International Journal of
Evidence-Based Healthcare
The Joanna Briggs Institute

doi:10.1111/j.1744-1609.2011.00237.x Int J Evid Based Healthc 2011; 9: 444–449

COMMENTARY

Comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: a randomised trial

Michael Crowe MIT BSc(Mgmt) ADMT,[1] Lorraine Sheppard PhD[1,2] and Alistair Campbell PhD[3]

[1]Discipline of Physiotherapy, James Cook University, Townsville, Queensland, [2]School of Health Sciences, University of South Australia, Adelaide, South Australia and [3]Discipline of Psychology, James Cook University, Townsville, Queensland, Australia

Abstract

In systematic reviews, evidence-based practice and journal clubs critical appraisal tools are used to rate research papers. However, little evidence exists on whether the critical appraisal tool, subject matter knowledge or research design knowledge affect the appraisal of research papers. A match paired randomised trial was conducted in August/September 2010 in the Faculty of Medicine, Health and Molecular Science, James Cook University, Australia. Ten participants in total were randomly assigned to two groups using either an informal appraisal of research (IA group) or the Crowe Critical Appraisal Tool (CCAT group), a general critical appraisal tool. Participant independently appraised five research papers, where each paper had a different research design. The scores allocated to the papers by each group were analysed. The intraclass correlation coefficient for absolute agreement was 0.76 for the informal appraisal group and 0.88 for the Crowe Critical Appraisal Tool group. The G study showed that in the informal appraisal group 24% of variance in scores was attributable to either the rater or paper × rater interactions, whereas this was 12% in the Crowe Critical Appraisal Tool group. Analysis of covariance showed that there were statistically significant results in the informal appraisal group for subject matter knowledge ($F(1,18) = 7.03$, $P < 0.05$ 1 tailed, partial $\eta^2 = 0.28$) and rater ($F(4,18) = 4.57$, $P < 0.05$ 1 tailed, partial $\eta^2 = 0.50$). Kendall's tau correlation coefficient also showed a significant weak positive relationship ($\tau = 0.38$, $P = 0.03$) between total score and subject matter knowledge for the informal appraisal group. The Crowe Critical Appraisal Tool was more reliable than an informal appraisal of the research papers. In the informal appraisal group, there were significant effects for rater and subject matter knowledge, whereas the Crowe Critical Appraisal Tool almost eliminated the rater effect, and no subject matter knowledge effect was apparent. There was no research design knowledge effect in either group. The Crowe Critical Appraisal Tool provided much better score reliability and should help readers with different levels and types of knowledge to reach similar conclusions about a research paper.

Key words: critical appraisal, evidence-based practice, qualitative research, quantitative research, systematic review.

Background

Critical appraisal tools (CATs) help readers to rate research papers and are used in systematic reviews, evidence-based practice and journal clubs.[1,2] There are many well-known CATs available such as the Jadad scale,[3] Maastricht scale,[4] Critical Appraisal Skills Programme tools,[5] Assessment of Multiple Systematic Reviews[6] and Single-Case Experimental Design scale.[7] However, these and other CATs suffer from similar problems. First, most CATs were designed to appraise either one or a small number of research designs.[1,8] When a reader wants to appraise many papers that use a diverse range of qualitative and quantitative research designs, or that use multiple or mixed methods, then they must use multiple CATs. The scores from multiple CATs cannot be compared because they may use different scoring systems, design features or assumptions that are incompatible. Second, the majority of CATs lack the depth to fully appraise research[8,9] or have scoring systems that are insufficient to accurately reflect the content of research papers.[10–12] In

Correspondence: Mr Michael Crowe, Discipline of Physiotherapy, James Cook University, 101 Angus Smith Drive, Townsville, Qld 4810, Australia. Email: michael.crowe@my.jcu.edu.au

either of these cases, the resultant score from the CAT can be compromised and, as a result, defects in the research may be hidden or not fully considered by a reader. Third, very few CATs have any validity and reliability data available.[1,13,14] This means that there may be no evidence that a particular CAT is effective or consistent in appraising research.

