BMJ open
accessible medical research

# Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and Newcastle–Ottawa Scales

Mark Oremus,[1,2] Carolina Oremus,[3,4] Geoffrey B C Hall,[3,4] Margaret C McKinnon,[3,4] ECT & Cognition Systematic Review Team[3,4,*]

**To cite:** Oremus M, Oremus C, Hall GBC, *et al*. Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and Newcastle–Ottawa Scales. *BMJ Open* 2012;**2**:e001368. doi:10.1136/bmjopen-2012-001368

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2012-001368).

*The ECT & Cognition Systematic Review Team includes Allyson Graham, Caitlin Gregory, Gagan Fervaha, Lindsay Hanford, Anthony Nazarov, Melissa Parlar, Maria Restivo, Erica Tatham and Wanda Truong.

Received 23 April 2012
Accepted 29 June 2012



For numbered affiliations see end of article.

**Correspondence to**
Dr Mark Oremus;
oremusm@mcmaster.ca

## ABSTRACT

**Introduction:** Quality assessment of included studies is an important component of systematic reviews.

**Objective:** The authors investigated inter-rater and test–retest reliability for quality assessments conducted by inexperienced student raters.

**Design:** Student raters received a training session on quality assessment using the Jadad Scale for randomised controlled trials and the Newcastle–Ottawa Scale (NOS) for observational studies. Raters were randomly assigned into five pairs and they each independently rated the quality of 13–20 articles. These articles were drawn from a pool of 78 papers examining cognitive impairment following electroconvulsive therapy to treat major depressive disorder. The articles were randomly distributed to the raters. Two months later, each rater re-assessed the quality of half of their assigned articles.

**Setting:** McMaster Integrative Neuroscience Discovery and Study Program.

**Participants:** 10 students taking McMaster Integrative Neuroscience Discovery and Study Program courses.

**Main outcome measures:** The authors measured inter-rater reliability using κ and the intraclass correlation coefficient type 2,1 or ICC(2,1). The authors measured test–retest reliability using ICC (2,1).

**Results:** Inter-rater reliability varied by scale question. For the six-item Jadad Scale, question-specific κs ranged from 0.13 (95% CI −0.11 to 0.37) to 0.56 (95% CI 0.29 to 0.83). The ranges were −0.14 (95% CI −0.28 to 0.00) to 0.39 (95% CI −0.02 to 0.81) for the NOS cohort and −0.20 (95% CI −0.49 to 0.09) to 1.00 (95% CI 1.00 to 1.00) for the NOS case–control. For overall scores on the six-item Jadad Scale, ICC(2,1)s for inter-rater and test–retest reliability (accounting for systematic differences between raters) were 0.32 (95% CI 0.08 to 0.52) and 0.55 (95% CI 0.41 to 0.67), respectively. Corresponding ICC(2,1)s for the NOS cohort were −0.19 (95% CI −0.67 to 0.35) and 0.62 (95% CI 0.25 to 0.83), and for the NOS case–control, the ICC(2,1)s were 0.46 (95% CI −0.13 to 0.92) and 0.83 (95% CI 0.48 to 0.95).

**Conclusions:** Inter-rater reliability was generally poor to fair and test–retest reliability was fair to excellent. A pilot rating phase following rater training may be one way to improve agreement.

## ARTICLE SUMMARY

### Article focus

- To examine the inter-rater and test–retest reliability of inexperienced raters' quality assessments of articles included in a systematic review.

### Key messages

- Among inexperienced raters, inter-rater reliability using the Jadad Scale and Newcastle–Ottawa Scale was generally poor to fair; test–retest reliability was fair to excellent.
- Systematic reviewers must pay special attention to training inexperienced quality raters; a pilot rating phase might be a helpful means of improving reliability among inexperienced raters, especially when rating observational study quality.

### Strengths and limitations of this study

- No other study has examined the reliability of quality assessments in a group of inexperienced raters.
- Results may differ depending on rater background and experience, rater training, quality assessment instruments and topic under study.

