Published as: Carroll, C., & Booth, A. (2015). Quality assessment of qualitative evidence for systematic review and synthesis: Is it meaningful, and if so, how should it be performed? *Research Synthesis Methods*, *6*(2), 149-154.

Method Note

# Quality assessment of qualitative evidence for systematic review and synthesis: Is it meaningful, and if so, how should it be performed?

Christopher Carrolla\*† and Andrew Boothb

a University of Sheffield, School of Health and Related Research, Regent Court, Regent Street, Sheffield, South Yorkshire S1 4DA, UK

b University of Sheffield, School of Health and Related Research (ScHARR), Sheffield, South Yorkshire, UK

\* Correspondence to: Christopher Carroll, University of Sheffield, School of Health and Related Research, Regent Court, Regent Street, Sheffield, South Yorkshire S1 4DA, UK.

† E-mail:c.carroll@shef.ac.uk

## Abstract

The critical appraisal and quality assessment of primary research are key stages in systematic review and evidence synthesis. These processes are driven by the need to determine how far the primary research evidence, singly and collectively, should inform findings and, potentially, practice recommendations. Quality assessment of primary qualitative research remains a contested area. This article reviews recent developments in the field charting a perceptible shift from whether such quality assessment should be conducted to how it might be performed. It discusses the criteria that are used in the assessment of quality and how the findings of the process are used in synthesis. It argues that recent research indicates that sensitivity analysis offers one potentially useful means for advancing this controversial issue.

**Keywords: qualitative evidence synthesis; critical appraisal; quality assessment; systematic review**

## 1. Background

The critical appraisal and quality assessment of primary research are key stages in systematic review and evidence synthesis. These processes are driven by the need to determine how far the primary research evidence, singly and collectively, should inform findings and, potentially, practice recommendations. Even in conventional quantitative systematic review, the tools and processes for critical appraisal are open to debate, particularly once reviewers reach beyond randomised controlled trials. The Cochrane Collaboration, for example, with its international pedigree in systematic review and its methodology, has only recently decided to develop a risk of bias tool for non-randomised trial evidence (Reeves et al., 2013). However, where quantitative systematic review of trials is concerned, there is a general consensus on which criteria are important and how and when they should be applied in any given review (Higgins et al., 2011). This consensus around the value of critical appraisal contrasts with the ongoing debate pertaining to its rightful place within qualitative evidence synthesis (QES). Nevertheless, recent developments indicate that even within QES, some aspects of the quality assessment process are becoming clearer. The aim of this article therefore was to conduct a narrative review of recent literature that researches or discusses the issue of quality assessment for QES and to establish where we are and where we go from here.

We therefore briefly rehearse previous views on this topic as to whether it should be performed and draw on perspectives from the recent literature to examine both the appraisal criteria themselves and how appraisals are used in evidence synthesis. Recent literature on quality assessment within QES was identified from an annual update performed by the information specialist for the Cochrane Collaboration Qualitative and Implementation Methods Group (AB). This update was used to inform the content of the Cochrane-Collaboration-affiliated workshop on Issues and Challenges for Qualitative Research in Evidence Synthesis (InCQuiRES) in September 2013. The procedure involved manual review of the previous 24 months of the Group’s Methodology Register, compiled from monthly updates run against MEDLINE, Web of Science and Scopus. Keywords included generic terms for QES (e.g. systematic review of qualitative research, qualitative meta-synthesis, qualitative research synthesis, etc.) and an exhaustive list of QES variants (e.g. meta-ethnography, critical interpretive synthesis, realist synthesis, framework synthesis, thematic synthesis, etc.; a full list of search variants is available on request from the second author). Topic-based searching was supplemented by citation searches of 20 key published works covering a wide range of methodologies (Appendix 1).

