Published in *KT Update* (Vol. 5, No. 5 – July 2017) [http://ktdrr.org/products/update/v5n5]
An e-newsletter from the Center on Knowledge Translation for Disability and Rehabilitation Research

How Good Is That Systematic Review?

Marcel Dijkers, PhD, FACRM
Icahn School of Medicine at Mount Sinai
Department of Rehabilitation Medicine

KT Update *presents another in a series of brief articles by Dr. Marcel Dijkers. This article discusses a number of checklists for evaluating systematic reviews and meta-analyses and how such quality evaluation cannot be done properly with a reporting checklist such as Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).*

Systematic reviews (SRs) are generally accepted to provide the highest level of evidence for Evidence Based Practice (EBP), or Evidence-Informed Practice, as some now refer to it. Many a "pyramid" depicting the ranking of designs in terms of the quality and dependability of the evidence produced by studies puts SRs at the very top—for example, the Oxford Centre for Evidence-Based Medicine's 2011 Levels of Evidence scheme places SRs (to include meta-analyses, or MAs) at the highest level for all question types that require empirical evidence, including issues of incidence/prevalence, screening, diagnosis, prognosis, and treatment harms (Oxford Centre for Evidence-Based Medicine, 2011.)

Such reliance on SRs makes the topic of the quality of these secondary studies (the research they summarize constitutes the primary studies) a significant issue. Surprisingly, the number of tools that are available to distinguish the very good SRs from the somewhat good, and the latter from the terrible, is limited. A recent SR exploring tools available to assess all design types by Zeng et al. (2015) came up with only six:

- A Measurement Tool to Assess Systematic Reviews (AMSTAR) (Shea et al., 2007; Shea et al., 2009);
- Critical Appraisal Skills Programme (CASP) (CASP Systematic Review Checklist, 2017);

- National Institute for Health and Care Excellence Methodology Checklist (NICE) (National Institute for Health and Care Excellence, 2014);
- Sacks's Quality Assessment Checklist (Sacks, Berrier, Reitman, Ancona-Berk, & Chalmers, 1987);
- Overview Quality Assessment Questionnaire (OQAQ) (Oxman, 1994; Oxman & Guyatt, 1991); and
- Joanna Briggs Institute (JBI) Checklist for Systematic Reviews and Research Syntheses (Joanna Briggs Institute, 2016).

The authors missed checklists from the Scottish Intercollegiate Guidelines Network (SIGN) (Scottish Intercollegiate Guidelines Network, 2017), the National Heart, Lung, and Blood Institute (NHLBI) (National Heart, Lung, and Blood Institute, 2014), and Assessing the Quality and Applicability of Systematic Reviews (AQASR) (Task Force on Systematic Review and Guidelines, 2013). Risk Of Bias In Systematic Reviews (ROBIS) was published after Zeng et al. went to print (Whiting et al., 2016). There are probably a few more tools hiding in the literature and on the websites of organizations dedicated to EBP.

The AQASR, NICE, SIGN, JBI, and CASP tools are similar in that they focus on helping readers of an SR or MA decide whether the study will aid them in answering their clinical question, if at all, and if so, whether it was done well enough so that the recommendations and conclusions can be trusted. For example, the CASP tool starts with five questions (see Table 1) that could be answered for any SR. (In their Appendix S7, Zeng et al. [2015] mention these five only, disregarding the other five.) They all are to be answered with *Yes*, *No*, or *Can't tell*. There are "hints" to help the potential SR user identify what information is important to focus on in selecting an answer. These five questions are followed by two for which there are no answers to check off, but there are hints to guide the user in coming up with an answer. Finally, there are three queries on the potential benefit of the use of the intervention studied, or the diagnostic, screening, assessment, procedure, or tool in the local situation.