## INTRODUCTION

Systematic reviews summarise healthcare research evidence, and they are useful for assessing whether treatment benefits outweigh risks.[1 2] Accordingly, conclusions drawn from systematic reviews may impact clinical care and patient outcomes, thereby necessitating high standards of methodological rigour.

One critical component of conducting systematic reviews involves evaluation of the

methodological quality of included studies. Study quality may influence treatment effect estimates and the validity of conclusions drawn from such estimates.[3] Through quality assessment, researchers identify strengths and weaknesses of existing evidence[4] and suggest ways to improve future research.

Careful work has identified key quality assessment domains.[1 5] For randomised controlled trials (RCTs), these domains include appropriate generation of random allocation sequences, concealment of allocation sequences, blinding (of participants, healthcare providers, data collectors and outcome assessors) and reporting of proportions of patients lost to follow-up.[1] For observational studies, key domains include the adequacy of case definition, exposure ascertainment and outcome assessment,[5] as well as selection and attrition biases.

Numerous scales exist to help raters assess study quality.[5–11] The majority of these scales list quality assessment domains and require raters to indicate whether each domain is present or absent from the studies under consideration. Some scales (eg, Jadad,[6] Newcastle–Ottawa Scale (NOS)[5]) assign points when quality domains are present, thus permitting the calculation of overall 'quality scores'. Other scales (eg, risk of bias[8]) ask raters to rank the degree of bias (high, low, unclear) associated with each quality domain.

Generally, quality scales demonstrate good inter-rater and test–retest reliability. Reliability coefficients such as κ are typically >0.60,[9–17] although recent work reports κs of <0.50 for eight of the nine questions on the NOS.[18]

Although quality assessment is now regarded as a standard component of systematic reviews, one issue that has received little attention in the literature is the effect of rater experience on the reliability of quality assessments. This issue is important because raters may be drawn from vast pools of persons with varying degrees of methods expertise, from experienced faculty to inexperienced students.

We investigated inter-rater and test–retest reliability for student raters with no previous experience in the quality assessment of RCTs and observational studies. To the best of our knowledge, no other study has examined this topic.

## METHODS

### Study design

In an ongoing systematic review of cognitive impairment following electroconvulsive therapy (ECT) to treat major depressive disorder, 78 published articles passed title and abstract and full-text screening. These articles formed the basis of this study. Fifty-five of the articles reported the results of RCTs, with one article containing results of five separate studies and two other articles each containing results of two separate studies, for a total of 61 RCTs. Fifteen articles reported on cohort studies and eight reported on case–control studies. Eleven articles were published prior to 1980, 17 between 1980 and 1989, 15 between 1990 and 1999, and 35 since 2000.

We invited all 10 students (three undergraduate and seven graduate) taking a 'special topics' course in the McMaster Integrative Neuroscience Discovery and Study Program to participate in this study. All 10 students accepted the invitation. One author (MO) with systematic review experience trained the students to rate the methodological quality of published study reports using the six-item Jadad Scale for RCTs[6 19] and the NOS for observational studies.[5] Training consisted of a 90 min didactic session divided into two parts: part one highlighted the importance of quality assessment in systematic reviews and part two contained a question-by-question description of the Jadad and NOS instruments. We provided a standardised tabular spreadsheet for student raters to use during quality assessment.

We used a random number table to assign the student raters into five pairs and we randomly distributed between 13 and 20 articles to each pair. None of the 78 articles was assigned to more than one pair; pairs received a mix of RCTs and observational studies. The number of articles assigned to the pairs depended on the amount of time each rater could devote to this study.

Raters determined the type of study design (ie, RCT or observational) for each of their assigned articles and one author (CO) verified their choices. Raters then independently rated their assigned articles to permit us to examine inter-rater reliability.

### Statistical analysis

We used κ (kappa)[20 21] to measure inter-rater reliability for individual Jadad and NOS questions. We interpreted κ values as follows: >0.80 was very good, 0.61–0.80 was good, 0.41–0.60 was moderate, 0.21–0.40 was fair and <0.21 was poor.[22]

For test–retest reliability, each rater re-assessed half of the articles to which they had been assigned during the inter-rater reliability phase. The re-assessments took place 2 months after the inter-rater reliability phase[13] to minimise the possibility that recall of the first assessments would influence the second assessments.