## 2. Critical appraisal: does it matter?

For many years, commentators have questioned the following: Is critical appraisal of qualitative research possible or even defensible? The issues have been well covered and include debates about the appropriateness of applying quantitative paradigms and approaches to qualitative research and review (Dixon-Woods et al., 2007a). A principal concern relates to the philosophical and epistemological diversity of qualitative research, which some researchers feel precludes meaningful appraisal of that research; approaches and purposes developed for quantitative review are believed to have little or no role to play in the review and synthesis of qualitative evidence. In contrast, other perspectives have acknowledged the need for critical appraisal, arguing that it is important to know whether the primary qualitative research is “good enough” to inform a synthesis and subsequent practice (Britten et al., 2011; Toye et al., 2013). In the early years, the problem of critical appraisal was invariably commented upon in published reviews and syntheses of qualitative evidence. Authors tended either to choose not to make an appraisal at all or to attempt some form of appraisal using arbitrarily selected criteria (Attree, 2006; Reiss et al., 2007; Ridd et al., 2009).

Despite such debate, recent years have witnessed an emerging trend for reviewers to perform critical appraisal rather than to elect not to do so. Hannes and Macaitis (2012) considered more than 80 qualitative evidence syntheses published between 1998 and 2008. They found that critical appraisal had been conducted in the majority of cases, with authors in only five reviews explicitly stating, and justifying, the absence of critical appraisal (Hannes and Macaitis, 2012). As a consequence, we conclude, with Hannes and Macaitis (2012), that “it becomes more important to shift the academic debate from whether or not to make an appraisal to what criteria to use”. The debate over what criteria to use therefore remains one of the issues currently exercising those conducting qualitative evidence syntheses.

## 3. Which criteria should be used?

Critical appraisal seeks to assess the validity and reliability of a primary research study and its findings. There appears to be little agreement on “standard” criteria, with the choice being largely left to reviewers. Within the set of articles considered by Hannes and Macaitis (2012), researchers used multiple critical appraisal checklists and criteria, often adapting existing checklists for their own review: 24 different tools were used in the 59 reviews that conducted critical appraisal. Recent practice remains consistent with this plurality of approach, as does current advice from the Cochrane Qualitative and Implementation Methods Group. For example, the Group’s Supplementary Guidance to the Cochrane Handbook states that critical appraisal of qualitative research should be undertaken using, “an instrument underpinned by a multi-dimensional concept of quality in research… According to several domains…including reporting, methodological rigour and conceptual depth and breadth…Reviewers need to decide which instrument appears to be most appropriate in the context of their review” (Hannes and Cochrane Collaboration Qualitative Methods Group, 2011).

More recently, Toye and colleagues (2013: p.11) also recommended selection of criteria depending on the question being asked and the evidence to be appraised: “[T]he criteria by which we judge quality are not fixed but shift and change over time in relation to context.” This view is echoed by Garside (2014), who states that the choice of criteria depends on the reviewers’ perspective. Other QES approaches seek to define quality in relation to the underlying purpose of each synthesis: Realist synthesis, for example, explores “quality” in terms of how well a study tests a particular theory of interest, that is, criteria of utility and relevance (Pawson et al., 2005). In contrast to the application of specific criteria, “expert opinion,” the intuitive assessment of the quality or value of a piece of research, has also been evaluated. Dixon-Woods et al. (2007a) and Toye et al. (2013) all compared different approaches to critical appraisal, including “expert opinion,” and reported inconsistencies in quality assessment both within and across different tools and techniques.