Table 1. Items in the CASP Tool, by Category

Question	Response
(A) Are the results of the review valid?	**Yes, Can't tell, or No**
1. Did the review address a clearly focused question?	
2. Did the authors look for the right type of papers?	
3. Do you think all the important, relevant studies were included?	
4. Did the review's authors do enough to assess the quality of the included studies?	
5. If the results of the review have been combined, was it reasonable to do so?	
(B) What are the results?	**(Open-ended)**
6. What are the overall results of the review?	
7. How precise are the results?	
(C) Will the results help locally?	**Yes, Can't tell, or No**
8. Can the results be applied to the local population?	
9. Were all important outcomes considered?	
10. Are the benefits worth the harms and costs?	

Note. CASP, 2017. *10 questions to help you make sense of a systematic review.*

Four other tools (SIGN, JBI, NICE, and ROBIS) do not have open-ended questions; they use checkboxes (respectively: *Yes* or *No*; *Yes*, *No*, *Unclear*, or *Not applicable*; or *Yes*, *Probably yes*, *Probably no*, *No*, or *No information*) without assigning a score value to the answers. In doing so the various authors clearly suggest that developing a quantitative total score is not an operationalization of the SR's quality, if such a thing is even possible. The ROBIS authors explicitly state: "We emphasize that ROBIS should not be used to generate a summary 'quality score' because of the well-known problems associated with such scores" (Whiting et al., 2016, p. 231).

The Sacks's, OQAQ, NHLBI, and AMSTAR checklists are similar in that the questions "Is this SR/MA applicable to my information need?" and "Are my patients/clients like the ones in the studies synthesized in this SR?" are not included. (In ROBIS, they are optional.) These checklists can be and are used to assess the quality of any single SR or MA, or any group of them, aside from immediate clinical application.

Sacks's checklist was used in one of the first studies reviewing published meta-analyses but apparently was forgotten, possibly because the list was not published in the article that presented the results of this "review of reviews," nor was it advertised as available from the authors (Sacks et al., 1987). It consisted of 23 items in six categories—prospective design, combinability, control of bias, statistical analysis, sensitivity analysis, and application of results—that apparently were scored, and the scores added a true measure of methodologic quality.

The OQAQ tool by Oxman and Guyatt is of a similar nature but has only 10 items (Oxman, 1994; Oxman & Guyatt, 1991). The same is true of AMSTAR, which without doubt is the most popular tool for evaluating these types of studies. It consists of 11 items that are scored *Yes*, *No*, *Can't answer*, and *Not applicable*. (Shea et al., 2007; Shea et al., 2009). AMSTAR has undergone extensive study of its metric qualities, which resulted in a judgment that the instrument is both reliable and valid (Pieper, Buechter, Li, Prediger, & Eikermann, 2015). There has been an apparently unsuccessful attempt to make scoring less subjective in R-AMSTAR (Kung et al., 2010). In recent years, there has been a trickle of papers arguing that improvements in AMSTAR are overdue (Burda, Holmer, & Norris, 2016; Faggion, 2015; Wegewitz, Weikert, Fishta, Jacobs, & Pieper, 2016). On their website (www.amstar.ca), the original authors have included a note: "We are improving on the original AMSTAR reflecting user comments and suggestions. The revised instrument will be even more user friendly, contain more guidance on its items, split to cater to different situations that may affect the direction of the review such [as] the included studies, grey literature, etc." (AMSTAR, 2015).

AQASR, which was developed by the Task Force on Systematic Review and Guidelines (2013) for the Center on Knowledge Translation for Disability and Rehabilitation Research (Center on KTDRR)—also the publisher of *KT Update*—is like the CASP tool but is enhanced. The reader receives a three-page introduction to how SRs are developed; a list of 112 questions to answer with respect to any SR or MA one is considering for advice on adoption of a measurement or diagnostic instrument or procedure; economic evaluation; prognosis; intervention, or practice for one's clinical,

research, or policymaking uses; and an 18-page glossary. In addition, there are 35 pages of guidance on answering the 112 questions, specifying what to look for and why the question and the details one is being told to look for are important (Task Force on Systematic Review and Guidelines, 2013). Answering 112 questions seems more than anyone would want to do before adopting (or rejecting) an SR, but Table 2 makes clear that in no instance is such a thing necessary.