The Crowe Critical Appraisal Tool (CCAT)[1,15,16] was designed to overcome the problems outlined above. First, the CCAT was built based on a review of the design of 44 CATs across all research designs.[1] These CATs were analysed using a combination of the constant comparative method,[17,18] standards for the reporting of research[19–24] and research methods theory.[25–27] This analysis led to the development of a tool that consisted of eight categories (*preliminaries*, *introduction*, *design*, *sampling*, *data collection*, *ethical matters*, *results* and *discussion*) divided into 22 items, which were further divided into 98 item descriptors.[1] The combination of categories, items and item descriptors allows for a wide range of qualitative and quantitative health research to be appraised using one tool.[1,15,16] Second, a comprehensive user guide was produced that is considered vital to obtaining valid scores from the CCAT. Scoring is described in the user guide as a combination of subjective and objective assessment where each category is scored from 0 (the lowest score) to 5 (the highest score). Third, an evaluation of score validity[15] and reliability[16] were completed for the CCAT. These preliminary assessments showed that the scores obtained had a reasonable degree of validity, and the CCAT could be considered a reliable means of appraising health research in a wide range of research designs. The CCAT and user guide, as used in this study, are available as additional material online (Appendix S1).

However, while undertaking research into the CCAT,[15,16] two questions arose with regards to CATs in general. First, a search of the literature revealed only one article that tested whether using a CAT is an improvement over not using a CAT to appraise research.[28] Therefore, although it has been assumed that using a CAT is a better option, there is little evidence to substantiate this assertion. The second question was whether a reader's subject matter knowledge or research design knowledge influence the scores awarded to a research paper. In other words, when a reader looks for evidence as a basis for their practice, does their subject matter or research design knowledge affect how they rate research papers? If subject matter knowledge or research design knowledge does affect appraisal, then this may lead to situations where only evidence that reinforces current knowledge is incorporated into practice while evidence that is new to or contradicts with a reader's knowledge may be discarded, no matter how worthy.

Teaching and implementation of evidence-based practice may be improved by exploring the relationship between using a CAT versus not using a CAT and the influence of subject matter knowledge and research design knowledge on the appraisal of research papers. Therefore, the aims of this study were:

1. to investigate whether using a CAT versus not using a CAT (i.e. informal appraisal) affected how readers appraise a sample of health research papers and
2. to examine whether subject matter knowledge or research design knowledge affected how readers appraise a sample of health research papers.

Methods

The CAT used in the study was the CCAT. The CCAT was used because it was known to the authors; score validity and reliability data were available; and the CCAT could be used across all health research designs, removing a potential confounder where a different CAT could be required for each research design. The alternative to using a CAT was an informal appraisal of research papers where no CAT was supplied to participants. The outcome measure used was rating (total score as a percent) of health research papers using either the CCAT or informal appraisal.

Design

Potential participants were asked to take part in the study through a series of invitations emailed to academic/research staff and postgraduate research students in: the School of Public Health, Tropical Medicine and Rehabilitation Science; the School of Nursing, Midwifery and Nutrition; and the School of Medicine and Dentistry, James Cook University, Australia.

Participants were match paired by the principle investigator (MC) based on their level of research experience so that participants with similar experience were allocated to each research group. Research experience was determined by a pre-enrolment questionnaire that asked the participants to indicated: how many years they had been involved in research; on how many research projects they had worked; on how many projects they had been lead or principal researcher and a subjective assessment of their level of research experience on a scale from 1 (novice) to 5 (expert). This measure of researcher experience was not validated because it was only used to match participants rather than as a conclusive measure of researcher experience. No additional inclusion or exclusion criteria were used.

When all participants had been match paired, they were randomly assigned by the principal investigator to either the informal appraisal (IA) group (control) or the CCAT group (intervention), using the random sequence generator available from http://www.random.org.[29] The principal investigator was not blinded to the groups participants were allocated. Blinding was not considered necessary because participants individually scored papers without input from the principal investigator. Participants were informed that they could contact the principal investigator if they had any general questions regarding the study. However, questions concerning how to score a research paper, whether using the CCAT or not, would not be answered because this could affect the scores awarded and bias the results obtained. Furthermore, participants were requested not to discuss the

study with other participants, if they became aware of them, until data collection was completed.

Sampling

A sample size calculation showed that six raters reading five papers each were required to achieve an intraclass correlation coefficient (ICC) of 0.90 ($\alpha = 95\%$, $1-\beta = 0.79$, $r_{min} = 0.55$).[30] Two separate groups were required, which meant a minimum of 12 participants in total.