Such plurality of criteria and inconsistency of findings begs the question as follows: Are reviewers producing meaningful assessments of the quality of the articles included in their review? The experience of reviewers suggests not. Toye et al. (2013) noted that, “You end-up measuring what you can measure rather than what is necessarily important.” Franzel et al. (2013) included criteria on openness and reflexivity in their appraisal tool while noting that these criteria were rarely satisfied by their included studies, principally because these elements were not reported. It appears that a review team can choose their own criteria, for reasons they deem important, but then hit the barrier of inadequate reporting of necessary detail. Indeed, poor reporting is endemic; a review team can only begin to make judgements about technical processes, conduct and trustworthiness, once they have arrived at a clear picture of what was performed (Carroll et al., 2012). Deficient reporting has led some academics to propose standardisation of reporting of “what is important” in qualitative research (Tong et al., 2007; Garside, 2014; Salzmann-Eriksen, 2013). Similar initiatives from quantitative research, for example, the Consolidated Standards of Reporting Trials statement for randomised controlled trials (CONSORT, 2010), have had a substantial impact on the reporting of such studies, albeit offering no guarantee of concordance between conduct and reporting (Vale et al., 2013). Garside (2014) has also pointed out that whether or not a study (as published) satisfies any given criteria can also depend, first, on the requirements placed upon the study’s authors (by the journal, by their discipline and by their research) and, second, on the fit between the primary study’s question and that posed by the review.

The problem of which criteria a review team should use is therefore being addressed, but no “standard” is yet emerging, should one be desired. Recent research appears to conclude that it is appropriate for criteria to be selected for each review based on both context and question. However, such an approach will continue to founder on the barrier of reporting. Given the persistence of inadequate reporting, we contend that it makes sense for quality of reporting criteria to precede the application of subsequent critical appraisal criteria (Carroll et al., 2012). These additional criteria should then be selected, first, on the basis of a team’s relevant expertise. For example, teams lacking a strong background in primary qualitative research should avoid seeking to answer questions about theoretical congruence. Second, additional criteria might relate to those factors thought most likely to influence the findings of the included studies, such as the relevance of the sample and the sampling strategy.

## 4. How should critical appraisal findings be used in evidence synthesis?

Findings from the critical appraisal process must be applied to or used in the synthesis in some meaningful way (Dixon-Woods et al., 2007a, 2007b). As noted by Garside (2014), “it is through doing synthesis that the critical elements of quality are illuminated.” The most common single reason behind appraisal in recent reviews of qualitative evidence appears to be the exclusion from the synthesis of papers deemed “not good enough”: Exclusion of poor-quality papers was performed in 30 of the 59 reviews considered by Hannes and Macaitis (2012). However, such a decision raises two questions. First, where should one set the threshold for exclusion? Second, does the final synthesis actually benefit from the exclusion of studies deemed “weak”?

Regarding thresholds, Alison Tong and colleagues state that the “rationale for weighting or excluding…should be explicit” (2012). As with critical appraisal generally, a degree of agreement exists: The threshold being applied is often reported by reviewers, but otherwise, there is no standardisation. For example, Franzel et al. (2013) excluded any study that failed to satisfy all of their criteria; Carroll et al. (2012) excluded those studies that satisfied only one of four criteria; Toye et al. (2013) “excluded studies from our meta-ethnography on the basis that the methodological report was insufficient to make a judgment on interpretive rigour…(by consensus on intuitive criteria).” Such divergence is not new. Noyes and Popay (2007) proposed exclusion based on deficiencies in the “thickness” of a study’s evidence, while Thomas and Harden (2008) excluded any study that failed to satisfy all seven of their pre-specified criteria. The exclusion of studies therefore presents the most common approach and, as with criteria, thresholds are review dependent.

However, is it ever meaningful to exclude such studies? After all, as noted previously, any appraisal is potentially only evaluating the reporting of the study rather than its actual conduct: The exclusion or inclusion of studies might therefore benefit or, equally, adversely affect the final synthesis. The inclusion of all studies, or the exclusion of some, therefore presents a problem. The threshold for exclusion is necessarily subjective. In

recognition of such subjectivity, many reviewers resist exclusion on the basis of quality (Hannes and Macaitis, 2012; Gallacher et al., 2013). However, even where authors reject exclusion, using appraisal simply “to inform the discussion,” they admit that this is vague and equally subjective. Such a verdict renders the value of the quality assessment process itself open to question (Gallacher et al., 2013). Faced with these issues, the natural response is

to seek to quantify the impact on QES of weighing or excluding studies. Boeije et al. (2013) explored a method of weighting studies and their contribution to a synthesis. Using scores achieved by checklist and expert judgement, the authors assigned more weight to findings from higher scoring studies, even if other findings appeared more frequently. They concluded that, “The outcomes of this method could be compared to sensitivity analysis…the

question to be answered is, what would happen to the results if studies below a certain established ‘quality threshold’ would be systematically excluded?” (Boeije et al., 2013).