Table 2. Categories of Questions in the AQASR Tool and the Number of Questions in Each Category

Category	Number of Questions
(A) Questions applicable to all SRs:	
SR question / clinical applicability	6
Protocol	5
Database searching	8
Other searches	2
Search limitations	6
Abstract and full paper scanning	8
Methodological quality assessment and use	6
Data extracting	4
Qualitative synthesis	6
Discussion	7
Various	3
(B) Questions only relevant to:	
SRs that incorporate a meta-analysis	7
SRs of studies of interventions/prevention	13
SRs of prognostic studies	6
SRs of studies of diagnostic accuracy	8
SRs of studies of measurement instruments	10
SRs of economic evaluations	7

Note. Task Force on Systematic Review and Guidelines. (2013). *Assessing the quality and applicability of systematic reviews (AQASR)*. Austin, TX: SEDL, Center on Knowledge Translation for Disability and Rehabilitation Research.

There are 61 "generic" questions, after which, depending on the issue the SR addresses, there are 6 to 13 more and an additional 7 if the SR incorporates an MA. That still is well more than the 10 questions of CASP, or any of the other instruments listed, but the advantage of AQASR is that it decomposes the questions present in the other checklists into many smaller, more concrete questions that may be easier to answer, especially for a neophyte in the dangerous SR waters.

One problem with all SR methodological quality tools is that, strictly speaking, one cannot answer their questions based on just a careful reading of the SR—one also should know quite a bit about the literature the SR authors looked at (or should have looked at) and whether they included the relevant primary studies in their review or rejected them. Take, for example, CASP Question 9: "Were all important outcomes considered?" One should know the literature to understand whether important outcomes were not considered. Or take NHLBI Question 3: "Did the literature search strategy use a comprehensive, systematic approach?" Someone wanting to assess an SR or MA needs to know rather well the literature relevant to the review's question—what terms are used in various disciplines, how the studies are indexed by the National Library of Medicine (PubMed) or other bibliographic database producers, etc. Not all questions require such broad knowledge, but in each assessment tool, at least a few do. The issue of a reader's need for knowledge of the literature does not exist at all, or on a much smaller scale, in evaluating primary studies. There, one needs to understand research design, and having a passing knowledge of the best outcome measures in a particular area may help, but one does not need semi-encyclopedic knowledge to be able to really evaluate a primary study. SRs and MAs may be at the top of the EBP pyramid, but determining whether any particular capstone is cut right is not easy.

And oh yes—Zeng et al. (2015) named one more tool for assessing the quality of SRs—Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). However, the readers of KTDRR's KT articles know that PRISMA is not a quality assessment tool but, rather, a guideline for complete and conscientious reporting of SRs and MAs (http://ktdrr.org/products/update/v5n3/index.html). The confusion is common, and there now are probably a few hundred articles in the literature that have

used PRISMA to assess the quality of SRs or MAs. However, it is a misuse of the tool and may lead to completely erroneous conclusions: A perfectly designed and implemented SR may be reported poorly, probably creating the impression that is was the product of shoddy science. On the other hand, by publishing a perfectly complete and honest written report, the authors of an MA may make clear how poorly their secondary study was done. Although there generally will be a correlation between the quality of the research and the quality of the reporting, in SRs and MAs, as in other research, there likely are exceptions—probably many. It may be easier to complete a PRISMA checklist for a SR of interest (all 27 items) than to complete, for example, an AQASR with more than 60 items, but one should not fool oneself into thinking the quality of the SR is known once there are 27 entries in PRISMA's boxes. Only careful completion of one of the SR quality assessment tools discussed can help make that determination.

