Using evidence in practice

Who will appraise the appraisers?—The paper, the instrument and the user

Andrew Booth

Introduction

Scenario

A group of local librarians decides to set up a monthly journal club. Each month they critically appraise a research study from the professional literature. As secretary to the group you volunteer to compile critical appraisal checklists for use by the fledgling journal club.

Critical appraisal is a key building block both of Evidence-Based Practice (EBP) and of Evidence-Based Library and Information Practice (EBLIP) which seeks to adopt and adapt methods from the wider paradigm.[1] How confident can we be about the reliability and robustness of the process of critical appraisal? The misconception persists that critical appraisal yields some universal truth—in short, that the outcome is all (i.e. is this a good or bad paper?; how can I apply these results?). In fact many who encounter critical appraisal for the first time feel slightly cheated in that they expect to discover answers—42, life, the universe and everything.[2] When the outcome is reduced uncertainty—or even generation of further questions!—many feel locked within an interminable cycle.

Concerns regarding critical appraisal centre around three variables—the paper, the instrument and the user. Let us consider each in turn.

The paper

It is self-evident, but often overlooked, that critical appraisal is limited by the quality of the reporting of a particular study. In other words, we view the paper through the lens, or 'rose-tinted spectacles' of its authors. Those who have been involved in primary research, or even those who have grappled with secondary research, know that writing an article is a form of academic cosmetic surgery. The anarchic and generally chaotic tendencies of a research project are typically superimposed with some semblance of order through historic 'revisionism'. Authors only tell us what they choose to tell us—either voluntarily or in response to often superficial peer review.[3]

While medical literature makes progress in seeking to improve the quality of its reporting, evidenced in a proliferation of standards rejoicing in the acronyms CONSORT, MOOSE and QUOROM,[4] the library literature lags some way behind. Structured abstracts remain the exception rather than the rule in library research.[5] In fact, standards may be a two-edged sword—standardizing reports of research may improve conformity but would not necessarily benefit critical appraisal. Often, discrepancies in an article illuminate doubts as to its quality—replacing unintentional self-revelation by formulaic descriptions of methods might perversely make it more difficult to discriminate articles on the basis of quality.

This potential confusion between the quality of *reporting* a study and the quality of its *conduct* extends to other characteristics. Often the length of an article is prescribed by the hosting journal, not by considerations of methodological quality or completeness of reporting. Recent trends to place supplementary materials on a website may offer elucidation for the purposive reader, but perversely decrease our ability to reach a verdict based simply on the published article.

This brings us to a final issue regarding the paper—how is a judgement of the 'quality' of a study to be made? Checklists take a 'mechanical' approach to appraisal, relying on the presence or

absence of particular features in determining the extent of bias. In contrast, increasing experience with qualitative research leads us to recognize that, perversely, a poor study may yield a meaningful insight, whereas a well-conducted study may add little of originality. In responding affectively rather than cognitively, we may find that a paper resonates with our own experience or that of others. If it has this ring of truth do we still relegate it to the 'discard' pile? Our mechanical analogy might consider the Volkswagen Beetle. If our question is 'Does this car have a reliable engine under its bonnet?' we might answer 'No' when, in fact, the engine *is* reliable but is located in the boot! We certainly need to increase our understanding of which features are most critical to the design, conduct and reporting of a study before we can construct checklists to discriminate between good and bad examples.

Perhaps we should invert our assessment of studies by asking a more holistic 'Do I believe this study?' before proceeding through features that detract from this original belief. Such a 'Bayesian approach' would examine how we revise our belief that our decision is appropriate in the light of both the strength of the findings and the methodological rigour of the research which produced those findings.[6] Appropriately, how we judge the overall value of a study would be determined more by our perception of how critical each flaw is rather than by the actual effect of that particular flaw.

The instrument

Instruments for critical appraisal generally employ one of three approaches:

1 question led;
2 study design led;
3 generic.

Within these three overarching categories, further subdivisions can be made; whether checklists are multi-disciplinary or discipline specific, whether they cover only quantitative studies, only qualitative studies or both, whether they are posed as questions or statements, whether they are designed for writers of research or for readers of research and whether they score or weight their components or simply record presence or absence of features (Yes, No, Not sure).

