BioMed Central

Research article

Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews

Beverley J Shea*[1,5], Jeremy M Grimshaw[†2], George A Wells[3], Maarten Boers[†4], Neil Andersson[5], Candyce Hamel[†5], Ashley C Porter[5], Peter Tugwell[2], David Moher[6] and Lex M Bouter[†1]

Address: [1]EMGO Institute, VU University Medical Center, Amsterdam, the Netherlands, [2]Institute of Population Health, Ottawa, Ontario, Canada, [3]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada, [4]Department of Clinical Epidemiology and Biostatistics. VU University Medical Center, Amsterdam, the Netherlands, [5]Community Information and Epidemiological Technologies (CIETcanada), Ottawa, Ontario, Canada and [6]University of Ottawa, Ottawa, Ontario, Canada

Email: Beverley J Shea* - bshea@ciet.org; Jeremy M Grimshaw - jgrimshaw@ohri.ca; George A Wells - gawells@ottawaheart.ca; Maarten Boers - mboers@vumc.nl; Neil Andersson - neil@ciet.org; Candyce Hamel - chamel@uottawa.ca; Ashley C Porter - ashleyclairereporter@hotmail.com; Peter Tugwell - ptugwell@uottawa.ca; David Moher - dmoher@uottawa.ca; Lex M Bouter - lm.bouter@dienst.vu.nl

* Corresponding author   †Equal contributors

Abstract

Background: Our objective was to develop an instrument to assess the methodological quality of systematic reviews, building upon previous tools, empirical evidence and expert consensus.

Methods: A 37-item assessment tool was formed by combining 1) the enhanced Overview Quality Assessment Questionnaire (OQAQ), 2) a checklist created by Sacks, and 3) three additional items recently judged to be of methodological importance. This tool was applied to 99 paper-based and 52 electronic systematic reviews. Exploratory factor analysis was used to identify underlying components. The results were considered by methodological experts using a nominal group technique aimed at item reduction and design of an assessment tool with face and content validity.

Results: The factor analysis identified 11 components. From each component, one item was selected by the nominal group. The resulting instrument was judged to have face and content validity.

Conclusion: A measurement tool for the 'assessment of multiple systematic reviews' (AMSTAR) was developed. The tool consists of 11 items and has good face and content validity for measuring the methodological quality of systematic reviews. Additional studies are needed with a focus on the reproducibility and construct validity of AMSTAR, before strong recommendations can be made on its use.

Background

It has been estimated that healthcare professionals attempting to keep abreast of their field would need to read an average of 17–20 original articles every day [1]. Increasingly, systematic reviews are being advocated as a way to keep up with current medical literature [2]. A well

conducted systematic review addresses a carefully formulated question by analyzing all available evidence. It employs an objective search of the literature, applying predetermined inclusion and exclusion criteria to the literature, critically appraising what is found to be relevant. It then extracts and synthesizes data from the available evidence base to formulate findings [3].

However, in spite of the care with which they are conducted, systematic reviews may differ in quality, and yield different answers to the same question [4]. As a result, users of systematic reviews should be critical and look carefully at the methodological quality of the available reviews [5].

A decade has passed since the initial development of tools to assess the quality of systematic reviews, such as those created by Oxman and Guyatt [6] and Sacks [7]. There are now more than 24 instruments available to assess the quality of systematic reviews [8]. Nevertheless, the majority of the available instruments are not widely used. Several are lengthy and include complicated instructions for their use. Furthermore, since their development, considerable empirical research has accumulated about potential sources of bias in systematic reviews. For example, recent methodological research has highlighted the potential importance of publication language and publication bias in systematic reviews [9-11].

Therefore, our goal was to develop a new instrument for assessing the methodological quality of systematic reviews by building upon empirical data collected with previously developed tools and utilizing expert opinion.

This goal was pursued by two study objectives. Our first objective was to assess a large sample of systematic reviews using an item pool drawn from two available instruments used to assess methodological quality, supplemented by additional items judged to be needed on the basis of recent publications. We used exploratory factor analysis to identify the underlying component structure. Our second objective was to build on the results of this factor analysis, by using experts in a nominal group technique (NGT) to reduce the items pool and to decide on a new assessment tool with face and content validity.

