By Kylie Porritt, PhD, Judith Gomersall, PhD, and Craig Lockwood, PhD

# Study Selection and Critical Appraisal

## The steps following the literature search in a systematic review.

This article is the fourth in a series on the systematic review from the Joanna Briggs Institute, an international collaborative supporting evidence-based practice in nursing, medicine, and allied health fields. The purpose of the series is to describe how to conduct a systematic review—one step at a time. This article focuses on the study selection and critical appraisal steps in the process. These steps ensure that the review produces valid results capable of providing a useful basis for informing policy, clinical practice, and future research.

In this article we offer guidance for conducting the fourth and fifth stages in the systematic review process, which together can be referred to as *study selection*. As explained in the previous articles in this series from the Joanna Briggs Institute (JBI), the systematic review is a rigorous form of literature review in which reviewers take the following steps:

- formulate a review objective and question
- define inclusion and exclusion criteria
- perform a comprehensive search of the literature
- select studies for critical appraisal
- appraise the quality of the selected studies using one or more standardized tools
- extract data according to a template
- analyze, synthesize, and summarize data
- write up findings and draw conclusions (and in some cases make recommendations for practice, policy, or research)

Study selection is a vital stage in the review process and should be conducted to ensure that results are credible and useful in informing health care policy, clinical practice, and future research. In this stage you'll include only papers that are relevant to the review question and ensure that any limitations of these studies are understood. There are two essential steps in the study selection process: screening, which involves reviewing the citations resulting from your search and selecting those deemed relevant for full-text retrieval, and critical appraisal of the selected studies.

While conducting a systematic review is a step-by-step process, it's also characterized by plurality.[1,2] No single methodology is advocated by all organizations that develop and conduct systematic reviews. Also, the instruments used in appraising quantitative evidence differ from those used to review qualitative evidence. A mixed-methods systematic review will differ from that designed to review one evidence type, most notably at the data synthesis stage.

In this article we'll employ the JBI approach to study selection and cover two types of evidence, quantitative (which measures the effectiveness of an intervention) and qualitative (which examines individual meaning and experience). We'll review the two stages of study selection: first, how reviewers choose from the studies identified by their search; and second, how reviewers critically appraise both quantitative and qualitative evidence chosen.

### PHASE 1: SELECTING STUDIES USING PREDEFINED CRITERIA

Study selection begins once you've completed database searches and hand searches. Using the inclusion and exclusion criteria, at least two reviewers will select the articles that merit critical appraisal from all the identified citations (usually stored in an electronic library such as EndNote). Ensuring the transparency and reproducibility of this part of the process is vital. That's why we recommend a two-reviewer or group process.

Many reviewers err on the side of caution in attempting to be comprehensive. For instance, in cases where it's unclear from title or abstract whether a paper is relevant, a copy of the full text of the study is sought for consideration. But this approach can be resource intensive: papers may need to be photocopied or requested from other libraries at considerable expense, and waiting for an article can hold up the review for several months.

The following questions may help in reviewing citations in the first phase of study selection:

- Is the article published in the time period covered in the protocol?
- Is the article published in a language specified in the inclusion criteria?
- Does the population studied meet the inclusion criteria (such as adults or children or both)?
- Does the study look at the phenomena stated in the review question?
- Has the study design been reported? Is it relevant to the review question?
- Is an outcome measured?

If the review protocol specifies a date range for included papers, you would exclude a paper published outside of that range (unless it is considered a primary or seminal source, in which case you could include a sentence in the review protocol stating that papers of this type may be included).

Once you have chosen the studies that should be critically appraised, you'll obtain and read the full-text articles, discarding any that on second consideration do not meet the inclusion criteria.

### PHASE 2: APPRAISING SELECTED STUDIES

The purpose of critical appraisal is twofold. First, you'll exclude studies that are of low quality (and whose results may therefore compromise the validity of the recommendations of the review). And second, you'll identify the strengths and limitations of the included studies. The latter is important: an interpretation of the studies' results must be sensitive to the characteristics of the studied populations, as well as to how weaknesses in the study designs have affected those results.

