Best-worst scaling identified adequate statistical methods and literature search as the most important items of AMSTAR2 (A measurement tool to assess systematic reviews)

Journal of Clinical Epidemiology 128 (2020) 74−82

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Best-worst scaling identified adequate statistical methods and literature search as the most important items of AMSTAR2 (A measurement tool to assess systematic reviews)

Victoria Leclercq[a,b,*], Mickaël Hiligsmann[b], Gianni Parisi[a], Charlotte Beaudart[a], Ezio Tirelli[c], Olivier Bruyère[a]

[a]*Division of Public Health, Epidemiology and Health Economics, University of Liège, Liège, Belgium. WHO Collaborating Center for Public Health aspects of musculo-skeletal health and ageing.*

[b]*Department of Health Services Research, Care and Public Health Research Institute (CAPHRI), Maastricht, the Netherlands*

[c]*Department of Psychology, University of Liège, Liège, Belgium*

Accepted 18 August 2020; Published online 20 August 2020

Abstract

Objective: To assess the relative importance of A MeaSurement Tool to Assess systematic Reviews 2 (AMSTAR2) items.

Study Design and Setting: A best-worst scaling object case was conducted among a sample of experts in the field of systematic reviews (SRs) and meta-analyses (MAs). Respondents were asked in a series of 15 choice tasks to choose the most and the least important item from a set of four items from the master list, which included the 16 AMSTAR2 items. Hierarchical Bayes analysis was used to generate the relative importance score for each item.

Results: The most important items highlighted by our 242 experts to conduct overview of reviews and critically assess SRs/MAs were the appropriateness of statistical analyses and adequacy of the literature search, followed by items regarding the assessment of risk of bias, the research protocol, and the assessment of heterogeneity (relative importance score >6.5). Items related to funding sources and the assessment of study selection and data extraction in duplicate were rated as least important.

Conclusion: Although all AMSTAR2 items can be considered as important, our results highlighted the importance of keeping the two items (the appropriateness of statistical analyses and the adequacy of the literature search) among the critical items proposed by AMSTAR2 to critically appraise SRs/MAs.

Keywords: Systematic review; Meta-analysis; AMSTAR2; Best-worst scaling; Meta-research; Expert survey

1. Introduction

Systematic reviews (SRs) and meta-analyses (MAs) are increasingly important to summarize and synthetize information about a specific research question and are therefore widely used in the literature and for clinical and policy decision-making [1]. Many authors have highlighted methodological weaknesses in SRs/MAs [2]. To provide reliable conclusions, a robust methodology for conducting SRs/MAs is therefore essential [3]. To critically appraise SRs/MAs and to enable users to carry out assessments of the conduct of SRs/MAs, several guidelines have been developed, including the A Measurement Tool to Assess systematic Reviews 2 (AMSTAR2) tool [4]. AMSTAR2 is a revision of the original AMSTAR instrument [5] developed by Shea et al. in 2007 [4]. Since its first publication, AMSTAR has been one of the most widely used instruments

Funding: This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors. VL is supported by a non-FRIA grant from the University of Liège.

Declarations of interest: None.

Authors' contribution: V.L., MH, CB, and OB conceived the project and developed the study hypotheses and the protocol. VL, MH, and GP were responsible for data collection, data management, and data analyses. VL, MH, and OB performed the statistical analysis and interpreted the data. VL wrote the drafts of the article under the supervision of MH, OB, ET, and CB. All authors have read, reviewed, and approved the final manuscript.

* Corresponding author. Division of Public Health, Epidemiology and Health Economics, University of Liège, CHU - Sart Tilman, Quartier Hôpital, Avenue Hippocrate 13 (Bât. B23), 4000 Liège, Belgium. Tel.: +32 43 66 25 20; fax: +32 43 66 28 12.

E-mail address: victoria.leclercq@uliege.be (V. Leclercq).

https://doi.org/10.1016/j.jclinepi.2020.08.011

What is new?

Key findings

- The items from AMSTAR2 regarding the appropriateness of statistical analyses and the adequacy of the literature search were considered the most important to critically assess SR/MA. The items regarding the assessment of risk of bias, the research protocol, and the assessment of heterogeneity were also very important to critically assess SR/MA.
- The least important items in the assessment of SR/MA were related to the description of included studies, the presence of funding sources and conflict of interest, the assessment of publication bias, the explanation of study designs, the presence of excluded studies lists, the study selection, and data extraction in duplicate.
- Subgroup analyses related to expert knowledge of AMSTAR2, difficulty completing the survey, and confidence in the results provided did not reveal major differences in the relative importance of AMSTAR2 items.

What this adds to what was known?