Methods
We designed a 37-item assessment tool that we developed by combining items from two available instruments: the enhanced Overview Quality Assessment Questionnaire (OQAQ) [8] containing 10 items and a checklist created by Sacks [7] containing 24 items. We supplemented this with three additional items based upon methodological advances in the field since the development of the original two instruments: *Language restriction*: Language restriction

in systematic reviews remains controversial. Some studies have suggested that systematic reviews that include only English language publications tend to overestimate effect sizes [10], whereas other studies suggest that such language restrictions may not do so [11]. An item was added to determine whether a language restriction was applied in selecting studies for the systematic review. *2) Publication bias*: Publication bias refers to the tendency for research with negative findings to get published less frequently, less prominently, or more slowly, and the tendency for research with positive findings to get published more than once. Publication bias has been identified as a major threat to the validity of systematic reviews. Empirical research suggests that publication bias is widespread, and that a variety of methods are now available to assess publication bias [12-19]. An item was added to determine whether the authors assessed the likelihood of publication bias. *3) Publication status* of studies suggests that published trials are generally larger and may show an overall greater treatment effect than studies published in the 'grey' literature [20]. The importance of including grey literature in all systematic reviews has been discussed [21]. The assessment of the inclusion of grey literature considers whether or not the authors reported searching for grey literature.

Objective 1
The 37-item assessment tool was used to appraise 99 paper-based reviews identified from a database of reviews and meta-analyses [22] and 52 Cochrane systematic reviews from the Cochrane Database of Systematic Reviews [9]. After the list of selected systematic reviews was generated, full copies of these were retrieved, copied, and masked to conceal author, institution, and journal. Reviews in languages other than English (i.e., French, German, and Portuguese) were translated into English with the assistance of colleagues before masking [23].

For each included systematic review, two reviewers independently assessed the methodological quality with the 37 items (CH, BS).

Statistical analyses and graphs displaying the results obtained were produced using SPSS version 13.0 for Windows. The 37 items were subjected to principal components analysis, and Varimax rotations were used to rotate the components. Items with low factor loadings of < 0.50 were removed.

Objective 2
We convened an international panel of eleven experts in the fields of methodological quality assessment and systematic reviews. The group was selected from three organizations involved both in the conduct of systematic reviews and in the assessment of methodological quality.

The group was made up of clinicians, methodologists and epidemiologists, and reviewers who were new to the field. Some individuals were previously involved in the Cochrane Collaboration, while a number were not. By examining the results of the factor analysis, they reflected critically on the components identified and decided on the items to be included in the new instrument. The nominal group process took place in San Francisco during a one day session.

We conducted the following NGT in order to achieve agreement. After delivery of an overview of the project and the planned process for the day, the panel reviewed the results of the factor analysis. The aim of the NGT was to structure interaction within the group. Firstly, each participant was asked to record his or her ideas independently and privately. The ideas were then listed in a round-robin format. One idea was collected from each individual in turn and listed in front of the group by the facilitator, and the process was continued until all ideas had been listed. Individuals then privately recorded their judgements. Subsequent discussions took place. The individual judgements were aggregated statistically to derive the group judgements. The nominal group was also asked to agree on a final label for each of the 11 components. A description was formulated for each of the items and a next-to-final instrument was assembled. This was circulated electronically to the group for a final round of fine tuning.

Results
Objective 1
The items were subjected to factor analysis, and only those items that loaded highly on one component (>.50) were retained. The described factor analysis made it possible to reduce the 37-item instrument to a shorter (29-item) instrument that measured 11 components (Table 1).

Objective 2
The nominal group discussed all 11 components (Table 1). The items most appropriate for the components (Table 2), were included in the draft instrument [also see Additional file 1]. The instrument is an 11-item questionnaire that asks reviewers to answer yes, no, can't answer or not applicable. A separate question on language was identified in the factor analysis as a significant issue, but the nominal group felt that the contradictory evidence in the literature warranted removing this item from the shortened item list and capturing it under the question on publication status.

Discussion
Strengths and Weaknesses
Our purpose was to help users of systematic reviews to critically appraise systematic reviews. Therefore, we set out with the goal of developing a new instrument for assessing

the methodological quality of systematic reviews, by building upon empirical data on previously developed tools, empirical evidence and utilizing expert opinion.