Health research papers to be scored were randomly selected using the random sequence generator available from http://www.random.org.[29] The research papers were selected from a larger pool of papers that was used in two other studies.[15,16] In brief, the larger pool of research papers was randomly selected from OvidSP's (Ovid, New York, NY, USA) full text articles subscribed to by James Cook University, Australia. Research papers in the larger pool were chosen based on the research design used in each paper, with possible categories of research designs being: true experimental, quasi-experimental, single system, descriptive, exploratory or observational, qualitative and systematic review. The five randomly selected papers were:

1. true experimental: Arts MP, Brand R, van den Akker EM, Koes BW, Bartels RH, Peul WC. Tubular diskectomy vs. conventional microdiskectomy for sciatica: a randomised controlled trial. *JAMA* 2009; **302**: 149–58;
2. quasi-experimental: Polanczyk G, Zeni C, Genro JP *et al.* Association of the adrenergic α2A receptor gene with methylphenidate improvement of inattentive symptoms in children and adolescents with attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* 2007; **64**: 218–24;
3. single system: Jais P, Haissaguerre M, Shah DC *et al.* A focal source of atrial fibrillation treated by discrete radiofrequency ablation. *Circulation* 1997; **95**: 572–76;
4. qualitative: Beck C. Postpartum depressed mothers' experiences interacting with their children. *Nurs Res* 1996; **45**: 98–104 and
5. systematic review: Singh S, Kumar A. Wernicke encephalopathy after obesity surgery: a systematic review. *Neurology* 2007; **68**: 807–11.

Data collection

All data were collected in August and September 2010. Each participant was supplied with a copy of the research papers to be appraised, instructions on what was required and forms to write their scores (see Appendix S1 online for copies of the forms). For the IA group, the participants were asked to read each research paper thoroughly and to rate each paper on a scale from 0 (the lowest score) to 10 (the highest score). No further instructions were given to the participants on how to determine the score for a paper other than to use their best judgement. For the CCAT group, the participants were asked to read each paper thoroughly and to fill out a CCAT form for each paper. The CCAT form was supplied with an extensive user guide to help participants use the tool as effectively as possible.

Participants in both groups were also asked to indicate their subject matter knowledge and their research design knowledge for each research paper. The scale used for both subject matter knowledge and research design knowledge was from 0 (no knowledge) to 5 (extensive knowledge).

Data analysis

When the appraisal forms were returned, the total scores for the CCAT group were checked by adding the individual category scores. Total scores for the research papers in the IA and CCAT groups were then converted to percentage scores so that the rating of papers could be compared. The reliability of scores was calculated using the ICC and generalisability theory (G theory). An analysis of covariance (ANCOVA) between the dependent variable (total score) and covariates (subject matter knowledge and research design knowledge) was also completed.

Ethics

Ethical approval for this study was obtained from James Cook University Human Ethics Committee (H3415) and the study conformed to the Declaration of Helsinki.[31] Written informed consent was obtained from each participant before they took part in the study. Participants could withdraw at any stage without explanation or prejudice. The authors have no potential conflicts of interest or sources of funding to declare.

Results

A total of 19 people responded to the invitation to participate in the study, and 10 participants (53%) completed the study. Despite repeated emails to attract further participants to the study, no other participants were forthcoming. Eight participants were academic/research staff and two were postgraduate students. Eight participants (not all of them staff) were from the School of Public Health, Tropical Medicine and Rehabilitation Science; one participant was from the School of Nursing, Midwifery and Nutrition; and one participant was from the School of Medicine and Dentistry. The flow of participants through the study is indicated in Figure 1.

Participants were match paired based on their responses to the pre-enrolment questionnaire. Four participants (two pairs) had a low level of research experience, four participants (two pairs) had a medium level of research experience and two participants (one pair) had a high level of research experience. There was no difference between the IA group and the CCAT group based on research experience. Subject matter knowledge for the IA group and the CCAT group were positively skewed (i.e. more participants stated they had low levels of knowledge rather than high levels of knowledge). For research design knowledge, both groups had normal distributions of data. There was no statistical difference (Mann–Whitney *U*-test) between the IA group and CCAT group for subject matter knowledge ($U = 298.5$, $z = -0.29$, $P = 0.78$ two-tailed) or research design knowledge ($U = 270.0$, $z = -0.85$, $P = 0.40$ two-tailed).