1 Question-led approaches

Question-led approaches to critical appraisal were pioneered by the Evidence-Based Medicine Working Party's User Guides series, disseminated as a series of journal articles and a subsequent manual[7] and guidebook.[8] The group started from the most common types of question asked in medicine (e.g. diagnosis and therapy) and devised a checklist for each type of question. A similar approach has been used within library and information practice by various colleagues and me, as seen in the CriSTAL checklists[9] and the new RELIANT checklist.[10]

2 Study design-led approaches

The Evidence-Based Medicine Working Party subsequently added checklists aimed at specific types of study, mainly for secondary studies such as systematic review, practice guideline and economic evaluation. This approach was subsequently adopted by the Critical Appraisal Skills Programme which developed a checklist for randomized controlled trials based on the User Guide for Treatment questions.[11] Such a study design approach has been extended to questionnaires[12] and qualitative research.[13] Appraisal of the latter is not unproblematic, however, as certain checklists may privilege certain types of qualitative research at the expense of other equally valid approaches.[14]

3 Generic approaches

The attraction of the generic approach is that it relieves the reader of the responsibility to select an appropriate checklist beforehand. Corresponding disadvantages may either be that the tool may make unwarranted assumptions about the optimal characteristics of research (e.g. size is important, research should be generalizable) or that, in seeking to be more catholic, the tool becomes unwieldy and complex. Interestingly, longer checklists are not believed to perform significantly better than shorter alternatives. Generic approaches are typically seen within a particular discipline—for example, in developing a checklist for educational and training interventions we found many

Table 1 Some examples of different approaches to appraisal checklists (accessed 8 November 2006)

1 Question-led checklists	
Users' Guides to Evidence-Based Practice	http://www.cche.net/usersguides/main.asp
University of Glasgow Department of General Practice and Primary Care	http://www.gla.ac.uk/departments/generalpractice/c._check.htm
CRISTAL Use Studies	http://www.shef.ac.uk/scharr/eblib/use.doc
CRISTAL Information Needs Analysis	http://www.shef.ac.uk/scharr/eblib/needs.doc
Reader's Guide to the Literature on Interventions Addressing the Need for education and Training (RELIANT)	http://eprints.rclis.org/archive/00007163/01/RELIANTfinal_.pdf
2 Study design-led checklists	
Critical Appraisal Skills Programme	http://www.phru.nhs.uk/casp/critical_appraisal_tools.htm
BestBETS	http://www.manchesterheartcentre.org/bestbets/bestbets.php#check
Questionnaire	http://www.bmj.com/cgi/content/full/328/7451/1312/DC1#e
3 Generic checklists	
Graphic Appraisal Tool for Epidemiology (GATE)	http://www.health.auckland.ac.nz/population-health/epidemiology-biostats/epiq/#Critical%20Appraisal
Standard Quality Assessment Criteria	http://www.ihe.ca/documents/hta/HTA-FR13.pdf
Critical appraisal tool for library and information research	http://www.newcastle.edu.au/service/library/gosford/ebl/toolkit/appraise.html

checklists specific to the domain of education.[10] A recent article in *Library Hi-Tech* introduces a promising generic tool for addition to the practitioner's toolkit (Table 1).[15]

Developing a usable checklist

Just as methodological filters can be characterized according to the rigour of their development, so too we can do the same for checklists.[16] 'First generation' checklists, such as those in the *Pocket Guide to Critical Appraisal*,[17] appear to be produced by one or two individuals. Items selected for inclusion in such a checklist are usually well informed, but have not necessarily been tested. Typically, these may fall into the trap of focusing on reporting rather than on the characteristics of rigorous design. For example, ethics questions are currently in favour and contemporary checklists often ask 'Was the research conducted ethically?'. While it is important to increase awareness of what is ethically acceptable research, such a question is largely irrelevant to the quality of the research itself. If research is unethical, we have a duty not to repeat it, but we should not ignore it. Joseph Lister did not obtain ethical consent for injecting the young boy, James Phipps, with cowpox but this does not mean that we disregard findings from his research on vaccination!

Superior to opinion-based checklist development is where a checklist is tested on candidate studies. Content of these 'second generation' checklists again relies on the expertise or experience of the developer, but at least the utility of the checklist is demonstrated. Ideally, checklists should assist the reader by raising issues in the order that they are typically encountered within a research article. This is not always evident. The checklist from *Library Hi-Tech*, for example, considers issues relating to Data Collection before those of Study Design which appears counter-intuitive.[15] While the author does describe consultation around usability, it would be interesting to learn whether sequencing of questions was specifically addressed.