We employed the intraclass correlation coefficient-model 2,1 or ICC(2,1)[23] to measure inter-rater and test–retest reliability for the Jadad and NOS total scores. We computed separate ICC(2,1) values for consistency (systematic differences between raters are considered irrelevant) and absolute agreement (systematic differences between raters are considered relevant).[24] ICC (2,1) values were interpreted as follows: >0.75 was excellent, 0.40–0.75 was fair to good and <0.40 was poor.[25]

We calculated two sets of ICC(2,1)s for the Jadad Scale. The first set pertained to the six-item Jadad Scale,[19] and the second set pertained to the original three-item Jadad Scale.[6]

SAS V.9.2 (The SAS Institute) was used to calculate κ; SPSS V.20 (IBM Corp.) was used to calculate ICC(2,1). The level of significance was $\alpha$=0.05.

## RESULTS

### Inter-rater reliability

For inter-rater reliability, agreement between raters on individual questions was generally poor (table 1). Half of the questions on the Jadad Scale had moderate κs and the other half had poor κs. On the NOS, all κs were poor for the cohort study questions (NOS cohort) and six of the eight κs were poor for the case–control study questions (NOS case–control).

Examining total scale scores within rater pairs (table 2), agreement was poor for the Jadad Scale (six- and three-item versions) and NOS cohort and fair for the NOS case–control. However, point estimate ICC(2,1)s for the NOS cohort and case–control were not statistically significantly different from zero. Point estimate ICC (2,1)s and 95% CIs did not appreciably differ according to calculation based on consistency or absolute agreement.

### Test–retest reliability

Test–retest reliability following a 2-month interval between assessments was fair to good for the Jadad Scale and NOS cohort and excellent for the NOS case–control (table 3). Test–retest reliability was slightly higher for the three-item Jadad Scale versus the six-item Jadad Scale. Point estimate ICC(2,1)s and 95% CIs calculated for consistency were similar to the results calculated for absolute agreement.

## DISCUSSION

### Overview and discussion of key findings

We investigated inter-rater and test–retest reliability for student raters with no previous experience in quality assessment. Our study is novel because, to the best of our knowledge, no other research has examined this issue. The raters used the Jadad Scale and NOS to assess the quality of studies on the topic of ECT and cognitive impairment. Inter-rater reliability was generally poor to fair and test–retest reliability was fair to excellent. Our results highlight the need for researchers to consider rater experience during the quality assessment of articles included in systematic reviews.

For inter-rater reliability, the poor κs on the Jadad Scale pertained to the questions about appropriateness of double blinding and the clarity of reporting withdrawals, inclusion/exclusion criteria and adverse effects. Often, authors did not report methods of blinding and raters had to make judgements about whether to award a point for the question on appropriateness of double blinding. Despite what we communicated during the training session, some raters may have given authors the benefit of the doubt and awarded the point for appropriateness if studies simply reported double blinding, even though another question on the Jadad Scale already asked whether authors reported their studies as blinded. Similarly, differences in rater opinion regarding what constitutes an 'adequate' description of withdrawals, inclusion/exclusion criteria or adverse effects led to poor agreement on these questions. To improve inter-rater agreement among inexperienced raters, we suggest a pilot phase wherein raters rate the quality of a subsample of articles to allow for the identification and clarification of areas of ambiguity.

We recognise that any strategy to improve reliability will be limited by instrument content and structure. Scales with larger numbers of interpretive questions will likely have lower reliability than scales with fewer interpretive questions, regardless of the efforts made to improve reliability.

