in the discussion.
GENERAL COMMENTS	This is the first study that examined inter-rater and test-retest reliability of the quality assessment by non experienced reviewers. “For observational studies, key domains include the adequacy of case definition, exposure ascertainment, and outcome assessment.[5]” Selection and attrition bias are also very important when evaluating internal validity of the observational studies of health c are

	interventions. “Articles fluctuated across pairs because of constraints on rater availability due to competing academic demands”. Please clarify what do you mean by article fluctuation and “competing academic demands”. Please clarify had senior reviewers evaluated quality of the articles before giving the articles to the students and had they compared own ranking with the ranking by non experienced raters? Please justify the same size and describe student invitation response rate and articles selection. Please clarify that your goal was quality evaluation of observational studies of health care interventions.

REVIEWER	Verhagen, Arianne Erasmus MC, University Medical Centre Rotterdam, the Netherlands, General Practice
REVIEW RETURNED	05-Jun-2012

THE STUDY	I think the design is flawed in a way that the authors actually evaluate the effect of a training course on quality assessment rather than the reliability. also the statistics need to be discussed with a statistician. They also do not seem to be very well documented on the topic, various key references are missing and they used a scale that it not very often used (modified Jadad scale).
RESULTS & CONCLUSIONS	I think the discussion lack clarity and does not discuss the main limitations of studies like these.
GENERAL COMMENTS	This paper is a reliability study evaluating the inter- and intrarater reliability of two quality assessment scales frequently used in systematic reviews. Quality assessment should be a valid and reliable exercise especially when it is used to exclude the studies with low quality. Therefore it is important to evaluate these issues. General comments 1. The authors do not explain why they study the reliability in (inexperienced) students. I cannot see what the rationale is to do so. What they actually do is to evaluate the output of the training course. I think this training course (of 90 minutes!) is not very good as the interrater reliability directly after the course is low. In 2 months time the authors do not expect any recall of the first score, I assume most of the information of the course might also be forgotten. Apparently there was not much information of the 90 minute course to forget because the intra-rater reliability higher that the interrater reliability. 2. For the assessment the authors used a modified Jadad scale. I do not think A. Jadad will be very pleased that this scale is chosen instead of the original one. The modifications are all related to external validity and have no clear relation with actual quality of the study. I recommend sticking to the original scale when studying reliability in a general way as the authors aimed to do. 3. Concerning the analysis the authors not only assessed reliability using Kappa scores and ICC, but also calculated mean differences between rater-paires. I do not see the rationale for this. This analysis does not add anything to the answer on the study question whether quality assessment done by inexperienced raters after a 90 min course is reliable. I should delete this from the manuscript as it

	confuses the reader and does not inform them. 4. One of the key messages is that the reliability between inexperienced raters is low. You do not have to do a study to show this, every course teacher knows, so whats new?

VERSION 1 – AUTHOR RESPONSE

Reviewer: Tatyana A Shamliyan

1. The abstract does not describe a selection of the articles (13-20) and design distribution.

We added these descriptions to the abstract (lines 9-11) and revised a small section of the text (p. 7, line 22) to clarify how the articles were distributed to raters.

2. The abstract does provide sample size calculation (10 students) and definitions of poor, fair, or excellent reliability.

Thank you for the comment.

3. Conclusions about low reliability rating for specific questions are not supported by result section.

We re-wrote the results section of the abstract to include summaries of the calculated Kappas and intraclass correlation coefficients. We removed reference to specific types of questions from the abstract conclusion. We also revised the first bullet under 'Key messages' of the 'Article Summary'.

4. Limitations of the study do not address limitations of the used scales (Jadad scale and Newcastle-Ottawa Scale). The scales with judgmental and interpretive questions could have poor reliability despite several phases of training.

We added mention of this issue to the discussion (p. 16, lines 18-20).

5. The training to clarify "ambiguity" in the methodological and reporting quality of the evaluated studies may not improve reliability testing that was due to scale content and structure.

We added mention of this issue to the discussion (p. 12, lines 1-4).

6. The recent publication by the Cochrane Bias Methods Group and the Cochrane Statistical Methods Group recommended :" Do not use quality scales.
Quality scales and resulting scores are not an appropriate way to appraise clinical trials. They tend to combine assessments of aspects of the quality of reporting with aspects of trial conduct, and to assign weights to different items in ways that are difficult to justify. Both theoretical considerations and empirical evidence suggest that associations of different scales with intervention effect estimates are inconsistent and unpredictable." BMJ 2011;343:d5928. This publication should be mentioned in the discussion.