**The optimal design for studies of the effects of interventions involves true randomization.**

Typically, two reviewers use checklists to appraise both quantitative and qualitative evidence. If the reviewers disagree and cannot resolve their differences through discussion, they consult a third reviewer. A number of checklists are available for assessing the many aspects of a study's quality, including its design, its methods and analysis, and its clinical relevance.[3] The goals and methods of quantitative and qualitative research differ, and so too do the checklists used to appraise them. The recently released 2014 version of the JBI's reviewers' manual offers checklists for appraising both types of studies (go to http://bit.ly/1h2F8RZ).[4]

Whether and how critical appraisal is conducted and reported is a significant indicator of quality in systematic reviews. At this stage, you're assessing full-text papers. For quantitative evidence you're identifying the risk of bias in the published research in order to decrease the possibility of including biased or misleading results. For qualitative evidence you're emphasizing the rigor of the research and the level of transferability. In the following two sections we'll look more closely at these two types of appraisal.

### QUANTITATIVE STUDIES: APPRAISING EVIDENCE OF EFFECTIVENESS

A range of study designs presents evidence on the effectiveness of interventions (therapies, technologies, or devices, for example). These include experimental, quasi-experimental, observational, and case reports. The study design used depends on the review question investigated and has its own advantages and limitations.

The ranking of evidence of effectiveness is generally linked to study design and the ability to maximize internal validity. For example, the randomized controlled trial (RCT) is ranked higher than a cohort study or case–control study; a systematic review of RCTs is ranked higher than a single RCT. Evidence hierarchies have been developed to be used as a tool to assist reviewers in the ranking of evidence. One such tool is the JBI Levels of Evidence; a new version was released in March (go to http://bit.ly/1qiic3Y).

There has been a surge of international interest in using GRADE (Grading of Recommendations Assessment, Development and Evaluation) when appraising RCTs for systematic reviews.[5] As explained by Goldet and Howick, GRADE differs from other appraisal tools by separating the quality of the evidence from the strength of the recommendation, assessing the quality of the evidence for each outcome, and upgrading observational studies that meet certain criteria.[5] (For more information, go to www.gradeworkinggroup.org.)

Two notions of validity guide reviewers seeking to examine the effectiveness of an intervention: internal validity and external validity. Internal validity refers to how good the study is—that is, how well a causal relationship between intervention and outcome can be inferred from the findings. For example, an internally valid RCT implies that the differences observed between groups receiving different interventions (apart from random error) are due to the intervention under investigation.[3] External validity, on the other hand, refers to the extent to which the results of the study can be generalized to groups, populations, and contexts that did not participate in the study. While it may

appear that there's a link between internal validity and generalizability, this is not the case. What a good study allows for is greater confidence in the findings when considering whether they are applicable to other populations. A good study does not automatically imply generalizability, but a poor study at significant risk for bias is not as useful in informing policy or practice because of its flaws.

many years. As an appraiser, you will need to assess how well attrition has been reported in the studies.

**Establishing external validity** involves reading in detail about the characteristics of the study population and how they were identified. It also encompasses identifying information about the study setting and whether it's sufficiently similar to the context to which the findings will be applied. Note that external validity is not about the accuracy or reliability of the results of a study. Rather, it's about the generalizability of the findings and the appropriateness of basing a change in practice on those findings.

You'll judge external validity by asking about the study's sampling method and sample characteristics, its context (cultural or organizational factors), and the intervention. How do these factors differ from those in the setting to which the findings will be applied? The optimal design for studies of the effects of interventions involves true randomization. True randomization affects external validity by increasing

**Internal validity refers to how good the study is—that is, how well a causal relationship between intervention and outcome can be inferred from the findings.**

**Establishing internal validity: assessing risk of bias.** The assessment of internal validity of quantitative studies involves determining whether the methods used in the study can be trusted to provide a genuine, accurate account of the intervention.[6-9]

The four following sources of bias may affect internal validity and can be addressed by questions asked in the study-selection process.