With regard to the NOS, question-specific inter-rater reliability was poorer than that of the Jadad Scale. We believe that the NOS's poor reliability may be explained

**Table 1** Inter-rater reliability for Jadad Scale and Newcastle–Ottawa Scale (NOS): by question

| Question—Jadad Scale | κ (95% CI) | Question—NOS cohort | κ (95% CI) | Question—NOS case–control | κ (95% CI) |
|---|---|---|---|---|---|
| Randomisation | 0.50 (−1.00 to 1.00) | Representativeness of exposed cohort | −0.13 (−0.36 to 0.11) | Case definition adequate | 1.00 (1.00 to 1.00) |
| Appropriate randomisation | 0.56 (0.29 to 0.83) | Selection of non-exposed cohort | −0.14 (−0.28 to 0.00) | Cases representative | −0.20 (−0.49 to 0.09) |
| Double blind | 0.41 (0.16 to 0.66) | Exposure ascertainment | 0.00 (0.00 to 0.00) | Control selection | 0.25 (−0.19 to 0.69) |
| Appropriate double blind | 0.17 (−0.07 to 0.41) | Outcome not present at baseline | 0.20 (−0.33 to 0.73) | Control definition | 0.14 (−0.54 to 0.82) |
| Description of withdrawals | 0.21 (−0.02 to 0.45) | Comparability of cohorts | 0.12 (−0.23 to 0.47) | Case and control comparability | 0.00 (0.00 to 0.00) |
| Description of inclusion/ exclusion criteria | 0.27 (−0.03 to 0.57) | Outcome assessment | 0.31 (−0.08 to 0.69) | Exposure ascertainment | −0.11 (−0.68 to 0.46) |
| Description of adverse effects | 0.13 (−0.11 to 0.37) | Follow-up long enough | −0.09 (−0.22 to 0.04) | Same ascertainment method for cases and controls | 0.60 (−0.07 to 1.00) |
| Description of statistical analysis | 0.49 (0.21 to 0.77) | Follow-up adequate | 0.39 (−0.02 to 0.81) | Non-response rate | −0.11 (−0.65 to 0.43) |

κ, Kappa.

**Table 2** Inter-rater reliability for Jadad and Newcastle–Ottawa Scales: total scale scores within rater pairs

| Scale | ICC(2,1) (95% CI), consistency* | ICC(2,1) (95% CI), absolute agreement† |
|---|---|---|
| Jadad—six item | 0.32 (0.08 to 0.53) | 0.32 (0.08 to 0.52) |
| Jadad—three item | 0.35 (0.11 to 0.56) | 0.35 (0.11 to 0.56) |
| Newcastle–Ottawa—cohort | −0.19 (−0.63 to 0.34) | −0.19 (−0.67 to 0.35) |
| Newcastle–Ottawa—case–control | 0.55 (−0.18 to 0.89) | 0.46 (−0.13 to 0.92) |

*ICC(2,1) where systematic differences between raters are irrelevant.
†ICC(2,1) where systematic differences between raters are relevant.
ICC, intraclass correlation coefficient.

in part by differences in how raters answered interpretive questions, for example, whether exposed cohorts are somewhat or truly representative of the average exposed person in the community (first question on NOS cohort).

Poor question-specific inter-rater agreement on the NOS also reflects an inherent challenge with rating the quality of observational studies compared with RCTs. This challenge is exemplified by the multiplicity of tools that exist to assess observational study quality. Two systematic reviews[26 27] each found over 80 such tools, which varied in design and content. Despite the cornucopia of tools, no gold standard scale exists to rate the quality of observational studies.[28]

Rater disagreements on interpretive questions and inherent challenges with assessing observational study quality explain the negative κs that were calculated for some NOS questions. Negative κs result when agreement occurs less often than predicted by chance alone. This suggests genuine disagreement between raters or an underlying issue with the instrument itself.[29] Indeed, Hartling *et al*[18] reported that raters had difficulty using the NOS because of uncertainty over the meaning of certain questions (eg, representativeness of the exposed cohort, selection of non-exposed cohort) and response options (eg, 'truly' vs 'somewhat' exposed). These difficulties existed despite Hartling *et al*'s use of a pilot training phase. Our raters' difficulties with the interpretative questions might have been a function of issues with the NOS, which could be related to the broader challenge of assessing the quality of observational studies.