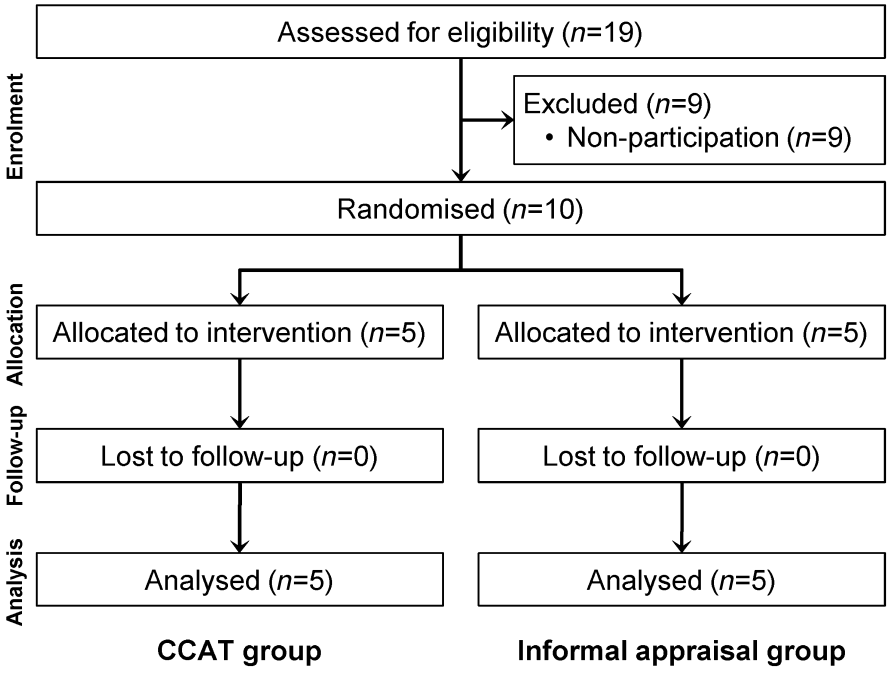

Figure 1 Flow of participants. CCAT, Crowe Critical Appraisal Tool.

The maximum score in the IA group was 90%, the minimum score was 30% (range 60%) and the average score was 67% with a standard deviation of 16%. The maximum score in the CCAT group was 98%, the minimum was 25% (range 73%) and the average score was 67% with a standard deviation of 22%. The total score for both groups was found to be normally distributed. In the IA group, Kendall's tau correlation coefficient showed a significant weak positive relationship ($\tau = 0.38$, $P = 0.03$) between total score and subject matter knowledge. There was no significant relationship between total score and subject matter knowledge for the CCAT group or between total score and research design knowledge for either group.

Reliability, based on total score, was calculated in SPSS version 18.02 (SPSS, Chicago, IL, USA) using the ICC for multiple raters. Reliability for the IA group showed an ICC for consistency of 0.84 and for absolute agreement of 0.76 (Table 1a). The CCAT group had an ICC for consistency of 0.89 and for absolute agreement of 0.88.

A G study (Table 1b), using G_String_III,[32] demonstrated where error occurred in the total scores. The IA group had 76% of variance attributable to the paper, 10% attributable to the rater and 14% attributable to paper × rater interaction. The CCAT group had 88% of variance attributable to the paper, 1% attributable to the rater and 11% attributable to paper × rater interaction. Taking an *a priori* minimum acceptable G coefficient of 0.75, a D (decision) study (Table 1c) showed that in the IA group, three raters would be required to achieve the relative G coefficient and five raters would be required for the absolute G coefficient. In the CCAT group, two raters would be required to achieve both the relative and absolute G coefficients.

Table 1 Reliability (total score %, k = 5, *n* = 5)

(a) Intraclass correlation coefficient (ICC)

	IA group	CCAT group
Consistency	0.84	0.89
Absolute agreement	0.76	0.88

(b) G study

	IA group	CCAT group
Paper	76	88
Rater	10	1
Paper × rater	14	11

(c) D study

	IA group		CCAT group	
k	$E\rho^2$	Φ	$E\rho^2$	Φ
1	0.51	0.38	0.62	0.60
2	0.68	0.55	0.76	0.75
3	0.76	0.65	0.83	0.82
4	0.81	0.71	0.87	0.86
5	0.84	0.76	0.89	0.88
6	0.86	0.79	0.91	0.90
7	0.88	0.81	0.92	0.91
8	0.89	0.83	0.93	0.92
9	0.90	0.85	0.94	0.93
10	0.91	0.86	0.94	0.94

k, no. raters per group; *n*, no. papers per rater; IA, informal appraisal; CCAT, Crowe Critical Appraisal Tool; $E\rho^2$, relative G coefficient; Φ, absolute G coefficient.