We added mention of this issue to the conclusion (p. 17, lines 18-23). We also cited the BMJ publication (reference 33 in bibliography).

7. "For observational studies, key domains include the adequacy of case definition, exposure ascertainment, and outcome assessment.[5]" Selection and attrition bias are also very important when evaluating internal validity of the observational studies of health c are interventions.

We added mention of this issue to the introduction (p. 5, line 17).

8. “Articles fluctuated across pairs because of constraints on rater availability due to competing academic demands”. Please clarify what do you mean by article fluctuation and “competing academic demands”.

We revised the sentence in question to enhance clarity (p. 8, lines 1-2).

9. Please clarify had senior reviewers evaluated quality of the articles before giving the articles to the students and had they compared own ranking with the ranking by non experienced raters?

Senior reviewers did not rate article quality. We added mention of this issue in the limitations section (p. 16, lines 14-16).

10. Please justify the same size and describe student invitation response rate and articles selection.

The 10 students in the study were a convenience sample and we added mention of this fact to the limitations section (p. 16, lines 12-13).
We added two sentences to the methods to describe student invitation and response rate (p. 7, lines 11-13).

We clarified article selection by editing the last paragraph of the introduction (p. 6, lines 10-13) and the first paragraph of the methods (p. 7, lines 3-5) to explain that the 78 articles in our study came from an ongoing systematic review of cognitive impairment and electroconvulsive therapy. These 78 articles were the included studies in the meta-analysis.

11. Please clarify that your goal was quality evaluation of observational studies of health care interventions.

Our goal was to examine the reliability of quality assessments done by inexperienced student raters. We included quality assessments of RCTs and observational studies in our examination. We clarified these points in the last paragraph of the introduction (p. 6, lines 10-13).

Reviewer: Arianne P Verhagen

1. I think the design is flawed in a way that the authors actually evaluate the effect of a training course on quality assessment rather than the reliability.

We agree that training programs may influence reliability and we added mention of this fact to the ‘strengths and limitations’ section of ‘article summary’ box and to the limitations section (p. 16, lines 11-12).

We disagree that the design is flawed. We patterned our study design on an approach used by several similar investigations, including references 12, 13, 14, 15, 16, 18, 31, and 32 from our bibliography. The primary objective of all of these studies was to calculate reliability, not to examine the impact of training programs on reliability.

2. [A]lso the statistics need to be discussed with a statistician.

Eleanor Pullenayegum and Harry Shannon, both listed in the acknowledgements, are statisticians who provided feedback on the manuscript prior to submission.

3. They [statistics] also do not seem to be very well documented on the topic.

We used standard statistics (Kappa and intraclass correlation coefficient) to calculate reliability. We referenced our sources for these statistics (references 20, 21, 23, and 24 in the bibliography). We also referenced the sources for our interpretations (e.g., 'poor', 'fair') of these statistics (references 22 and 25 in the bibliography).

4. [V]arious key references are missing.

We would be happy to look into these references if the reviewer could provide us with a list.

5. [T]hey used a scale that it not very often used (modified Jadad scale).

The modified Jadad scale contains the original three questions proposed by Jadad et al. (http://ac.els-cdn.com/0197245695001344/1-s2.0-0197245695001344-main.pdf?_tid=7e939adc61c0566105605a2bc2682525&acdnat=1339428722_689c0e9c11502cca0c52cacb7a1e1d76). The modified scale also contains three additional questions considered by Jadad et al. in their original scale development work. The additional three questions were added to the Jadad scale for a systematic review of Alzheimer's disease medications (http://www.cadth.ca/media/pdf/106_alzheimers1_tr_e.pdf). The modified Jadad scale had excellent interrater reliability (ICC=0.90) in this systematic review (see reference 19 in the bibliography) and was subsequently used in a range of other systematic reviews, e.g., Testing for BNP and NT-proBNP in the Diagnosis and Prognosis of Heart Failure (http://www.ahrq.gov/downloads/pub/evidence/pdf/bnp/bnp.pdf), Diagnosis and Treatment of Secondary Lymphedema (https://www.cms.gov/Medicare/Coverage/DeterminationProcess/downloads/id66aTA.pdf), Pharmacological Treatment of Dementia (http://www.ahrq.gov/downloads/pub/evidence/pdf/dempharm/dempharm.pdf).