Because we had already created a dataset of 151 systematic reviews assessed using 37 completed items for each review, we were able to conduct a factor analysis as the first step in the creation of the new tool. A more commonly used approach would have been to harvest appropriate items from existing questionnaires. This method has been used extensively in the development of instruments for assessing the quality of both randomized and non-randomized studies of health care interventions [24-26]. The disadvantage of harvesting appropriate items from existing questionnaires is that it relies heavily on the validation of the source questionnaires [27]. Conducting a factor analysis made it possible to determine whether the measured dimensions could in principle be assessed using a smaller number of items.

Traditionally, factor analysis is divided into two types of analyses: exploratory and confirmatory. As its name indicates, exploratory factor analysis aims to discover the main constructs or dimensions of a concept by conducting a preliminary investigation of the correlations between all the identified variables. This process is also known as Principal Components Analysis (PCA). PCA has been recommended for use in test construction by Kline, as a means of condensing the correlation matrix, rather than as an aid to the interpretation of the factor-structure of a questionnaire [28]. Items with low factor loadings tend to be weakly correlated with other items, and therefore were removed. Various rotational strategies have also been proposed. The goal of all of them is to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others [29,30]. We used this approach because it is useful when a body of theory or principles has been established, but has not yet been operationalised into an evaluative framework [31].

The structured-discussion format employed in this project enabled all participants to contribute to the refining of the assessment tool. The nominal technique followed involved experts, discussion, and a consensus that was qualitative in nature. Consequently, it complemented the quantitative nature of factor analysis, and as a result the final tool had face and content validity as judged by the nominal consensus panel.

We recognize the need for further testing of AMSTAR. Additional studies are necessary with a focus on the reproducibility and construct validity of AMSTAR before strong recommendations can be made on its use. The Canadian Agency for Drugs and Technologies in Health (CADTH)

Table 1: Items identified through the factor analysis

	Original instrument (item no)	1	2	3	4	5	6	7	8	9	10	11
1 Protocol	Sacks	.58										
2 Literature Search	Sacks				.82							
3 List of Trials Analyzed	Sacks		.75									
4 Log of Rejected Trials	Sacks								.68			
5 Treatment Assignment	Sacks		.80									
6 Ranges of Patients	Sacks											
7 Range of Treatment	Sacks		.88									
8 Range of Diagnosis	Sacks		.80									
9 CombinabilityCriteria	Sacks									.88		
10 Measurement	Sacks				.57							
11 Selection Bias	Sacks	.85										
12 Data abstraction	Sacks	.50										
13 Inter-observer Agreement	Sacks	.65										
14 Sources of Support	Sacks								.64			
15 Statistical Methods	Sacks			.81								
16. Statistical Errors	Sacks											
17 Confidence Intervals	Sacks			.73								
18 Subgroup Analysis	Sacks											
19 Quality Assessment	Sacks					.77						
20 Varying Methods	Sacks					.63						
21 Publication Bias 1	Sacks						.77					
22 Caveats	Sacks											
23 Economic Impact	Sacks											.84
24 Language 1	Added to Sacks							.79				
25 Search Strategy	OQAQ (1)				.81							
26 Was the search comprehensive	OQAQ (2)											
27 Criteria used for deciding which studies to include	OQAQ (3)											
28 Was bias in the selection avoided	OQAQ (4)	.81										
29 Were the criteria used for assessing the validity reported?	OQAQ (5)					.75						
30 Was the validity of all studies referred to in the text assessed using appropriate criteria	OQAQ (6)				.53						.60	
31 Were the methods used to combine the finding of the relevant studies reported	OQAQ (7)											
32 Were the findings of the relevant studies combined appropriately	OQAQ (8)			.78								
33 Were the conclusions made by the author supported by the data	OQAQ (9)			.68								
34 Overall Summary	OQAQ (10)											
35. Publication Bias 2	Additional (1)						.80					
36 Publication Status	Additional (2)							.77				
37 Language 2	Additional (3)							.63				

Table 2: AMSTAR is a measurement tool created to assess the methodological quality of systematic reviews.