- *Selection bias* refers to the researchers' allocation of participants to groups that favor one of the treatments. This can be avoided by randomization and concealment of participant allocation, a form of blinding. Randomization ensures that every participant has an equal chance of being selected for any group. When appraising a study, your goal is to determine how well randomization has been achieved in order to ascertain whether bias has influenced study results. Randomization may not be possible in all study designs; for example, case–control design is inherently prone to selection bias (also known as *allocation bias*).
- *Performance bias* refers to the differences between groups in the care received. It can be avoided by blinding—the concealment of the treatment group from both participant and investigator.
- *Detection bias* arises when outcomes are assessed differently for treatment and control groups. Blinding is a recognized means of alleviating this type of bias; if researchers are unaware of which group a participant is assigned to, they will be more likely to deal with that participant impartially. Detection bias may also be referred to as *measurement bias*.
- *Attrition bias* refers to the differences in losses of subjects between groups. Losses to follow-up should be reported, though this is often difficult to do in longitudinal studies, which may last

**Table 1.** Critical Appraisal of Quantitative Evidence: A Checklist from JBI[4]

- Is the assignment to treatment groups truly random?
- Are participants blind to treatment allocation?
- Is allocation to treatment groups concealed from the allocator?
- Are the outcomes of people who withdrew described and included in the analysis?
- Are those assessing the outcomes blind to the treatment allocation?
- Are the control and treatment groups comparable at entry?
- Are groups treated identically other than for the named interventions?
- Are outcomes measured in the same way for all groups?
- Are outcomes measured in a reliable way?
- Is appropriate statistical analysis used?

the likelihood that participants in each group in the sample reflect the population they were recruited from, hence affecting the ability to generalize beyond the sample studied.

**Assessing characteristics of sample, culture, geography.** Fortunately, most journals require study authors to include a table that shows age, sex, and other relevant clinical or demographic information on participants. For example, a study of hemodialysis patients should include information on creatinine clearance levels, comorbidities, types of fistula, and other details that you can compare with your own patients to see how similar or dissimilar they are to the study sample. Also, identifying characteristics of the setting will inform you as to whether the study has relevance to your own practice setting. Interesting results from a small community pharmacy will have less relevance if you work in a large tertiary care center with automated dispensing.

**External validity is not about the accuracy or reliability of the results of a study. Rather, it's about the generalizability of the findings and the appropriateness of basing a change in practice on those findings.**

Cultural and geographic differences can have a major impact on external validity. Unique cultural practices or beliefs can exist between different groups, even between professions within the same hospital or within a profession across countries. Knowing that cultural differences exist and are not limited to racial characteristics is an important step in establishing external validity. Geographic differences across countries can be overt such as in prevalence studies of tropical or contagious diseases, or they can be more subtle such as socioeconomic differences between states or boroughs. If you find a study from a country where there are different funding models for health care provision, consider the relevance of public versus private funding models and their potential impact on outcomes in your own context.

Table 1 shows JBI's checklist for appraising internal and external validity of RCTs and quasi-RCTs.[4]

## QUALITATIVE STUDIES: APPRAISING EVIDENCE OF EXPERIENCES

Methods for establishing credibility in systematic reviews using qualitative studies have been developed and debated.[10-13] In a quantitative review, studies are appraised to identify sources of bias (selection, performance, and attrition). But what constitutes quality in a qualitative study? Should it even be assessed? And if so, how? These are highly contentious questions, and there's little consensus in the debate, raising as it does issues of ontology, epistemology, and methodology.[14]

Qualitative research is characterized by a wide-ranging methodological tradition, explained in part by ontological, epistemological, and philosophical perspectives. Ontological assumptions—asking whether something exists and how we can know it exists—influence why and how a qualitative researcher seeks knowledge about human consciousness. Approaches to qualitative research are informed by varying ontological positions and are the source of much debate, questioning, and contention among qualitative researchers, since they deal with fundamentals of the meaning and the nature of knowledge. Constructivism, for example, is informed by an ontology that says how we know something is shaped through our interaction with it, while an interpretivist perspective is based on the notion that meaning is subjective, with an emphasis on individual meaning. Such differences—socially developed versus individual perspectives on how meaning is made—reflect the diversity among qualitative researchers and in how they investigate knowledge.

Those who believe that qualitative research should be assessed for quality take that position because qualitative research can be flawed.[13] Averis and Pearson state that the critical appraisal of qualitative research contributes to its ongoing credibility, transferability, and theoretical potential.[11,15] Some researchers have attempted to develop criteria for appraising qualitative studies. In a review examining the layperson experience of diabetes and diabetes care, a modified version of the Critical Appraisal Skills Programme (CASP) was used to assist with critically appraising each paper.[16] The authors found the level of agreement between the assessors, when using the CASP, was reasonable.