- This study prioritizes the relative importance of 16 items of AMSTAR2 in a sample of 242 experts.

What is the implication and what should change now?

- The findings of this study may provide evidence to weigh the different AMSTAR2 items when critically appraising SR/MA.
- The two items considered as the most important ones (i.e., literature search and statistical analyses) could be reformulated to provide extra details to appropriately judge their proper applicability.

to assess SRs/MAs. The relevance of the 11 original items was confirmed, and some were refined, leading to the AMSTAR2 tool in 2017 that comprises 16 items [4]. Interestingly, AMSTAR2 proposes a classification by identifying seven critical and 9 noncritical items to determine the rating of overall confidence in SR/MA results (high, moderate, low, or critically low). The suggested rating of overall confidence is based on the presence or absence of critical domains. When an SR/MA presents at least one critical weakness, the quality is considered low or critically low. Because of the tool's novelty, evidence for its psychometric properties is sparse [6−9]. The choice of the critical items that lead to the classification of SR/MA quality was made based on the advice of a small group of experts and can raise some questions. Because the classification of items can significantly influence the way SRs and MAs are appraised, it seems crucial to identify this classification in the most objective way. For this reason, the authors of AMSTAR2 encouraged to adjust their selection of critical items by the context of the application of the critical appraisal [4].

One potential adjustment could be to identify the items considered most important by a larger number of experts. To ensure involvement of experts in the improvement of SR/MA methodology, the use of preference studies and, more specifically, the best-worst scaling (BWS) approach may help to measure preferences and prioritize the items by asking respondents to select them in different choice sets [10,11]. This method, developed by Finn and Louviere, is an increasingly popular conjoint analysis technique in health care research [12−14] and is very useful for eliciting preferences and rankings for a large set of items.

This study aimed to assess the relative importance of the 16 AMSTAR2 items. The results could be helpful for the people who conduct overviews of reviews and need to apply a critical appraisal of SRs/MAs using AMSTAR2 [15]. For this purpose, a BWS approach was used among many experts in the field of SRs/MAs. In addition, we aimed to assess whether the perception of the importance of items differs as per some experts' characteristics.

2. Methods

2.1. Registration and protocol

The study protocol is available on the platform Open Science Framework: https://osf.io/6fr48/.

2.2. Study design & questionnaire

To elicit experts' opinions about the relative importance of the 16 items of the AMSTAR2, a survey including a BWS object case (also named BWS case 1) was conducted [14]. In the BWS approach, a person faces choices among scenarios with three or more items and identifies the most and least preferred item in each scenario [11]. In this study, participants received a series of 15 scenarios (i.e., choice tasks), each with four items derived from the 16 AMSTAR2 items (see Table 1), and in each scenario were asked to identify which item should receive the highest importance and which should receive the lowest importance, to critically appraise methodological characteristics of SRs/MAs.

Sawtooth Software's SSI Web platform was used to design the survey. Four BWS versions were generated, each consisting of 15 unique choice tasks with four items per choice task, to obtain the most efficient, fractional design. The design of the BWS was characterized by orthogonality (items are paired approximately an equal number of times), minimal overlap (minimizing the number of times each item appears within the same set across the design), potential balance (items appear approximately an equal number

Table 1. AMSTAR 2 items

Item 1	Did the research questions and inclusion criteria for the review include the components of PICO?
Item 2	Did the report of the review contain an explicit statement that the review methods were established before the conduct of the review and did the report justify any significant deviations from the protocol?
Item 3	Did the review authors explain their selection of the study designs for inclusion in the review?
Item 4	Did the review authors use a comprehensive literature search strategy?
Item 5	Did the review authors perform study selection in duplicate?
Item 6	Did the review authors perform data extraction in duplicate?
Item 7	Did the review authors provide a list of excluded studies and justify the exclusions?
Item 8	Did the review authors describe the included studies in adequate detail?
Item 9	Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?
Item 10	Did the review authors report on the sources of funding for the studies included in the review?
Item 11	If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?
Item 12	If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?
Item 13	Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review?
Item 14	Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?
Item 15	If they performed quantitative synthesis, did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?
Item 16	Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?

of times in each position), connectivity (items are directly and indirectly linked), and stability (four different versions of the questionnaire are used to increase variation) [16]. Overall, each item was presented approximately 15 times, was combined at least twice (mean of 3 ± 0.4) with other items and appeared approximately 3 times in each position (mean of 3.75 ± 0.5) in the choice tasks. An example of a BWS choice task is presented in Table 2. The questionnaire, which was developed online using Qualtrics, began with some background characteristics (employment status, experience in method, years of experience, type of expertise, and level of expertise or knowledge of AMSTAR2). Then, the Qualtrics systems randomly assigned each participant to one of the four versions of the questionnaire. Finally, the survey ended with some general questions (confidence in the results, difficulty with the survey, and additional comments). An e-mail invitation to participate in the survey was sent to all identified experts between December 3, 2019 and January 12, 2020 using Qualtrics software. If there was no response, two reminders were sent. The questionnaire was pretested among 13 experts in SRs/MAs and/or BWS from the University of Liège and Maastricht University. Only minor textual modifications were made. The study was fully anonymous; no personal data were collected.