1. Was an 'a priori' design provided? The research question and inclusion criteria should be established before the conduct of the review.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
2. Was there duplicate study selection and data extraction? There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
3. Was a comprehensive literature search performed? At least two electronic sources should be searched. The report must include years and databases used (e.g. Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion? The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
5. Was a list of studies (included and excluded) provided? A list of included and excluded studies should be provided.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
6. Were the characteristics of the included studies provided? In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions and outcomes. The ranges of characteristics in all the studies analyzed e.g. age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
7. Was the scientific quality of the included studies assessed and documented? 'A priori' methods of assessment should be provided (e.g., for effectiveness studies if the author(s) chose to include only randomized, double-blind, placebo controlled studies, or allocation concealment as inclusion criteria); for other types of studies alternative items will be relevant.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
8. Was the scientific quality of the included studies used appropriately in formulating conclusions? The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
9. Were the methods used to combine the findings of studies appropriate? For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e. Chi-squared test for homogeneity, I^2). If heterogeneity exists a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e. is it sensible to combine?).	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
10. Was the likelihood of publication bias assessed? An assessment of publication bias should include a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test).	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable
11. Was the conflict of interest stated? Potential sources of support should be clearly acknowledged in both the systematic review and the included studies.	☐ Yes ☐ No ☐ Can't answer ☐ Not applicable

[32] undertook an independent assessment of available quality assessment criteria for systematic reviews. Feedback from CADTH reviewers has been very positive. Further preliminary experience suggests that AMSTAR has good reliability and convergent validity also suggesting that appraisers can apply it in a consistent way.

AMSTAR, if used widely after external validation, could also enable methodological research (i.e. meta-regression of item of AMSTAR and effect size of reviews). Our instrument is an attempt to achieve consensus amongst current mainstream opinions. Inevitably, new evidence will modify current thinking in some areas and at that point the AMSTAR will be updated. This is indeed likely to be the case with techniques to identify and quantify publication bias [33]. Although a number of alternative tests for publication bias exist, none has yet been validated [34].

Publication bias remains an area of contention amongst those who assess the quality of systematic reviews. It remains a research priority because it is unclear what the impact of publication bias is on making decisions in health care. We are aware of the 20 years of work that has gone in this area of research. This has given us some clear answers as to the effect publication bias may have on the overall results of estimating the impact of interventions.

AMSTAR will remain a living document and advances in empirical methodological research will be reflected in further improvements to the instrument.

Conclusion
A measurement tool for *assessment of multiple systematic reviews* (AMSTAR) was developed. The tool consists of 11 items and has good face and content validity for measuring the methodological quality of systematic reviews. Additional studies are needed with a focus on the reproducibility and construct validity of AMSTAR, before strong recommendations can be made on its use.

Competing interests
The author(s) declare that they have no competing interests.

Authors' contributions
BS designed and conducted the study and wrote the manuscript. MB, JG and LB participated in the design and coordination and assisted with writing the manuscript. CH and AP carried out the quality assessments and assisted with the writing. NA assisted in the conduct of the study and commented on earlier drafts. GW assisted in the design of the study and commented on the statistical analysis. PT helped with the design of the study and assisted with the nominal group process. All authors read and approved the final manuscript.

Additional material

Additional File 1
AMSTAR. AMSTAR is a measurement tool created to assess the methodological quality of systematic reviews.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-7-10-S1.doc]