References

AMSTAR. (2015). Developments. Retrieved from http://www.amstar.ca/Developments.php

Burda, B. U., Holmer, H. K., & Norris, S. L. (2016). Limitations of A Measurement Tool to Assess Systematic Reviews (AMSTAR) and suggestions for improvement. *Systematic Reviews, 5*, 58-016-0237-1. doi:10.1186/s13643-016-0237-1

Critical Appraisal Skills Programme (CASP). (2017). *10 questions to help you make sense of a systematic review* (CASP Systematic Review Checklist). Retrieved from http://docs.wixstatic.com/ugd/dded87_7e983a320087439e94533f4697aa109c.pdf

Faggion, C. M., Jr. (2015). Critical appraisal of AMSTAR: Challenges, limitations, and potential solutions from the perspective of an assessor. *BMC Medical Research Methodology, 15*, 63-015-0062-6. doi:10.1186/s12874-015-0062-6

Joanna Briggs Institute. (2016). Checklist for systematic reviews and research syntheses. Retrieved from https://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Systematic_Reviews.pdf

Kung, J., Chiappelli, F., Cajulis, O. O., Avezova, R., Kossan, G., Chew, L., & Maida, C. A. (2010). From systematic reviews to clinical recommendations for evidence-based

health care: Validation of Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) for grading of clinical Relevance. *The Open Dentistry Journal, 4*, 84–91. doi:10.2174/1874210601004020084

National Heart, Lung, and Blood Institute. (2014). *Quality assessment of systematic reviews and meta-analyses*. Retrieved from https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools/sr_ma

National Institute for Health and Care Excellence (NICE). (2014). *The guidelines manual: Appendices B–I. Appendix B: Methodology checklist: Systematic reviews and meta-analyses*. Retrieved from https://www.nice.org.uk/process/pmg10/chapter/appendix-b-methodology-checklist-systematic-reviews-and-meta-analyses

Oxford Centre for Evidence-Based Medicine. (2011). *2011 levels of evidence*. Retrieved from http://www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf

Oxman, A. D. (1994). Checklists for review articles. *BMJ (Clinical Research Ed.), 309*(6955), 648–651.

Oxman, A. D., & Guyatt, G. H. (1991). Validation of an index of the quality of review articles. *Journal of Clinical Epidemiology, 44*(11), 1271–1278.

Pieper, D., Buechter, R. B., Li, L., Prediger, B., & Eikermann, M. (2015). Systematic review found AMSTAR, but not R(evised)-AMSTAR, to have good measurement properties. *Journal of Clinical Epidemiology, 6*8(5), 574–583. doi:10.1016/j.jclinepi.2014.12.009

Sacks, H. S., Berrier, J., Reitman, D., Ancona-Berk, V. A., & Chalmers, T. C. (1987). Meta-analyses of randomized controlled trials. *The New England Journal of Medicine, 31*6(8), 450–455.

Scottish Intercollegiate Guidelines Network. (2017). *Critical appraisal notes and checklists*. Retrieved from http://www.sign.ac.uk/checklists-and-notes.html

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C.,…Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*, 10. doi:10.1186/1471-2288-7-10

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J.,…Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 6*2(10), 1013–1020. doi:10.1016/j.jclinepi.2008.10.009

Task Force on Systematic Review and Guidelines. (2013). *Assessing the quality and applicability of systematic reviews (AQASR).* Austin, TX: SEDL, Center on Knowledge Translation for Disability and Rehabilitation Research. Retrieved from http://ktdrr.org/aqasr/

Wegewitz, U., Weikert, B., Fishta, A., Jacobs, A., & Pieper, D. (2016). Resuming the discussion of AMSTAR: What can (should) be made better*? BMC Medical Research Methodology, 1*6(1), 111-016-0183-6. doi:10.1186/s12874-016-0183-6

Whiting, P., Savović, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B.,…ROBIS group. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology, 69*, 225–234. doi:10.1016/j.jclinepi.2015.06.005

Zeng, X., Zhang, Y., Kwong, J. S., Zhang, C., Li, S., Sun, F.,…Du, L. (2015). The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: A systematic review. *Journal of Evidence-Based Medicine, 8*(1), 2–10. doi:10.1111/jebm.1214. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/jebm.12141/ful

The contents of this article were developed under grant number 90DP0027 from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR). NIDILRR is a Center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). The contents of this article do not necessarily represent the policy of NIDILRR, ACL, HHS, and you should not assume endorsement by the Federal Government.