Sensitivity analysis has been reported in two recent studies (Carroll et al., 2012; Franzel et al., 2013). In each case, the authors conducted post hoc sensitivity analyses by, first, identifying studies that failed to achieve the required “threshold” based on critical appraisal criteria and, second, by removing these studies from the synthesis. The authors (Carroll et al., 2012) deemed studies in their review to be “inadequately reported” if they failed to satisfy at least two of four reporting criteria. Franzel et al. (2013) “excluded” studies based on a combination of reporting and more subjective elements. The post hoc “exclusion” of studies’ findings from the synthesis was made possible because both reviews performed a type of thematic synthesis, so the supporting studies and evidence for each theme or sub-theme were recorded. It was therefore possible to gauge the impact on the synthesis of excluding individual studies. The authors (2012) found that studies deemed “inadequate” neither

contributed any unique themes to the synthesis nor offered any original perspectives on any identified theme.

The superfluity of inadequate studies was echoed in the analysis by Franzel et al. (2013), who found that, “The themes from the excluded papers would not have altered the meta-synthesis. Theoretical study saturation was achieved because no new relevant data emerged regarding a category, either to extend or to contradict it.” Tong et al. (2012) also concluded, based on earlier examples of similar types of sensitivity analysis, that, “studies with

sparse detail about the conduct of research tend to contribute less to the synthesis” (Noyes and Popay, 2007; Thomas and Harden, 2008; Morton et al., 2010). Offering unequivocal direction is problematic given that both the reviews conducted sensitivity analysis using

different critical appraisal criteria. Carroll et al. (2012) applied quality of reporting criteria only, while Franzel et al. (2013) considered criteria relating to the plausibility and coherence of results and reflexivity, as well as reporting.

One strength of sensitivity analysis is that the critical appraisal criteria and the quality “threshold” must be specified but need not be standardised. Indeed, if a review team deemed a single methodological criterion to be the most important confounder of findings, then that alone could be used as the threshold for exclusion from the sensitivity analysis. A single criterion-based approach is analogous to clinical experts considering blinding of

patients to be the single biggest confounder of certain quantitative trial outcomes, and so a review team conducts sensitivity analysis based on that criterion alone. For a QES, such a criterion might be the adequacy and appropriateness of the sampling frame, or the data collection method. Sensitivity analysis need not be based on an “overall” critical appraisal score for each study, which can mask studies’ different relative strengths and weaknesses. Superficially, therefore, based on the studies described previously, the principle of excluding studies on the basis of study quality is supported. However, Garside (2014) offers an alternative interpretation: If findings are similar between studies of varying quality, there might be no reason to exclude those apparently weaker studies; indeed, the synthesis is not improved by doing so but might be harmed if these studies “emanated from different

disciplines or used different methods.” Such a conclusion is supported by our own sensitivity analysis (Carroll et al., 2012). In a QES of health professionals’ experiences of online learning, studies focusing exclusively on nurses tended to perform least well against the appraisal criteria. As a consequence, such studies might justifiably have been excluded; after all, they merely duplicated the findings of the better reported studies. However, the exclusion of these “inadequately reported” studies of nurses would have minimised the transferability of the synthesis findings, which would otherwise have been based almost exclusively on studies of doctors. Therefore, there is evidence that the exclusion of studies can adversely affect the generalisability of a review and synthesis.