Question-specific differences between raters also led to poor inter-rater agreement on total scores for the Jadad Scale and NOS cohort. This may not be evident by comparing the κs and ICC(2,1)s calculated for the Jadad. κs for four of the eight Jadad questions were moderate yet the ICC(2,1) for total score was poor. However, since total scores are computed using raters' answers to all of the questions on a scale (some answers are awarded one point and others zero points), raters who disagree on small numbers of questions (eg, two of the eight questions) will nonetheless show poor agreement on total scores.

Conversely, for the NOS case–control, κs for six of the eight questions were poor yet the ICC(2,1) was fair. In this situation, no 'reliability' relation exists between responses to questions and total scores. For example, rater 1 might answer 'yes' (one point per 'yes' response) and rater 2 might answer 'no' (zero points per 'no' response) to even-numbered questions. For odd-numbered questions, the pattern is reversed. Assuming eight questions, inter-rater reliability at the question level will be poor because the raters did not agree on their responses, but their overall scores will be equivalent.

Many authors base their discussions of study quality in systematic reviews on raters' responses to individual questions on quality assessment scales. Given that we found generally poor inter-rater reliability on answers to questions, the process of resolving conflicts between raters becomes important. Many reviews simply report that raters solved disagreements by consensus without describing specific procedures. We speculate that conflict resolution may occasionally be approached in an ad hoc nature or treated as a nuisance to be dealt with as expeditiously as possible. We suggest the process of conflict resolution should be more of a formalised endeavour requiring raters to set aside some 'resolution time' and articulate their reasons for choosing specific

**Table 3** Test–retest reliability for Jadad and Newcastle–Ottawa Scales: comparison of total scale scores for individual raters after two assessments

| Scale | ICC(2,1) (95% CI), consistency* | ICC(2,1) (95% CI), absolute agreement† |
|---|---|---|
| Jadad—six item | 0.56 (0.42 to 0.67) | 0.55 (0.41 to 0.67) |
| Jadad—three item | 0.67 (0.55 to 0.76) | 0.67 (0.55 to 0.76) |
| Newcastle–Ottawa—cohort | 0.61 (0.24 to 0.82) | 0.62 (0.25 to 0.83) |
| Newcastle–Ottawa—case–control | 0.85 (0.55 to 0.95) | 0.83 (0.48 to 0.95) |

*ICC(2,1) where systematic differences between raters are irrelevant.
†ICC(2,1) where systematic differences between raters are relevant.
ICC, intraclass correlation coefficient.

answers. In the event the raters do not agree, a third party may be asked to listen to each rater's opinion and make a decision. Although space restrictions in journals might prevent authors from reporting such procedures (when they exist) in manuscripts, the move towards publication of systematic review protocols, for example, as mandated by the United States Agency for Healthcare Research and Quality's Effective Health Care Program,[30] provides authors with an opportunity to elaborate on their consensus processes.

Test–retest reliability was better than inter-rater reliability. Individual raters appeared to adopt a uniform approach to assessing the quality of articles assigned to them. Each rater had her or his own understanding of the interpretive questions and applied this point-of-view consistently throughout the rating process. The issue was the difference in interpretations between raters.

## Comparison with other studies

To the best of our knowledge, no other study has examined inter-rater and test–retest reliability for a group of novice student quality assessors. Two published studies[31 32] of rater agreement included persons with different levels of experience, although the focus was on extraction of article data (eg, info on study design, sample characteristics, length of follow-up, definition of outcome and results) rather than quality assessment. Horton *et al*[31] classified rater experience as minimal, moderate or substantial and asked raters to extract data from three studies on insomnia therapy. They found no statistically significant differences in error rates according to experience. Haywood *et al*[32] trained two experienced raters and one inexperienced rater to independently extract data from seven studies. Agreement between raters was largely perfect.

A recent AHRQ methods report had 16 raters assess the quality of 131 cohort studies using the NOS. Rater experience ranged from 4 months to 10 years; 13 raters had formal training in systematic reviews.[18] Ks were <0.50 for eight of the nine NOS questions, although the authors did not break down their results by rater experience.