6. I think the discussion lack clarity and does not discuss the main limitations of studies like these.

We would be happy to clarify any section of the discussion that may be lacking clarity. We encourage the reviewer to point out any sections that she feels may require more clarity.

We provided a limitations section in the initial manuscript and we added to this section in response to both reviewers' comments.

General comments
1. The authors do not explain why they study the reliability in (inexperienced) students. I cannot see what the rationale is to do so.

The introduction to our original manuscript contained two paragraphs that explained our rationale for studying reliability in inexperienced students. With some modifications, we retained these paragraphs in the current version of the manuscript (p. 6, lines 5-13).

What they actually do is to evaluate the output of the training course. I think this training course (of 90 minutes!) is not very good as the interrater reliability directly after the course is low.

We agree that training programs may influence reliability and we added mention of this fact to the 'strengths and limitations' section of 'article summary' box and to the limitations section (p. 16, lines 11-12).

Some of the poor reliability scores may also result from the difficulty of using the NOS, which we addressed in the discussion of the original manuscript. We retained this section in the current manuscript (p. 12, lines 16-23; p. 13, lines 1-4).

In 2 months time the authors do not expect any recall of the first score, I assume most of the information of the course might also be forgotten.

The literature provided very little guidance on an adequate time frame for measuring test-retest reliability in studies such as ours. We based our two-month interval on a study that did utilize methods similar to ours (reference 13 in the bibliography).

Since our purpose was not to evaluate a training program, we did not assess recall of course content.

2. For the assessment the authors used a modified Jadad scale. I do not think A. Jadad will be very pleased that this scale is chosen instead of the original one. The modifications are all related to external validity and have no clear relation with actual quality of the study. I recommend sticking to the original scale when studying reliability in a general way as the authors aimed to do.

Table 1 already presents interrater reliability for the 'original' three Jadad questions (randomization, double-blinding, description of withdrawals), along with the follow-up questions on appropriateness of double-blinding and randomization, which are also part of Jadad et al.'s initial (3-item) scale. To account for the reviewer's comment, we calculated interrater and test-retest reliability for total scores based on the original 3-item Jadad scale (p. 8, lines 21-23; p. 10, lines 10 & 19-20; Table 3).

3. Concerning the analysis the authors not only assessed reliability using Kappa scores and ICC, but also calculated mean differences between rater-pairs. I do not see the rationale for this. This analysis does not add anything to the answer on the study question whether quality assessment done by inexperienced raters after a 90 min course is reliable. I should delete this from the manuscript as it confuses the reader and does not inform them.

We removed the mean differences comparison from the manuscript.

4. One of the key messages is that the reliability between inexperienced raters is low. You do not have to do a study to show this, every course teacher knows, so what's new?

This article is the first research to quantify test-retest and interrater reliability for inexperienced student raters. This study empirically tests what "every course teacher knows". Prior to this study, no one could say for sure whether reliability for inexperienced raters was low, nor could anyone estimate 'how low' this reliability might be.

Sincerely,

Mark Oremus, PhD
McLaughlin Foundation Professor of Population and Public Health & Assistant Professor, Department of Clinical Epidemiology & Biostatistics
Co-Associate Director, McMaster Evidence-based Practice Centre
Associate Scientific Director, Canadian Longitudinal Study on Aging
McMaster University

VERSION 2 – REVIEW

REVIEWER	Shamliyan, Tatyana (R0) University of Minnesota, health Policy and Managment

REVIEW RETURNED	28-Jun-2012

GENERAL COMMENTS	The authors revised the manuscript responding to reviewer's comments and clarified limitations of the study. Reliability in quality assessment and ranking strength of evidence in systematic reviews is poorly reported in systematic reviews of health care interventions. This article is the first attempt in quantifying reliability in quality evaluation for inexperienced raters. The authors suggested future transparent reliability testing for various tools used in systematic reviews.

REVIEWER	Verhagen, Arianne Erasmus MC, University Medical Centre Rotterdam, the Netherlands, General Practice
REVIEW RETURNED	28-Jun-2012

THE STUDY	This study aims at assessing inter-rater reliability, but they also evaluate their training course. Indeed inter-rater reliability is nicely assessed, but no attention (or hardly any) is paid to the effect of the training course.
RESULTS & CONCLUSIONS	I miss a critical evaluation of what is done and why. I still think you do not need to study this to know for sure that inexperienced raters have a low reliability. OK, the authors are right, we now know how unreliable, but it does not add anything to what we already know. Also no recommendations can be made for future studies on how to make sure that quality assessment is reliable.