

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

Beverley Shea<sup>1,5§</sup>, Jeremy M Grimshaw<sup>2\*</sup>, George A Wells<sup>3</sup>, Maarten Boers<sup>4\*</sup>, Candyce Hamel<sup>5\*</sup>, Neil Andersson<sup>5</sup>, Ashley Porter<sup>5</sup>, David Moher<sup>6</sup>, Peter Tugwell<sup>2</sup>, Lex M Bouter<sup>1\*</sup>

<sup>1</sup> EMGO Institute, VU University Medical Center, Amsterdam, the Netherlands

<sup>2</sup>Institute of Population Health, Ottawa, Ontario, Canada

<sup>3</sup> Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada

<sup>4</sup> Department of Clinical Epidemiology and Biostatistics. VU University Medical Center, Amsterdam, the Netherlands

<sup>5</sup> Community Information and Epidemiological Technologies (CIETcanada), Ottawa, Ontario, Canada

<sup>6</sup> University of Ottawa, Ottawa, Ontario, Canada

\*These authors contributed equally to this work

§Corresponding author

Email address:

BS: [bevshea@ciet.org](mailto:bevshea@ciet.org)

JG: [jgrimshaw@ohri.ca](mailto:jgrimshaw@ohri.ca)

GW: [gawells@ottawaheart.ca](mailto:gawells@ottawaheart.ca)

MB: [mboers56@xs4all.nl](mailto:mboers56@xs4all.nl)

CH: [chamel@ciet.org](mailto:chamel@ciet.org)

NA: [CIETinter@compuserve.com](mailto:CIETinter@compuserve.com)

AP: [ashleyclaireporter@hotmail.com](mailto:ashleyclaireporter@hotmail.com)

DM: [dmoher@uottawa.ca](mailto:dmoher@uottawa.ca)

PT: [ptugwell@uottawa.ca](mailto:ptugwell@uottawa.ca)

LB: [lm.bouter@planet.nl](mailto:lm.bouter@planet.nl)

## **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) scale, 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.

### **Conclusions**

A measurement tool for the 'assessment of 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 on its use can be made.

## **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, and employs an objective search of the literature, applying predetermined inclusion and exclusion criteria to the literature, critically appraising what is found to be relevant, and extracting and synthesizing 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 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 its 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 [9,10, 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 publication. 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: 1)

*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 often. 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*: Studies suggest that trials published in scientific journals are generally larger and may show a greater overall treatment effect than trials available from the grey literature [20]. The importance of including grey literature in systematic reviews has been addressed [21]. An item was added to ascertain whether 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 new reviewers 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 idea independently and privately. The ideas were then listed in a round-robin format, that is 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 were 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-items) 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 as outlined in Table 2, were included in the draft instrument . 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 a goal to develop 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, 25, 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 on its use can be made. The Canadian Agency for Drugs and Technologies in Health (CADATH) [32], undertook an independent assessment of available quality assessment criteria for systematic reviews. Feedback from CADATH 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).

## **Conclusions**

A measurement tool for *assessment of 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 on its use can be made.

## **Competing interests**

The authors declare that they have no competing interests

## **Authors' contributions**

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

## **Acknowledgements**

The authors would like to thank the following members of the San Francisco Nominal Group for their participation: Gail Kennedy, George Rutherford, George Wells, Peter Tugwell, Maarten Boers, Madhukar Pai, Kirby Lee, Maria Suarez-Almazor, Kaveh Shojania, Tara Horvath, and Kathryn McDonald. Thank you to Vivian Robinson, Adriene Steven and David Henry for their helpful comments.

The authors also thank Drs. Andy Oxman, Gord Guyatt and Henry Sacks for their methodological instruments and for providing additional information.

## 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**(6995):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.* Egger M, Smith GD, Altman DG, Eds. London: BMJ books; 2001:122-139.
9. The Cochrane Library. Chichester, UK: John Wiley & Sons Ltd, **3**.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 of Publication Bias in Systematic Reviews on HIV/AIDS.**  
[[http://www.igh.org/Cochrane/pdfs/MSRI\\_workshop\\_talk\\_abstract.pdf](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. 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.
16. Pham B, Platt R, McAuley L, Sampson M, Klassen T, Moher D: Detecting and minimizing publication bias. A systematic review of methods. 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 J.A.C., 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*, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.
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 controlled trials: 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.** 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 H.C.W., de Bie RA, Kessels A.G.H., 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*, 2<sup>nd</sup> 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 [[www.cadth.ca](http://www.cadth.ca)].

## Tables

**Table 1** – Results from the factor analysis (37 items reduced to 29, and 11 components)

**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. Combinability Criteria	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	OQAQ (6)				.53						.60	

all studies referred to in the text assessed using appropriate criteria												
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** – A measurement tool created to assess the methodological quality of systematic reviews.

**Table 2: Assessment of Systematic Reviews (AMSTAR)**

<p><b>1. Was an ‘a priori’ design provided?</b> The research question and inclusion criteria should be established before the conduct of the review.</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Can’t answer <input type="checkbox"/> Not applicable</p>
<p><b>2. Was there duplicate study selection and data extraction?</b> There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Can’t answer <input type="checkbox"/> Not applicable</p>
<p><b>3. Was a comprehensive literature search performed?</b> 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.</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Can’t answer <input type="checkbox"/> Not applicable</p>
<p><b>4. Were reports included irrespective of their status of publication (i.e. grey literature was not used as exclusion criterion)?</b> 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.</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Can’t answer <input type="checkbox"/> Not applicable</p>
<p><b>5. Was a list of studies (included and excluded) provided?</b> A list of included and excluded studies should be provided either within the article or on a designated website. Included studies are only those incorporated in the systematic review and/or the meta-analysis. Specifically, for excluded studies the authors should list those studies seriously considered, meaning selected on the basis of title or abstract, but after studying the full text rejected for the review.</p>	<p><input type="checkbox"/> Yes <input type="checkbox"/> No <input type="checkbox"/> Can’t answer <input type="checkbox"/> Not applicable</p>

**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 the 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) and/or statistical tests (e.g., Egger regression test).

- Yes
- No
- Can't answer
- Not applicable

**11. Were potential conflicts of interest stated?**

Sources of support and other potential conflicts should be clearly acknowledged both in terms of the systematic review and the included studies.

- Yes
- No
- Can't answer
- Not applicable