2.3. Participants

Experts in SRs/MAs were recruited from various public sources (e.g., scientific articles) using a judgment sampling strategy. Respondents were eligible to complete our survey and considered experts if they confirmed they belonged to one of the following categories: authors of metaresearch studies on SRs/MAs (i.e., all authors who have explored the methodological quality and/or the reporting of SRs/MAs in metaresearch studies identified through a scoping review performed by Page et al. [1] and its update in September 2019 available in the research protocol), members of Cochrane Methods Groups (i.e., experts in the Cochrane Methods Groups found on the Cochrane website), first authors of Cochrane SR/MA (i.e., all first authors of SRs/MAs published between January 2018 and August 2019), and any other additional experts who had experience in SRs/MAs who were not identified in the first three categories (i.e., authors of checklists, tools or guidelines about SRs/MAs (e.g., PRISMA), or authors of a theoretical book about SRs/MAs). All bibliographies of previously identified articles were checked to identify any other potentially relevant experts. Moreover, snowball sampling was used by inviting all respondents to provide the e-mail address of someone who could be interested in completing the survey and competent to do so. As the literature provides no guidance regarding the minimal sample size for BWS, with previous sample sizes ranging from 15 to 803 participants [10], we aimed to recruit as many participants as possible.

2.4. Analyses

Only fully completed surveys were included in the analyses. Descriptive statistics were used to present the

Table 2. Example of a BWS choice task

Among the following four items from AMSTAR 2, please indicate the most important and the least important for you:	
Least important	**Most important**
Did the research questions and inclusion criteria for the review include the components of PICO?	
Did the review authors report on the sources of funding for the studies included in the review?	
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	
Did the review authors use a comprehensive literature search strategy?	

Abbreviation: PICO, population intervention comparaison outcome.

background characteristics of the respondents. Frequency and percentage values for categorical items and mean ± SD values for continuous items were reported. BWS data were analyzed using a hierarchical Bayes model using a multinomial logit procedure [14], using Sawtooth SSI Web version 8.2.0. The mean relative importance score (RIS) with its 95% confidence interval was estimated for each item. This score represents the item's relative importance to critically assess SRs/MAs. The RIS for each individual summed to 100 [12,17]. We then ordered all these RIS from the highest to lowest to obtain a ranking of all items (from the item with the highest RIS to the item with the lowest RIS). The higher the RIS score was, the higher the relative importance of an item. Items with a score of 6.25 were regarded as of average importance (100 points divided by 16 items). In the lack of consensus on the minimally important difference in RIS, if the confidence intervals of two consecutively ranked items did not overlap, we considered them to be of different importance in the critical assessment of SRs/MAs [17]. The internal consistency for each respondent was checked based on the individual's fit statistics (ranging from 0 to 1). If responses had a fit statistic below 0.336 [18], they were omitted from the analyses because they would suggest random responses to the choice tasks. In addition to the hierarchical Bayes method, a best-worst count analysis was conducted to check the sensitivity of the result. This means that for each factor, the frequency of the best choice minus the frequency of the worst score was calculated. To adjust the score, this best-worst score was divided by the frequency of occurrence of each item across the four versions of the questionnaire.

To assess the impact of participant characteristics, subgroup analyses were conducted. None of the quantitative variables followed a normal distribution. First, a Kruskal−Wallis test was used to evaluate differences among experts regarding their knowledge with the AMSTAR2 tool (i.e., "I know and I use AMSTAR 2," I know but I have not yet used AMSTAR2," "No, I don't know AMSTAR2"). Second, the Mann−Whitney test was conducted for continuous data to evaluate differences among experts on the level of expertise, the difficulty with the survey, and the confidence in the results as per the RIS. A P-value ≤ 0.05 was considered statistically significant for the analysis. A Bonferroni-adjusted alpha of 0.003 (for 16 comparisons) was used to assess whether differences in the RIS of items between subgroups were statistically significant using SAS 9.4.