Consequently, we propose that some form of sensitivity analysis based on a priori critical appraisal criteria, focused initially on quality of reporting, offers the most promising alternative to straightforward exclusion of some studies (or the inclusion of all). Such a process can be both transparent and systematic, that is, specified and justified criteria can be applied, and the effect on the synthesis of any studies of questionable merit can be

demonstrably evaluated. Unlike such instances where appraisals simply “inform the discussion,” sensitivity analysis offers less scope for selectivity when applying appraisal judgments because studies are demonstrated either to satisfy the pre-specified criteria or else they do not. Sensitivity analysis offers the potential to give form and meaning to the critical appraisal process, while also allowing reviewers to achieve a deeper and clearer

understanding of the included papers from a methodological perspective.

We contend that sensitivity analysis performed post-synthesis, based on specified critical appraisal criteria, as described previously and in detail elsewhere (Noyes and Popay, 2007; Thomas and Harden, 2008; Carroll et al., 2012; Franzel et al., 2013), should be used more extensively and/or new methods developed. Sensitivity analysis as a “default” position is a conservative and risk-averse proposal that recognises the fact that the current evidence

does not permit us to reach a firm conclusion on best practice. Rather, it suggests that sensitivity analysis by study quality offers an extremely useful means of applying critical appraisal findings to a QES. However, more research is needed to test the value of sensitivity analysis and to determine whether it might be applicable only to certain types of synthesis, evidence or questions. For example, the contribution of individual studies to a synthesis is more easily assessed in aggregative approaches and in thematic synthesis than in more interpretive syntheses such as meta-ethnography.

## 5. Conclusions

Within the specific context of health care, peer reviewers and editors appear to expect that some form of critical appraisal be undertaken when conducting QES. However, the variety and subjectivity of appraisal tools and criteria present a problem for those seeking to conduct such work. We propose therefore that criteria take into account both the limitations of reporting and how the quality assessment will be used within the synthesis.

Indeed, linking the choice of quality assessment approach to the synthesis offers the most worthwhile route for establishing the judicious selection of appropriate methods for investigation through future research.

Secondary researchers face three principal options when conducting QES, each with strengths and weaknesses. First, they might decide not to make an appraisal at all, arguing that quality assessment is too subjective, time consuming and likely to be of questionable value to a synthesis, but with a possible consequent loss of credibility with users of such syntheses. Second, reviewers could opt to exclude studies based on specific criteria. However, these criteria might be subjective and uncertain, and the external validity of the synthesis might be affected by arbitrary exclusions. Third, they might seek to conduct a quality assessment using some form of sensitivity analysis. Conducting such an assessment similarly suffers from potential issues relating to inappropriate use of criteria. However, it fulfils a requirement that critical appraisal be conducted; no study is lost to the synthesis; and the reviewers can evaluate the relative contribution of studies based on specified criteria. At least for the moment, this third option, namely to conduct a sensitivity analysis as a matter of routine, would seem to be the most risk-averse strategy.

## Acknowledgements

A version of this paper was presented by the first author at the Cochrane Collaboration workshop on InCQuiRES in September 2013.

## References

Attree P 2006. The social costs of child poverty: a systematic review of the qualitative evidence. Children and Society 20: 54–66.

Boeije HR, van Wesel F, Alisic E. 2013. Making a difference: towards a method for weighing the evidence in qualitative synthesis. Journal of Evaluation in Clinical Practice 17: 657–663.

Britten N, Campbell R, Pound P, Morgan M, Daker-White G, Pill R, Yardley L. 2011. Evaluating meta-ethnography: systematic analysis and synthesis of qualitative research. Health Technology Assessment 15(43): 1–164.

Carroll C, Booth A, Lloyd-Jones M. 2012. Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. Qualitative Health Research 22: 1425–1434.

CONSORT. 2010. CONSORT statement. Transparent reporting of trials. Available at: <http://www.consort-statement.org/> .