'Third generation checklists' ideally use rigorous scientific methods in both development and testing. For example, a Delphi process of consensus could be complemented by a literature-driven approach that reviews items from existing checklists using categorical analysis. This might be followed by random selection of candidate articles for appraisal, with half used to develop the checklist and the remainder used to test it. Such a method mirrors the process recommended for testing and retesting methodological filters.[16]

Whichever method we use, we need to remember that instrument development is a science not an art. Psychometric development of instruments

is a research field in its own right and it is naïve to believe that we could stumble on the perfect tool simply by accident. Research demonstrates that, not only do different appraisers using the same approach produce different verdicts,[18] but even the same appraiser using different methods may produce different results.[19] If (and this is a big if) our intent is to devise a scientifically robust tool, then we need to become familiar with issues relating to instrument validity.[20]

That, as they say, is a completely different subject. However, we can illustrate such complexity with the recently published checklist already mentioned above.[15] This particular checklist uses a scoring system which states that, if the number of 'Yes' responses is less than 75% of the total questions (or, conversely, if the number of 'No' responses and 'Unclear' responses combined is greater than 25%), then the study is of questionable validity. Aside from leading us to question 'Where does this magic 75% come from?' this scoring approach assumes that the answer to 'Is the instrument included in the publication?' should carry the same importance as 'Is there external validity?' To compound this problem, the same scoring threshold is given for each of four component sections. This not only assumes equal weight between each individual item, but also allows the relative weight of each section to be determined by the number of individual questions it addresses. In other words, as already observed when discussing questionnaires,[21] avoid scoring unless you have a valid basis for weighting. A further consideration relates to the question 'Why are we scoring anyway?' Is it to exclude the study?—as, by implication, for this checklist. Alternatively, is it to determine a relative value for different articles if they conflict or contradict? In either case, scoring is problematic, but the overall purpose of the appraisal must be factored into our instrument.

The user

A favourite ploy when running appraisal sessions is to ask people 'What is the strongest determinant of whether you will act on this study once you have appraised it?' After several suggestions, connected with either the properties of the article (e.g. its authors or its source; the pre-EBM ancien regime) or the checklist (e.g. its design, its quality; the EBM regime), we usually arrive at the answer—our prior beliefs or prejudices. Yes, we select, evaluate and remember information to support our individual preferences.[22] This means that we fail to look for evidence that disconfirms our pet hypotheses. It also means that we cannot spot errors in our own reasoning. We examine evidence contradicting our views more critically than that supporting them.[22] This observation confirms the value of appraisal as a group process for, as individuals, we may otherwise miss important information. However, a group may be led into errors through group dynamics (e.g. the personality of the facilitator or of one vocal or influential member). It may arrive at an uneasy consensus that all 'support' but no one individually 'believes'! Time constraints may even force the group to be active (we must make a decision—right?) rather than reflective (we must make a right decision).

Conclusion

In summary, critical appraisal is a flawed 'technology' with limitations surrounding the paper itself, the appraisal instrument and the appraisers, either collectively or individually. To the danger, reported by Sackett, of 'critical appraisal nihilism'[23]—the perception that no paper is ever good enough—we add two further dimensions—no instrument is good enough and no appraiser is good enough!

Before plunging into despair, we should examine the wording of our verdict—what is 'good enough'? A chink of light lies in focusing on the relative 'enough', not the absolute 'good'. We must use the best available evidence and the best available tool to reach the best available decision. In fact, we simply require what we are doing to be 'fit for purpose'.[22] As Petticrew emphasizes 'Critical appraisal is not carried out in pursuit of some "holy grail" of perfection'.[22] It suffices if the process is 'evidentially adequate'.[22] This enables us to operate in the here and now, while continuing our ongoing quest for improvement, namely better papers—better tools—even better appraisers!

In actuality, critical appraisal is an enabling mechanism, a focus for systematic exploration of the strengths and weaknesses of a research study.

That is why we must see a checklist as a navigational aid not an anchor.[24] This voyage of discovery *for* the reader and *about* the paper is its most important feature—evidence-based library and information practice is a journey not a destination!

References

1 Booth, A. & Brice, A. Appraising the evidence. In: Booth, A. & Brice, A. (eds). *Evidence-Based Practice for Information Professionals: A Handbook*. London: Facet Publishing, 2004: 104–18.