3. Results

3.1. Population

Of the 1,777 experts contacted by e-mail, 305 started the survey, and 245 completed the survey between December 3, 2019 and January 12, 2020. The overall fit statistic (0.61 ± 0.1) was considered good, but three respondents had a fit statistic lower than 0.336 and were excluded from the analyses. Table 3 presents an overview of the characteristics of the 242 experts included. Most experts had an academic position (57%), and they have experience with both SR and MA methods (65%). The majority of the respondents who completed the survey were authors of an SR/MA work (93%) and/or were Cochrane members (61%). The AMSTAR2 tool was known by 76% of the experts, and 39% had already used it. The years of experience in SR/MA varied between the experts, but most of them (81%) had worked in SRs/MAs for at least 6 years. Overall, experts self-reported a mean level of expertise of 5.5 on a 7-point Likert scale. They also reported a mean confidence in their results of 4.87 and a mean level of difficulty completing the survey of 3.8 (over a 7-point Likert scale).

3.2. Importance of AMSTAR2 items

Table 4 presents the RIS of AMSTAR2 items for all participants and the RIS for subgroup analyses based on the knowledge of AMSTAR2 by the respondents. Fig. 1 shows a graphical representation of the overall results. The ranking of AMSTAR2 items, as per the mean RIS, showed that statistical analyses (item 11, RIS 12.19, 95% CI 11.6−12.8) and a comprehensive literature search strategy (item 4, RIS 11.79, 95% CI 11.1−12.5) were the most important items. The following five important items with

Table 3. Experts characteristics (n = 242)

Characteristics	*N* (%)
Employment status	
PhD student	18 (7.4)
Postdoctoral researcher	38 (15.7)
Academic	138 (57.0)
Private companies	5 (2.0)
Clinical practitioner	16 (6.6)
Government employee	6 (2.5)
Other (consultant, librarian, retired, etc.)	21 (8.7)
Method experience	
Systematic review	61 (25.2)
Meta-analysis	23 (9.5)
Systematic review and meta-analysis	158 (65.3)
Experience time in SRs/MAs	
0–5 yr	45 (18.6)
6–10 yr	85 (35.1)
11–15 yr	45 (18.6)
16–20 yr	36 (14.9)
21–30 yr	22 (9.1)
>30 yr	9 (3.7)
Type of expertise[a]	
Author of SRs/MAs	225 (93.0)
Cochrane member	147 (60.7)
Author of metaresearch on SRs/MAs	96 (39.7)
Author of guideline or tools on SRs/MAs	88 (36.4)
Methodologist of SRs/MAs	115 (47.5)
Other (advisor, librarian, etc.)	5 (2.0)
Level of expertise, mean ± SD (from 0 = very low expertise to 7 = excellent expertise)	5.43 ± 1.1
Knowledge of the AMSTAR 2	
I know and I use AMSTAR2	94 (38.8)
I know, but I have not yet used AMSTAR2	90 (37.2)
No, I don't know AMSTAR2	58 (24.0)
Confidence in the result, mean ± SD (from 0 = very low confidence to 7 = excellent confidence)	4.87 ± 1.22
Difficulty with the survey, mean ± SD (from 0 = very low difficulty to 7 = high difficulty)	3.84 ± 1.72

[a] The respondents could choose more than one type of expertise.

a mean RIS higher than 6.25 were the "impact of the risk of bias (RoB)" (item 12, RIS 10.24, 95% CI 9.6–10.8), "RoB assessment" (item 9, RIS 10.13, 95% CI 9.6–10.7), "interpretation of the RoB" (item 13, RIS 8.79, 95% CI 8.2–9.4), "research protocol cited" (item 2, RIS 7.8, 95% CI 7.0–8.6), and "explanation of heterogeneity" (item 14, RIS 7.3, 95% CI 6.6–7.7). The following items were least important in the assessment of SRs/MAs, with no 95% CI overlap, including "research question" (item 1), "description of included studies" (item 8), "conflict of interest of SR/MA authors" (item 16), "publication bias" (item 15), "study designs" (item 3), "list of excluded studies" (item 7), "study selection" (item 5), "funding sources of included studies" (item 10), and "data extraction" (item 6). The results of the count analyses showed similar results (Table 1 in the supplementary file A).