Oremus *et al* examined the inter-rater reliability of the Jadad Scale using three raters (two experienced faculty members and one inexperienced PhD student), who read the methods and results of 42 Alzheimer's disease drug trials.[19] The ICC(2,1) for total scores on the Jadad Scale was 0.90. Al-Harbi *et al*[12] engaged two paediatric surgeons to rate 46 cohort studies that were presented at Canadian Association of Pediatric Surgeons annual meetings and later published in the *Journal of Pediatric Surgery*. The authors did not specify whether the surgeons received training in quality assessment. The ICC between surgeons, calculated on NOS total scores, was 0.94.

The lower inter-rater reliability of the novice student raters in this study, compared with the raters in the Oremus *et al*[19] and Al-Harbi *et al*[12] studies, may be explained by topic familiarity and similarity of expertise. The faculty raters in the Oremus *et al* study had previously worked on a systematic review of Alzheimer's disease medications and their expertise lay in two domains of epidemiology, that is, neuroepidemiology and pharmacoepidemiology. The paediatric surgeons in Al-Harbi *et al* may have possessed at least a general familiarity with the types of cohort studies conducted in their specialty. These characteristics may have predisposed the raters to adopt more uniform opinions on the questions contained in the Jadad and NOS. In contrast, the novice student raters in our study had for the most part not been exposed to systematic reviews and quality assessment in the past. Also, seven of these raters were recent entrants to graduate school, and they came from a variety of undergraduate backgrounds such as medicine, psychology and basic science.

## Limitations

Readers should exercise caution when generalising the results of our study to other types of raters. Reliability could differ according to raters' disciplines and levels of training. Reliability in our study also could have been affected by the specific training programme we gave to the students. Additionally, the 10 student raters in this study were a convenience sample that might not represent all raters with similar disciplines and training.

We did not compare the students' rankings with the rankings of more experienced raters (eg, faculty who conduct systematic reviews). Thus, we could not assess the relative differences in reliability between experienced raters and inexperienced students.

Reliability is also partly a function of the instruments used in the quality assessment. Indeed, instruments with many interpretive questions (eg, appropriateness of randomisation and double-blinding, representativeness of exposed cohort or adequacy of case definition) could have poor reliability, despite several phases of training.

Furthermore, the topic under study could influence reliability, as could certain methodological decisions related to the systematic review. For example, the systematic review of ECT and cognition, upon which we based this study, included 28 papers published prior to 1990. Since the style of reporting in older papers does not always facilitate quality assessment or data extraction, systematic reviews that include older papers could present challenges for maintaining acceptable levels of inter-rater and test–retest reliability.

## Conclusions

In conclusion, we asked a group of 10 novice students to rate the quality of 78 articles that contained data on cognitive impairment following the use of ECT to treat major depressive disorder. Overall, inter-rater reliability on the Jadad Scale and NOS was poor to fair and test–retest reliability was fair to excellent. We trained the raters prior to the quality assessment exercise yet inter-rater agreement was low for several questions that required a certain degree of interpretation to answer. This was especially so for the NOS and underscores an

inherent greater difficulty with assessing the quality of observational studies compared with RCTs.

In addition to standardised training prior to commencing quality assessment, a pilot rating phase may also be necessary to discuss scale questions that generate disagreement among novice student raters. This procedure could help the raters develop standardised interpretations to minimise disagreement.

While the Cochrane Collaboration has stated that quality scales and scale scores are inappropriate means of ascertaining study quality,[33] our results are relevant because many researchers continue to use the Jadad Scale and NOS in their systematic reviews. Indeed, our work suggests an area of future research. The Cochrane Collaboration has proposed a 'risk of bias' tool to assess the quality of RCTs.[33] The reliability of the risk of bias tool should be assessed in raters with different levels of experience.

**Author affiliations**
[1]McMaster Evidence-based Practice Centre, McMaster University, Hamilton, Ontario, Canada
[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[3]McMaster Integrative Neuroscience Discovery and Study (MINDS) Program, Hamilton, Ontario, Canada
[4]Department of Psychiatry and Behavioural Neuroscience, Hamilton, Ontario, Canada

**Acknowledgements** Special thanks to Eleanor Pullenayegum and Harry Shannon for their helpful comments on an earlier draft of this manuscript.