Table 2 Analysis of covariance (total score %)

Main effects	IA group				CCAT group			
	F	Sig	Part η^2	*f*	*F*	Sig	Part η^2	*f*
Subject matter knowledge (*df* 1,18)	7.03	0.02	0.28	0.63	0.33	0.57	0.02	0.14
Research design knowledge (*df* 1,18)	2.34	0.14	0.12	0.36	1.18	0.29	0.06	0.26
Rater (*df* 4,18)	4.57	0.01	0.50	1.00	0.27	0.89	0.06	0.25

CCAT, Crowe Critical Appraisal Tool; *df*, degrees of freedom; *f*, effect size; *F*, *F* statistic; IA, informal appraisal; Part η^2, partial eta squared; Sig, significance; α, 0.05 one-tailed.

Analysis of covariance (Table 2) was used to determine whether raters (considered a random factor) were influenced by their subject matter knowledge or research design knowledge in appraising each paper. Assumptions of independence, normality, linearity, homogeneity and independence of covariates were met before analysis of covariance was undertaken. There were significant results in the IA group for subject matter knowledge ($F(1,18) = 7.03$, $P < 0.05$ one-tailed, partial $\eta^2 = 0.28$) and rater ($F(4,18) = 4.57$, $P < 0.05$ one-tailed, partial $\eta^2 = 0.50$). There were no significant results for the CCAT group.

Discussion

Even though both groups had the same average score, the range for the IA group was narrower than that for the CCAT group; the CCAT group had a lower minimum and higher maximum scores. Therefore, it could be concluded that the CCAT had better discriminatory power than informal appraisal. In other words, finer distinctions could be made between papers using the CCAT.

With regards to reliability, it was expected that the scores from CCAT group would be more reliable than the IA group because there was a more structured approach to appraising the papers. This expectation was borne out with the CCAT group having an ICC for consistency 0.05 higher than the IA group and an ICC for absolute agreement which was 0.12 higher than the IA group. Furthermore, the CCAT almost eliminated the rater effect (variance in total scores because of variability in how a rater scored a paper), with the CCAT group having a rater effect of 1% and the IA group's was 10%. Also, the D study showed that fewer raters would be required to achieve similar reliability using the CCAT than using informal appraisal especially where absolute agreement was sought (two vs. five raters).

In the IA group, there was a significant subject matter knowledge effect ($f = 0.63$) and a weak positive Kendall's tau correlation between total score and subject matter knowledge ($\tau = 0.38$, $P = 0.03$). This meant that taking rater variance and research design knowledge variance into account, knowledge of subject matter had a significant effect on total scores for the IA group and the greater the rater's subject matter knowledge, the higher the score they will give a paper. The ANCOVA also reinforced the significant rater effect ($f = 1.00$) for the IA group, as was apparent in the G study, and also that the rater effect was larger than the subject matter knowledge effect. This was as expected considering that subject matter knowledge is a characteristic of a rater.

The G study, ANCOVA and D study results show that using the CCAT appeared to neutralise any effects the raters or their subject matter knowledge had on the appraisal of the research papers. In other words, using the CCAT instead of an informal appraisal of research papers should help raters with different subject matter knowledge reach similar conclusions about a paper. This, in turn, has the potential to reduce poor conclusions being drawn from research papers and may even improve the implementation of evidence into practice.

The results did not show what other characteristics of the raters, besides subject matter knowledge (a significant effect) or research design knowledge (no effect), influenced the IA group's appraisal of the research papers. The level of research experience, which was used to match pair participants, could not be used because fewer participants were recruited than initially hoped for and the method used to determine researcher experience was not validated. Another limitation of this study was the small number of papers appraised. The same result may not be found if a large number of papers were appraised. Future research should address these two issues.

Conclusion

For the researcher, the decision on whether to use a CAT or an informal appraisal of research papers is clear: a structured approach was better. The CCAT was developed from theory and empirical evidence to work across multiple research designs, has a substantial user guide and has a published body of score validity and reliability data. The CCAT was shown to reduce the influence raters and subject matter knowledge had on the research papers being appraised. Finally, by being a consistent and structured tool, using the CCAT may in turn lead to improved understanding of findings and application of the evidence.