3.3. Comparison of experts regarding their knowledge and use of AMSTAR2

Subgroup analysis found few significant differences in the RIS between the three groups of experts regarding their knowledge and use of AMSTAR2 (Table 4). The RIS of all RoB-related items (items 9, 12, and 13) differed between groups. Indeed, the experts who have knowledge of and use AMSTAR2 give more importance to the "RoB

assessment'' (item 9, RIS 11.53, 95% CI 10.8−12.3), ''impact of the RoB'' (item 12, RIS 11.77, 95% CI 10.9−12.7), and ''interpretation of the RoB'' (item 13, RIS 10.43, 95% CI 9.5−11.3) than the two other groups. However, experts who do not use the tool and have no knowledge of AMSTAR2 give more importance to ''study designs'' (item 3, RIS 5.15, 95% CI 3.9−6.4), ''list of excluded studies'' (item 7, RIS 4.36, 95% CI 3.2−5.5), and ''publication bias'' (item 15, RIS 5.24, 95% CI 3.9−6.6). Nevertheless, the ranking of the items did not change. Finally, other subgroup analyses regarding the difficulty of completing the study, confidence in the results, and the level of expertise among experts did not influence the ranking of items (Table 2 in the supplementary file A).

4. Discussion

In 2017, the AMSTAR2 tool proposed a new classification to determine the overall confidence in SR/MA results [4]. However, the tool was not focused on the weighted importance of each item, and authors were encouraged to adjust the selection of critical items by the context of the application of the critical appraisal. With the aim to add to the knowledge of AMSTAR2, as suggested by the authors of the tool, this study investigated the relative importance of AMSTAR2 items with a large panel of experts, using a BWS approach.

In-line with the seven critical items proposed by AMSTAR2 [4], five items were also considered important in our study: the presence of a research protocol, the adequacy of the literature search, the assessment of the RoB in individual studies, the appropriateness of meta-analytical methods, and the interpretation of the RoB in the results. As expected, the two most important items were the adequacy of the literature search and the appropriateness of the statistical method. These two items are the cornerstone of SR/MA but may also be the two most difficult items to apply and assess with a low to moderate reliability [7,8]. Recently, Page et al. [19] highlighted the need to provide more specific statistical guidance in SRs/MAs. For example, authors should provide an explication for their choice of meta-analysis model (e.g., clinical justification) used, subgroup analysis (e.g., choice of covariate) performed, and sensitivity analysis (e.g., one-way sensitivity) conducted. Regarding the components of item four in the literature search, recommendations from studies are not sufficiently clear. For example, AMSTAR2 recommended searching at least 2 databases and providing key words and/or a search strategy. However, to check the reproducibility and comprehensiveness of the search strategy, much more information in addition to key words would be needed. These 2 items could be considered as superficial to really be able to critically assess SRs/MAs. Further research is needed to better define the content of these items to apply them more easily and improve their reliability. Otherwise, these two aspects have relatively little importance in AMSTAR2 with only one item each, in comparison with the RoB assessment in individual studies. Indeed, three items are dedicated to the RoB assessment (i.e., the use of an appropriate tool, the impact of the RoB, and the interpretation of the results). Although the RoB assessment in SRs/MAs is highly relevant, the frequency of its occurrence (three times in the tool) may lead to an overestimation of its importance in comparison with the adequacy of statistical methods and literature searches.

Our results also showed that some items were rated by the experts as least important to critically appraise SR/MA, in comparison with the other items. However, this does not mean that these items are not important at all to conduct a good SRs/MAs. A BWS does not aim to assess the importance of each item separately, but rather evaluate the relative importance of items. In comparison with AMSTAR2 items, two critical items (i.e., assessment of publication bias and justification for excluding individual studies) are less important in our results, coming in 11th and 13th place, respectively. One possible explanation for the least importance of publication bias may be that this item is related to the literature search and, more specifically, to the search for relevant unpublished literature. Publication bias is difficult to assess because even when a researcher sets out to find all potentially eligible unpublished articles, the possibility remains that some studies meeting the inclusion criteria will be missed [20]. The lower importance of the full report of excluded studies, a process mainly used in the Cochrane review [21], may be explained by the fact that the non-Cochrane reviewer has more restrictions (i.e., the number of words or number of appendices) than the Cochrane reviewer for the publication of their SRs/MAs. In addition, it is possible that the authors choose to list the reasons for exclusions in the flowchart, as recommended by PRISMA checklist [22], instead of providing a detailed table of the reasons for exclusions listed by article. Surprisingly, items related to study selection and data extraction are also rated as least important. A possible explanation for these results may be the fact that these two practices in the conduct of SRs/MAs are well known and easier to apply than other more technical items (e.g., statistical methods). AMSTAR2 is particularly interested in carrying out these steps in duplicate; however, the explanation of conflict management between reviewers could be even more important. Recently, Gartlehner and his team [23] highlighted that single-reviewer abstract screening misses approximately 13% of relevant studies and could not be appropriate for the SR/MA process. Once again, few studies document these steps for SR/MA practices [24].