**Contributors** MO and CO conceived and designed the study. MO analysed the data. MO, CO, MCM, GBCH and the ECT & Cognition Systematic Review Team interpreted the data. MO drafted the manuscript. CO, MCM, GBCH and the ECT & Cognition Systematic Review Team critically revised the manuscript for important intellectual content. All authors approved the final version of the manuscript.

**Funding** This study did not receive funds from any sponsor. No person or organisation beyond the authors had any input in study design and the collection, analysis and interpretation of data and the writing of the article and the decision to submit it for publication.

**Competing interests** None.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data available.

## REFERENCES

1. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
2. Agency for Healthcare Research and Quality (AHRQ). *Systems to Rate the Strength of Scientific Evidence. Evidence Report/ Technology Assessment No. 47*. Rockville, MD: Agency for Healthcare Research and Quality, 2002.
3. Verhagen AP, de Vet HC, de Bie RA, *et al*. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54:651–4.
4. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988;138:697–703.
5. Wells GA, Shea B, O'Connell D, *et al*. *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-analyses*. Ottawa: Ottawa Hospital Research Institute. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (accessed 2 Apr 2012).
6. Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1–12.
7. Chalmers TC, Smith H Jr, Blackburn B, *et al*. A method for assessing the quality of a randomized control trial. *Control Clin Trials* 1981;2:31–49.
8. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0. [Updated March 2011]*. The Cochrane Collaboration, 2011. http://www.cochrane-handbook.org (accessed 23 Apr 2012).
9. Maher CG, Sherrington C, Herbert RD, *et al*. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther* 2003;83:713–21.
10. Kocsis JH, Gerber AJ, Milrod B, *et al*. A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Compr Psychiatry* 2010;51:319–24.
11. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377–84.
12. Al-Harbi K, Farrokhyar F, Mulla S, *et al*. Classification and appraisal of the level of clinical evidence of publications from the Canadian Association of Pediatric Surgeons for the past 10 years. *J Pediatr Surg* 2009;44:1013–17.
13. Berard A, Andreu N, Tetrault J, *et al*. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Ann Epidemiol* 2000;10:498–503.
14. Hartling L, Ospina M, Liang Y, *et al*. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.
15. Hartling L, Bond K, Vandermeer B, *et al*. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6: e17242.
16. Tooth L, Bennett S, McCluskey A, *et al*. Appraising the quality of randomized controlled trials: inter-rater reliability for the OTseeker evidence database. *J Eval Clin Pract* 2005;11:547–55.
17. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med* 2001;135:982–9.
18. Hartling L, Hamm M, Milne A, *et al*. *Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments*. Rockville, MD: Agency for Healthcare Research and Quality, 2012.
19. Oremus M, Wolfson C, Perrault A, *et al*. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. *Dement Geriatr Cogn Disord* 2001;12:232–6.
20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
21. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd edn. Hoboken, NJ: John Wiley & Sons, 2003.
22. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.
23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
24. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th edn. Oxford: Oxford University Press, 2008.
25. Fleiss J. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.
26. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *J Clin Epidemiol* 2010;63:1061–70.
27. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;36:666–76.
28. Lang S, Kleijnen J. Quality assessment tools for observational studies: lack of consensus. *Int J Evid Based Healthc* 2010;8:247.
29. Juurlink DN, Detsky AS. Kappa statistic. *CMAJ* 2005;173:16.
30. Agency for Healthcare Research and Quality (AHRQ). *Effective Health Care Program*. Rockville, MD: Agency for Healthcare Research and Quality. http://www.effectivehealthcare.ahrq.gov (accessed 2 Apr 2012).
31. Horton J, Vandermeer B, Hartling L, *et al*. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* 2010;63:289–98.
32. Haywood KL, Hargreaves J, White R, *et al*. Reviewing measures of outcome: reliability of data extraction. *J Eval Clin Pract* 2004;10:329–37.
33. Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.