References
1. Davidoff F, Haynes B, Sackett D, Smith R: **Evidence-based medicine: a new journal to help doctors identify the information they need.** *BMJ* 1995, **310:**1085-6.
2. Lau J, Ioannidis JPA, Schmid CH: **Summing up evidence: one answer is not always enough.** *Lancet* 1998, **351:**123-127.
3. **Systematic Review definition** [http://www.nlm.nih.gov/nichsr/hta101/ta101014.html]
4. Moher D, Soeken K, Sampson M, Campbell K, Ben Perot L, Berman B: **Assessing the quality of reports of systematic reviews in pediatric complementary and alternative medicine.** *BMC Pediatr* 2002, **2(2):**.
5. Jadad A, Moher M, Browman G, Booker L, Sigouin C, Fuentes M: **Systematic reviews and meta-analyses on treatment of asthma: critical evaluation.** *BMJ* 2000, **320:**537-540.
6. Oxman AD: **Checklists for review articles.** *BMJ* 1994, **309:**648-651.
7. Sacks H, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC: **Meta-analyses of randomized controlled trials.** *N Engl J Med* 1987, **316(8):**450-455.
8. Shea B, Dubé C, Moher D: **Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools.** In *Systematic Reviews in Health Care: Meta-analysis in context* Edited by: Egger M, Smith GD, Altman DG. London: BMJ books; 2001:122-139.
9. **The Cochrane Library.** *Volume 3*. Chichester, UK: John Wiley & Sons Ltd; 2004.
10. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G: **Language bias in randomised controlled trials published in English and German.** *Lancet* 1997, **350:**326-329.
11. Moher D, Pham B, Klassen T, Schulz K, Berlin J, Jadad A, Liberati A: **What contributions do languages other than English make to the results of meta-analyses?** *Journal of Clinical Epidemiology* 2000, **53:**964-972.
12. Pai M, McCulloch M, Colford J Jr, Bero LA: **Assessment ofPublication Bias in Systematic Reviews on HIV/AIDS.** [http://www.igh.org/Cochrane/pdfs/MSRI_workshop_talk_abstract.pdf].
13. Dickersin K: **The existence of publication bias and risk factors for its occurrence.** *JAMA* 1990, **263:**1385-1389.
14. Dickersin K: **How important is publication bias? A synthesis of available data.** *AIDS Educ Prev* 1997, **9:**15-21.
15. Phillips C: **Publication bias in situ.** *BMC Med Res Method* 2004, **4:**20.
16. Pham B, Platt R, McAuley L, Sampson M, Klassen T, Moher D: **Detecting and minimizing publication bias. A systematic**

review of methods. In *Technical Report* Thomas C. Chalmers Centre for Systematic Reviews, Ottawa, Canada; 2000.

17. Stern JM, Simes RJ: **Publication bias: evidence of delayed publication in a cohort study of clinical research projects.** *BMJ* 1997, **315:**640-5.

18. Sterne JAC, Gavaghan D, Egger M: **Publication and related bias in meta-analysis – power of statistical tests and prevalence in the literature.** *Journal of Clinical Epidemiology* 2000, **53:**1119-1129.

19. Sutton A, Duval S, Tweedie R, Abrams K, Jones D: **Empirical assessment of effect of publication bias on meta-analyses.** *BMJ* 2000, **320:**1574-1577.

20. Hopewell S, McDonald S, Clarke M, Egger M: **Grey literature in meta-analyses of randomized trials of health care interventions.** In *The Cochrane Library Volume 1*. Chichester, UK: John Wiley & Sons, Ltd; 2004.

21. McAuley L, Pham B, Tugwell P, Moher D: **Does the inclusion of grey literature influence the estimates of intervention effectiveness reported in meta-analyses?** *Lancet* 2000, **356:**1228-1231.

22. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP: **Assessing the quality of randomized controlledtrials: implications for the conduct of meta-analyses.** *Health Technol Assess* 1999, **3(12):**i-iv. 1–98

23. Shea B: **Assessing the quality of reporting meta-analyses of randomized controlled trials.** In *MSc thesis* University of Ottawa, Department of Epidemiology and Community Medicine; 1999.

24. Downs S, 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:**377-384.

25. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ: **Assessing the quality of randomized clinical trials: is blinding necessary?** *Control Clin Trials* 1996, **17:**1-12.

26. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG: **The Delphi list: a consensus list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus.** *J Clin Epid* 1998, **51:**1235-41.

27. Cluzeau FA: **Development and application of an appraisal instrument for assessing the methodological quality of clinical practice guidelines in the United Kingdom.** London, England: St.George's Hospital Medical School, University of London; 2001.

28. Kline P: *An Easy Guide to factor Analysis* London: Routledge; 1994.

29. Norman GR, Streiner DL: *Biostatistics: The Bare Essentials* 2nd edition. St. Louis: Mosby; 2000.

30. Streiner DL, Norman GR: *Health Measurement Scales: A Practical Guide to their Development and Use* 3rd edition. UK: Oxford University Press; 2003.

31. McDowell I, Jenkinson C: **Development standards for health measures.** *Journal of health Services Research and Policy* 1996, **1(4):**238-246.

32. **Canadian Agency for Drugs and Technologies in Health** [http://www.cadth.ca]

33. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I: **The case of the misleading funnel plot.** *BMJ* 2006, **333:**597-600.

34. Rothstein HR, Sutton AJ, Borenstein M, eds: *Publication bias in meta-analysis: prevention, assessment and adjustments* Sussex: John Wiley and Sons; 2005.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/7/10/prepub