To the best of our knowledge, this is the first study investigating the relative importance of AMSTAR2 items

Table 4. Overall relative importance score of AMSTAR 2 items to critically assess SRs/MAs and the results of the subgroup analysis with the Kruskal-Wallis test for the comparison between groups

Items	Overall ($N = 242$)	
	Rank	RIS (95% CI)
1. Did the research questions and inclusion criteria for the review include the components of PICO	8	6.03 (5.3–6.8)
2. Did the report of the review contain an explicit statement that the review methods were established before the conduct of the review and did the report justify any significant deviations from the protocol?	6	7.80 (7.0–8.6)
3. Did the review authors explain their selection of the study designs for inclusion in the review?	12	3.39 (2.9–3.9)
4. Did the review authors use a comprehensive literature search strategy?	2	11.79 (11.1–12.5)
5. Did the review authors perform study selection in duplicate?	14	2.31 (1.9–2.7)
6. Did the review authors perform data extraction in duplicate?	16	1.74 (1.4–2.1)
7. Did the review authors provide a list of excluded studies and justify the exclusions?	13	2.8 (2.3–3.2)
8. Did the review authors describe the included studies in adequate detail?	9	5.08 (4.5–5.7)
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	4	10.13 (9.6–10.7)
10. Did the review authors report on the sources of funding for the studies included in the review?	15	1.86 (1.4–2.3)
11. If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?	1	12.19 (11.6–12.8)
12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	3	10.24 (9.6–10.8)
13. Did the review authors account for RoB in individual studies when interpreting/discussing the results of the review?	5	8.79 (8.2–9.4)
14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	7	7.13 (6.6–7.7)
15. If they performed quantitative synthesis, did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	11	3.99 (3.5–4.5)
16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	10	4.77 (4.1–5.5)

Abbreviations: RIS, relative importance score; SD, standard deviation.

[a] *P*-value for the Kruskal-Wallis test conducted to compare the RIS of items between the 3 groups of the knowledge of AMSTAR2.

[b] Items statistically significant with the Bonferroni correction for multiple testing ($P \leq 0.003$). Interpretation of Eta^2: 0.01–<0.06 is a small effect size, 0.06–<0.14 is a moderate effect size, and >0.14 is a large effect size.

to critically appraise SRs/MAs. Our results provide robust information by identifying very important AMSTAR2 items confirmed by sensitivity and subgroup analyses. However, some limitations of this study are worth noting. Although the number of respondents is high, the sample may not be representative of the entire population of SR/MA experts. The survey was completed by only 14% (245/1777) of the contacted experts, and some potential experts may have been missed. Another limitation is that participants could not be familiar with the BWS tasks, and some of them may have had difficulties discriminating between items. However, subgroup analyses showed no difference between experts having more or less difficulty completing the survey, experts having more or less confidence in their results, and between the four versions of the questionnaires used. Moreover, some items may have been misunderstood because the formulation of the AMSTAR2 items may not be familiar for respondents and because, to facilitate the survey, the subdescription of the items was not presented. In addition, the responses of participants who were familiar with the tool could be possibly biased by knowledge of critical items, which could explain why 5 of the seven critical items have a high relative importance. However, as per the results of the subgroup analysis, participants with limited knowledge had overall similar ranking than those with AMSTAR2 knowledge. Finally, this study used the AMSTAR2 tool. Other guidelines for critical appraisal of SRs/MAs are available, and experts could also value other recommendations that were not included in this tool.

Although our study did not intend to develop a scoring system for an SR/MA based on the relative weight of AMSTAR2, our study provides relevant information to determine the most important criteria. Further work is needed to explore the possibility of proposing a weighted score in accordance with the importance of the items to critically appraise SRs/MAs. Moreover, future research should focus on the important items such as literature search process and an appropriate application of statistical analyses to improve the recommendations.