2 Booth, A. The Unteachable in Pursuit of the Unreadable? *Evidence Based Library and Information Practice* 2006, **1**, 51–6. Available from: http://ejournals.library.ualberta.ca/index.php/eblip/article/view/48/118 (accessed 8 November 2006).

3 Emden, C. Manuscript reviewing: too long a concealed form of scholarship? *Nursing Inquiry* 1996, **3**, 195–9.

4 Golder, S., Mather, L., Wright, K. & Glanville, J. What's on the Web: recommendations for reporting research. *Journal of Health Service Research Policy* 2005, **10**, 124–5.

5 Hartley, J. Current findings from research on structured abstracts. *Journal of the Medical Library Association* 2004, **92**, 368–71.

6 Schwartz, A. & Hupert, J. A decision-making approach to assessing critical appraisal skills. *Medical Teacher* 2005, **27**, 76–80.

7 Guyatt, G. & Rennie, D. (eds) *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, IL: AMA Press, 2002.

8 Guyatt, G. & Rennie, D. *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice.* Chicago, IL: AMA Press, 2002.

9 Booth, A. & Brice, A. Clear-cut?: facilitating health librarians to use information research in practice. *Health Information and Libraries Journal* 2003, **20**(Suppl. 1), 45–52.

10 Koufogiannakis, D., Booth, A. & Brettle, A. ReLIANT: Reader's guide to the Literature on Interventions Addressing the Need for education and Training. *Library and Information Research* 2006, **94**, 44–51.

11 Critical Appraisal Skills Programme. *Ten Questions to Help You Make Sense of Randomized Controlled Trials.* Available from: http://www.phru.nhs.uk/casp/critical_appraisal_tools.htm#rct (accessed 8 November 2006).

12 Boynton, P. M. & Greenhalgh, T. Selecting, designing, and developing your questionnaire. *British Medical Journal* 2004, **328**, 1312–5. Table E: Critical appraisal checklist for a questionnaire study. Available from: http://www.bmj.com/cgi/content/full/328/7451/1312/DC1#e (accessed 8 November 2006).

13 Greenhalgh, T. & Taylor, R. Papers that go beyond numbers (qualitative research). *British Medical Journal* 1997, **315**, 740–3.

14 Dixon-Woods, M., Bonas, S., Booth, A., Jones, D. R., Miller, T., Sutton, A. J., Shaw, R. L., Smith, J. A. & Young, B. How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research* 2006, **6**, 27–44.

15 Glynn, L. A critical appraisal tool for library and information research. *Library Hi-Tech* 2006, **24**, 387–99.

16 Jenkins, M. Evaluation of methodological search filters—a review. *Health Information and Libraries Journal* 2004, **21**, 148–63.

17 Crombie, I. K. *The Pocket Guide to Critical Appraisal*, London: BMJ Publishing Group, 1996.

18 MacRae, H. M., Regehr, G., Brenneman, F., McKenzie, M. & McLeod, R. S. Assessment of critical appraisal skills. *American Journal of Surgery* 2004, **187**, 120–3.

19 Dixon-Woods, M., Sutton, A. J., Shaw, R. L., Miller, T., Smith, J. A., Young, B., Bonas, S., Booth, A. & Jones, D. R. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. *J Health Serv Res Policy* 2007, **12**, 41–7.

20 Shaneyfelt, T., Baum, K. D., Bell, D., Feldstein, D., Houston, T. K., Kaatz, S., Whelan, C. & Green, M. Instruments for evaluating education in evidence-based practice: a systematic review. *Journal of the American Medical Association* 2006, **296**, 1116–27.

21 Booth, A. Mind your Ps and Qs (pitfalls of questionnaires). *Health Information and Libraries Journal* 2005, **22**, 228–31.

22 Petticrew, M. & Roberts, H. Chapter 5: How to appraise the studies: an introduction to assessing quality. In: Petticrew, M & Roberts, H. (eds). *Systematic Reviews in the Social Sciences: A Practical Guide.* Oxford: Blackwell Publishing, 2006: 125–63.

23 Sackett, D. L., Richardson, S. R., Rosenberg, W. & Haynes, R. B. *Evidence-Based Medicine: How to Practice and Teach EBM*. London: Churchill Livingstone, 1997, 188–99.

24 Eakin, J. M. & Mykhalovskiy, E. Reframing the evaluation of qualitative health research: reflections on a review of appraisal guidelines in the health sciences. *Journal of Evaluation in Clinical Practice* 2003, **9**, 187–94.