Hannes and colleagues compared three qualitative-appraisal instruments: the CASP checklist, the Evaluation Tool for Qualitative Studies (ETQS), and JBI's Qualitative Assessment and Review Instrument (JBI-QARI) for Interpretive and Critical Research.[10] The study found that CASP was less sensitive to validity than either the JBI-QARI or the ETQS, and while the ETQS had a clear instruction set, the JBI-QARI, with its congruity among philosophical perspective, methodology, and the methods used to conduct the research, was the most coherent of the three instruments.

Some researchers resist the notion of critical appraisal for qualitative research, saying that relevant findings or a "golden nugget" of information may be missed if papers are excluded because of their quality.[17-20] Others argue that because qualitative research represents a unique form of science, its appraisal requires unique criteria. Walsh and Downe write that the "epistemological status of most qualitative research makes the indiscriminate transferral" of criteria evaluating validity and reliability "inappropriate."[13]

Traditionally, the terms used to measure research quality in quantitative research are *reliability* and *validity*. Reliability is the extent to which the results of a study are repeatable in different circumstances; validity is the degree to which a study reflects or assesses the concept the researcher is attempting to measure. Analogous terms relevant to qualitative research have been developed, and these are generally well accepted by qualitative researchers.

*Dependability* in qualitative research closely corresponds to the notion of reliability in quantitative research.[21] To maintain dependability, the qualitative research process should be logical, traceable, and clearly documented.

**To maintain dependability, the qualitative research process should be logical, traceable, and clearly documented.**

*Credibility* in qualitative research addresses whether a finding has been represented correctly; it corresponds to internal validity in quantitative studies. Credibility depends on the researcher's ability to address the "fit" between respondents' views and the researcher's representation of them; strategies used to ensure credibility include member checks (returning to participants after data analysis), peer checking (using outsiders to reanalyze data), prolonged engagement, persistent observation, and audit trails.

*Transferability* in qualitative research refers to the generalizability of results, an area of contention among researchers. It corresponds to external validity in quantitative research and might be thought of as a matter of "fit" between the situation studied and others to which one might be interested in applying the concepts and conclusions of that study; this is sometimes referred to as *cross-case generalizations*.[22]

At JBI, we consider the critical appraisal of identified studies as a required stage in the process of conducting a qualitative synthesis using meta-aggregation, though it is not necessary in some approaches to qualitative synthesis—meta-ethnography, for example. Meta-aggregation is a structured

**Table 2.** Critical Appraisal of Qualitative Evidence: A Checklist from JBI[4]

- There is congruity between the stated philosophical perspective and the research methodology.
- There is congruity between the research methodology and the research question or objectives.
- There is congruity between the research methodology and the methods used to collect data.
- There is congruity between the research methodology and the representation and analysis of data.
- There is congruence between the research methodology and the interpretation of results.
- There is a declaration of the researcher's cultural or theoretical orientation.
- The influence of the researcher on the research, and vice versa, is addressed.
- There is representation of participants and their voices.
- There is ethical approval by an appropriate body.
- There is a relationship between the conclusions of the study and the analysis or interpretation of the data.

approach in which findings of high-quality studies are integrated; in meta-ethnography, the findings of qualitative studies are reinterpreted so that new knowledge or theory can be generated. From the JBI perspective, critical thinking should be applied to studies before they are included in a review and should focus on congruity between the following:

- epistemology and theoretical perspective—that is, there is agreement between philosophy and the set of assumptions
- theoretical perspective and methodology—that is, there is agreement between the set of assumptions aligned with the particular perspective within the research and the theoretical underpinning of the research
- methodology and methods—that is, there is agreement between the theoretical underpinning of the research and the methods used within the research

Table 2 presents points to consider when appraising qualitative research from the JBI-QARI.[4]

## NEXT STEPS

Once you've critically appraised the studies found in the literature search, your next steps are data extraction and analysis. How will you decide which studies are of sufficient quality for the data extraction stage of the review? This is a question your entire review team will take up, and no single approach is considered "best practice." Wherever you draw the line for quality, though, you'll have to apply a uniform standard for all studies considered. Only then will the conclusions and recommendations you draw in the review be valid and useful. That will be the topic for the next article in this series. ▼