Table 4. Continued

Use of AMSTAR2 ($N = 94$)		Don't use of AMSTAR2 ($N = 90$)		No knowledge of AMSTAR2 ($N = 58$)		Difference between groups	Effect size
Rank	RIS (95% CI)	Rank	RIS (95% CI)	Rank	RIS (95% CI)	*P*-value[a]	Eta^2
8	5.34 (4.2–6.5)	8	6.76 (5.5–8.0)	9	6.00 (4.6–7.5)	0.188	0.006
6	8.24 (6.9–9.5)	7	7.61 (6.3–9.0)	5	7.40 (5.7–9.1)	0.736	0.006
13	2.39 (1.7–3.0)	12	3.31 (2.4–4.2)	12	5.15 (3.9–6.4)	<0.001[b]	0.069
4	11.16 (10.1–12.3)	2	12.24 (11.2–13.3)	2	12.11 (10.8–13.4)	0.445	0.002
12	2.48 (1.7–3.2)	14	2.43 (1.7–3.1)	15	1.85 (1.1–2.7)	0.654	0.005
14	1.83 (1.2–2.5)	15	1.6 (1.1–2.1)	16	1.82 (1.0–2.6)	0.918	0.008
16	1.74 (1.1–2.4)	13	2.79 (2.0–3.6)	13	4.36 (3.2–5.5)	<0.001[b]	0.087
9	4.79 (3.9–5.7)	9	4.93 (4.0–5.9)	10	5.77 (4.5–7.0)	0.280	0.002
3	11.53 (10.8–12.3)	3	10.04 (9.2–10.9)	4	7.99 (6.7–9.3)	<0.001[b]	0.079
15	1.80 (1.1–2.5)	16	1.54 (1.0–2.1)	14	2.44 (1.3–3.6)	0.729	0.006
1	11.96 (11.0–13.0)	1	12.40 (11.4–13.4)	1	12.19 (10.7–13.7)	0.595	0.004
2	11.77 (10.9–12.7)	4	9.97 (9.0–10.9)	3	8.15 (6.9–9.4)	<0.001[b]	0.086
5	10.44 (9.5–11.3)	5	8.42 (7.5–9.4)	6	6.71 (5.4–8.0)	<0.001[b]	0.086
7	6.76 (5.9–7.6)	6	7.79 (6.9–8.7)	7	6.69 (5.5–7.9)	0.166	0.007
11	3.61 (2.8–4.4)	11	3.59 (2.8–4.4)	11	5.24 (3.9–6.6)	0.146	0.008
10	4.16 (3.1–5.2)	10	4.54 (3.4–5.7)	8	6.13 (4.4–7.9)	0.750	0.006

[a] *P*-value for the Kruskal-Wallis test conducted to compare the RIS of items between the 3 groups of the knowledge of AMSTAR2.

[b] Items statistically significant with the Bonferroni correction for multiple testing ($P \leq 0.003$). Interpretation of Eta^2: 0.01–<0.06 is a small effect size, 0.06–<0.14 is a moderate effect size, and >0.14 is a large effect size.

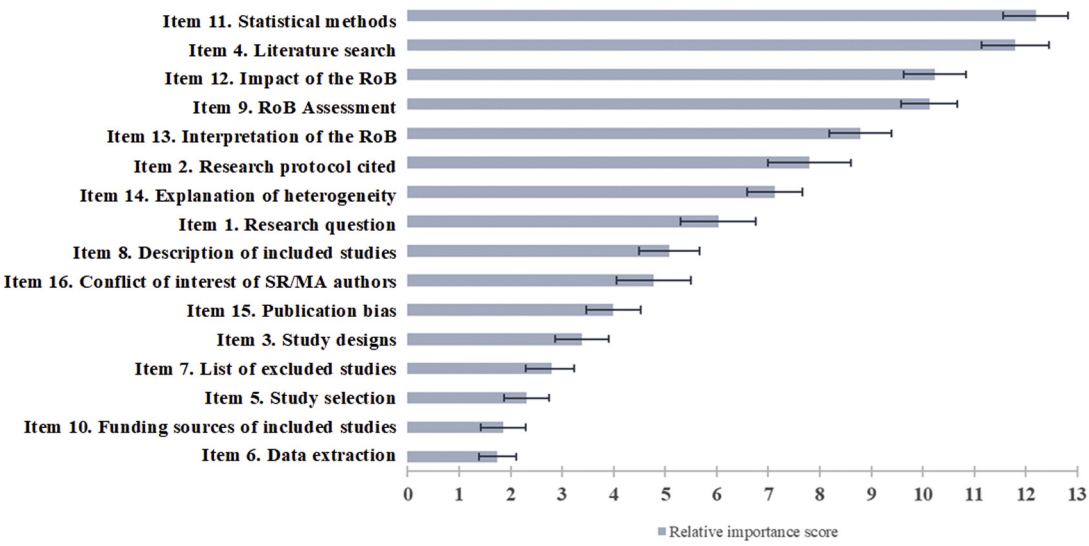

Fig. 1. Relative importance score of the 16 AMSTAR2 items as per the 242 experts.

5. Conclusion

In conclusion, this study elicited preferences from experts for AMSTAR2 items to critically assess SRs/MAs. The results of this study highlight the importance of keeping the two items (the adequacy of the literature search and the appropriateness of the statistical method) among the critical items proposed by AMSTAR2 because they have been highlighted as the two most important ones by a large panel of experts. Moreover, our results bring an additional piece of evidence for people that conduct overviews of reviews and need to apply a critical appraisal of SRs/MAs using AMSTAR2.