Acknowledgement

The authors wish to thank Anne Jones (James Cook University) for her contribution to this paper.

References

1. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigour: alternative tool structure is proposed. *J Clin Epidemiol* 2011; **64**: 79–89.
2. Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijen J. *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for Those Carrying Out or Commissioning Reviews (CRD Report 4)*. York: University of York, 2001.
3. Jadad AR, Moore RA, Carroll D *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996; **17**: 1–12.
4. de Vet HCW, de Bie RA, van der Heijden GJMG, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodological criteria. *Physiotherapy* 1997; **83**: 284–9.
5. NHS Public Health Resources Unit. CASP: Critical Appraisal Skills Programme. 2010. Accessed 29 Jan 2011. Available from: http://www.sph.nhs.uk/what-we-do/public-health-workforce/resources/critical-appraisals-skills-programme/
6. Shea B, Grimshaw J, Wells G *et al*. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007; **7**: 10. doi:10.1186/1471-2288-7-10.
7. Tate RL, McDonald S, Perdices M, Togher L, Schultz R, Savage S. Rating the methodological quality of single-subject designs and n-of-1 trials: introducing the single-case experimental design (SCED) scale. *Neuropsychol Rehabil* 2008; **18**: 385–401.
8. Deeks JJ, Dinnes J, D'Amico R *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003; **7**: iii–x, 1–173.
9. Moyer A, Finney JW. Rating methodological quality: toward improved assessment and investigation. *Account Res* 2005; **12**: 299–313.
10. Heller RF, Verma A, Gemmell I, Harrison R, Hart J, Edwards R. Critical appraisal for public health: a new checklist. *Public Health* 2008; **122**: 92–8.
11. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999; **282**: 1054–60.
12. Kuper A, Lingard L, Levinson W. Critically appraising qualitative research. *BMJ* 2008; **337**: 687–9.
13. Burnett J, Kumar S, Grimmer K. Development of a generic critical appraisal tool by consensus: presentation of first round Delphi survey results. *Internet J Allied Health Sci Pract* 2005; **3**(1).
14. Maher CG, Sheerington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. *Phys Ther* 2003; **83**: 713–21.
15. Crowe M, Sheppard L. A general critical appraisal tool: an evaluation of construct validity. *Int J Nurs Stud* 2011; doi: 10.1016/j.ijnurstu.2011.06.004.
16. Crowe M, Sheppard L, Campbell A. Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *J Clin Epidemiol* 2011; doi:10.1016/j.jclinepi.2011.08.006.
17. Boeije H. A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Qual Quant* 2002; **36**: 391–409.
18. Dye JF, Schatz IM, Rosenberg BA, Coleman ST. Constant comparison method: a kaleidoscope of data. *Qual Rep* 2000; **4**: Available from http://www.nova.edu/ssss/QR/QR4-1/dye.html.
19. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999; **354**: 1896–900.
20. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001; **285**: 1992–5.
21. Ogrinc G, Mooney SE, Estrada C *et al*. The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: explanation and elaboration. *Qual Saf Health Care* 2008; **17** (Suppl. 1): i13–32.
22. Stroup DF, Berlin JA, Morton SC *et al*. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA* 2000; **283**: 2008–12.
23. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007; **19**: 349–57.
24. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* 2007; **4**: e296.
25. Campbell DT, Stanley JC. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin, 1966.
26. Creswell JW. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 3rd edn. Thousand Oaks, CA: Sage, 2008.
27. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*, 3rd edn. Upper Saddle River, NJ: Prentice Hall, 2008.
28. MacAuley D, McCrum E, Brown C. Randomised controlled trial of the READER method of critical appraisal in general practice. *BMJ* 1998; **316**: 1134–7.
29. Haadr M. Random.org: Random sequence generator. 2009. Accessed 29 Jan 2011. Available from: http://www.random.org/sequences/
30. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998; **17**: 101–10.
31. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects. 59th World Medical Association General Assembly; 2008; Seoul, South Korea; 2008.
32. Bloch R. G_String_III. 5.4.6 ed. Hamilton, ON: Programme for Educational Research and Development; 2010.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Additional material. A comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: a randomised trial.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.