**Keywords:** quality, quality assessment, systematic review

*Kylie Porritt is a research fellow at the Joanna Briggs Institute in the School of Translational Health Science, University of Adelaide, South Australia, where Judith Gomersall is a research fellow at the National Health and Medical Research Council Centre for Research Excellence in Aboriginal Chronic Disease Knowledge Translation and Exchange, and Craig Lockwood is an associate professor in Implementation Science. Contact author: Kylie Porritt, kylie.porritt@adelaide.edu.au.* *The authors have disclosed no potential conflicts of interest, financial or otherwise.*

*The Joanna Briggs Institute aims to inform health care decision making globally through the use of research evidence. It has developed innovative methods for appraising and synthesizing evidence; facilitating the transfer of evidence to health systems, health care professionals, and consumers; and creating tools to evaluate the impact of research on outcomes. For more on the institute's approach to weighing the evidence for practice, go to http://joannabriggs.org/jbi-approach.html.*

## REFERENCES

1. Lockwood C, et al. *Synthesizing quantitative evidence*. Adelaide, SA: Lippincott Williams and Wilkins/Joanna Briggs Institute; 2011. Synthesis science in healthcare series.
2. Pearson A, et al. *Synthesizing qualitative evidence*. Adelaide, SA: Lippincott Williams and Wilkins/Joanna Briggs Institute; 2011. Synthesis science in healthcare series.
3. Juni P, et al. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323(7303):42-6.
4. Joanna Briggs Institute. *Joanna Briggs Institute reviewers' manual: 2014 edition*. Adelaide, SA; 2014. http://joannabriggs.org/assets/docs/sumari/ReviewersManual-2014.pdf.
5. Goldet G, Howick J. Understanding GRADE: an introduction. *J Evid Based Med* 2013;6(1):50-4.
6. Joanna Briggs Institute. An introduction to systematic reviews. *Changing practice: evidence-based practice information sheets for health professionals*. 2001;5(Suppl 1):1-6.
7. Khan KS, et al. Five steps to conducting a systematic review. *J R Soc Med* 2003;96(3):118-21.
8. Pearson A, et al. The JBI model of evidence-based healthcare. *Int J Evid Based Healthc* 2005;3(8):207-15.
9. Tricco AC, et al. The art and science of knowledge synthesis. *J Clin Epidemiol* 2011;64(1):11-20.
10. Hannes K, et al. A comparative analysis of three online appraisal instruments' ability to assess validity in qualitative research. *Qual Health Res* 2010;20(12):1736-43.
11. Pearson A. Balancing the evidence: incorporating the synthesis of qualitative data into systematic reviews. *JBI Reports* 2004;2:45-64.
12. Sandelowski M. Rigor or rigor mortis: the problem of rigor in qualitative research revisited. *ANS Adv Nurs Sci* 1993;16(2):1-8.
13. Walsh D, Downe S. Appraising the quality of qualitative research. *Midwifery* 2006;22(2):108-19.
14. Campbell R, et al. Evaluating meta-ethnography: systematic analysis and synthesis of qualitative research. *Health Technol Assess* 2011;15(43):1-164.
15. Averis A, Pearson A. Filling the gaps: identifying nursing research priorities through the analysis of completed systematic reviews. *JBI Reports* 2003;1(3):49-126.
16. Campbell R, et al. Evaluating meta-ethnography: a synthesis of qualitative research on lay experiences of diabetes and diabetes care. *Soc Sci Med* 2003;56(4):671-84.
17. Sandelowski M. "To be of use": enhancing the utility of qualitative research. *Nurs Outlook* 1997;45(3):125-32.
18. Shaw RL, et al. Finding qualitative research: an evaluation of search strategies. *BMC Med Res Methodol* 2004;4:5.
19. Sherwood G. Meta-synthesis: merging qualitative studies to develop nursing knowledge. *International Journal for Human Caring* 1999;3(1):37-42.
20. Walsh D, Downe S. Meta-synthesis method for qualitative research: a literature review. *J Adv Nurs* 2005;50(2):204-11.
21. Lincoln YS, Guba EG. *Naturalistic inquiry*. Newbury Park, CA: Sage Publications; 1985.
22. Salmond SW. Steps in the systematic review process. In: Holly C, et al., eds. *Comprehensive systematic review for advanced nursing practice*. New York, NY: Springer Publishing; 2012.