Dixon-Woods M, Booth A, Sutton A. 2007a. Synthesizing qualitative research: a review of published reports. Qualitative Research 7: 375–422.

Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B, Jones D. 2007b. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. Journal of Health, Service, Research and Policy 12: 42–47.

Franzel B, Schwiegershausen M, Heusser P, Berger B. 2013. How to locate and appraise qualitative research in complementary and alternative medicine. BMC Complementary and Alternative Medicine 13: 125.

Gallacher K, Jani B, Morrison D, Macdonald S, Blane D, Erwin P, May CR. et al. 2013. Qualitative systematic reviews of treatment burden in stroke, heart failure and diabetes-Methodological challenges and solutions. BMC Medical Research Methodology 13: 10.

Garside R. 2014. Should we appraise the quality of qualitative research reports for systematic reviews, and if so, how? Innovation: European Journal of Social Sciences 27: 67–79.

Hannes K. Cochrane Collaboration Qualitative Methods Group. 2011. Chapter 4: critical appraisal of qualitative research. In Noyes J, Booth A, Hannes K, Harden A, Harris J, Lewin S, Lockwood C (eds.), Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions Version 1. Available from URL <http://cqrmg.cochrane.org/supplemental-handbook-guidance>. updated August 2011.

Hannes K, Macaitis K. 2012. A move to more systematic and transparent approaches in qualitative evidence synthesis: update on a review of published papers. Qualitative Research 12: 402–442.

Higgins JPT, Altman D, Goetzche P, Juni P, Moher D, Oxman A, et al. 2011. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ 343: d5928.

Morton RL, Tong A, Howard K, Snelling P, Webster AC. 2010. The views of patients and carers in treatment decision making for chronic kidney disease: systematic review and thematic synthesis of qualitative studies. BMJ 340:c112.

Noyes J, Popay J. 2007. Directly observed therapy and tuberculosis: how can a systematic review of qualitative research contribute to improving services? A qualitative meta-synthesis. Journal of Advanced Nursing 57: 227–243.

Pawson R, Greenhalgh T, Harvey G, Walshe K. 2005. Realist review-a new method of systematic review designed for complex policy interventions. Journal of Health, Service, Research and Policy 10: 21–34.

Reeves BC, Higgins JPT, Ramsey C, Shea B, Tugwell P, et al. 2013. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. Special Issue: Inclusion of Non-Randomized Studies in Systematic Reviews 4: 1–11.

Reiss S, Hermoni D, Van Raalte R, Dahan R, Borkan JM. 2007. Aggregation of qualitative studies—From theory to practice: patient priorities and family medicine/general practice evaluations. Patient Education and Counseling 65: 214–222.

Ridd M, Shaw A, Lewis G, Salisbury C. 2009. The patient–doctor relationship: a synthesis of the qualitative literature on patients’ perspectives. British Journal of General Practice, 59: e116–e133.

Salzmann-Eriksen M 2013. IMPAD-22: A checklist for authors of qualitative nursing research manuscripts. Nurse Education Today 33: 1295–1300.

Thomas J, Harden A. 2008. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Medical Research Methodology 8: 45.

Tong A, Sainsbury P, Craig J. 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care 19: 349–357.

Tong A, Flemming K, McInnes E, Oliver S, Craig J. 2012. Enhancing transparency in reporting the synthesis of qualitative research (ENTREQ). BMC Medical Research Methodology 12: 181.

Toye F, Seers K, Allcock N, Briggs M, Carr E, Andrews J, Barker K. 2013. ’Trying to pin down jelly’ – exploring intuitive processes in quality assessment for meta-ethnography. BMC Medical Research Methodology 13: 46.

Vale CL, Tierney JF, Burdett S. 2013. Can trial quality be reliably assessed from published reports of cancer trials: evaluation of risk of bias assessments in systematic reviews. BMJ 346: f1798.

## Supporting information

Additional supporting information may be found in the online version of this article at the publisher’s web site.