Credit authorship contribution statement

VL, MH, CB, GP, ET, and OB contributed to conceptualization. VL, MH, CB, and OB contributed to methodology. VL, MH, and GP contributed to investigation. VL and MH contributed to formal analysis & software. VL contributed to writing−original draft. VL, MH, CB, GP, ET, OB contributed to writing−review & editing. MH, ET, and OB contributed to supervision.

Acknowledgments

The authors would like to acknowledge the participation of all experts who took part in this research. Moreover, we would like to thank Ingrid Kremer and Luca Janssen for their help with the management of the data and statistical analyses.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.08.011.

References

[1] Page MJ, Moher D. Evaluations of the uptake and impact of the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement and extensions: a scoping review. Syst Rev 2017;6:263.
[2] Lane PW, Higgins JPT, Anagnostelis B, Anzures-Cabrera J, Baker NF, Cappelleri JC, et al. Methodological quality of meta-analyses: matched-pairs comparison over time and between industry-sponsored and academic-sponsored reports. Res Synth Methods 2013;4:342−50.
[3] Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nat Publ Gr 2018;555:175−82.
[4] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.
[5] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.
[6] Lorenz RC, Matthias K, Pieper D, Wegewitz U, Morche J, Nocon M, et al. A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool. J Clin Epidemiol 2019;114:133−40.
[7] Leclercq V, Beaudart C, Tirelli E, Bruyère O. Psychometric measurements of AMSTAR 2 in a sample of meta-analyses indexed in PsycINFO. J Clin Epidemiol 2019;119:144−5.
[8] Pieper D, Puljak L, González-Lorenzo M, Minozzi S. Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. J Clin Epidemiol 2019;108:26−33.
[9] Gates M, Gates A, Duarte G, Cary M, Becker M, Prediger B, et al. Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. J Clin Epidemiol 2020;125:9−15.
[10] Cheung KL, Wijnen BFM, Hollin IL, Janssen EM, Bridges JF, Evers SMAA, et al. Using best−worst scaling to investigate preferences in health care. Pharmacoeconomics 2016;34:1195−209.
[11] Louviere JJ, Flynn TN, Marley AAJ. Best-worst scaling: theory, methods and applications. Cambridge: Cambridge University Press; 2015.
[12] Cheung KL, Mayer S, Simon J, de Vries H, Evers SMAA, Kremer IEH, et al. Comparison of statistical analysis methods for object case best−worst scaling. J Med Econ 2019;22:509−15.
[13] Finn A, Louviere JJ. Determining the appropriate response to evidence of public concern: the case of food. J Public Policy Mark 1992;11:12−25.
[14] Mühlbacher AC, Kaczynski A, Zweifel P, Johnson FR. Experimental measurement of preferences in health and healthcare using best-worst scaling: an overview. Health Econ Rev 2016;6:1−14.
[15] Pollock M, Fernandes RM, Becker LA, Pieper DHL. Chapter V: overviews of reviews. cochrane handb. syst. rev. interv. version 6.0 (updated March 2020). London: Cochrane; 2020.
[16] Cheung KL, Evers SMAA, de Vries H, Hiligsmann M. Most important barriers and facilitators regarding the use of health technology assessment. Int J Technol Assess Health Care 2017;33:183−91.
[17] Kremer IEH, Evers SMAA, Jongen PJ, Van Der Weijden T, Van De Kolk I, Hiligsmann ML. Identification and prioritization of important attributes of disease-modifying drugs in decision making among patients with multiple sclerosis: a nominal group technique and best-worst scaling. PLoS One 2016;11:1−16.
[18] Orme B. Lighthouse Studio Help. Sawtooth; 2014. https://sawtoothsoftware.com/help/lighthouse-studio/manual/. Accessed March 1, 2020.
[19] Page MJ, Altman DG, McKenzie JE, Shamseer L, Ahmadzai N, Wolfe D, et al. Flaws in the application and interpretation of statistical analyses in systematic reviews of therapeutic interventions were common: a cross-sectional analysis. J Clin Epidemiol 2018;95:7−18.
[20] Rothstein HR, Bushman BJ. Publication bias in psychological science: comment on Ferguson and brannick (2012). Psychol Methods 2012;17:129−36.
[21] Faggion CM, Huivin R, Aranda L, Pandis N, Alarcon M. The search and selection for primary studies in systematic reviews published in dental journals indexed in MEDLINE was not fully reproducible. J Clin Epidemiol 2018;98:53−61.
[22] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6:e1000100.
[23] Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Journal pre-proof Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol 2020;121:20−8.
[24] Robson RC, Pham B, Hwee J, Thomas SM, Rios P, Page MJ, et al. Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review. J Clin Epidemiol 2019;106:121−35.