The design and evaluation of a critical appraisal tool for qualitative and quantitative health research

Submitted by

Michael Crowe

MIT – National University of Ireland, Galway

BSc (Mgmt) – Trinity College Dublin, Ireland

ADMT – Dublin Institute of Technology, Ireland

in

May 2011

for the degree of

Doctor of Philosophy

in the school of

Public Health, Tropical Medicine and Rehabilitation Sciences, James Cook University

Supervisor Prof Lorraine Sheppard

Associate Supervisor Dr Alistair Campbell

Declarations

RELEASE

Except where otherwise acknowledge herein, this thesis is the author's own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education.

In accordance with James Cook University's *Intellectual Property Policy* (2010) all intellectual property created by the author in the course of undertaking this PhD project belongs to the author and the author owns copyright to this thesis.

COPYRIGHT

Every reasonable effort has been made to gain permission from and acknowledge the owners of copyright material. Any copyright owner who has been omitted or incorrectly acknowledged can contact the author to remedy this situation. Extracts of copyright permissions are included in Appendix A.

ETHICS

The research presented and reported in this thesis was conducted within the guidelines for research ethics outlined in the *National Statement on Ethical Conduct in Human Research* (2007), the *Australian Code for the Responsible Conduct of Research* (2007), and James Cook University's *Code for the Responsible Conduct of*

Research (2009). The proposed research received authorisation from James Cook University's Human Research Ethics Committee (approval number H3415). A copy of the approval is reproduced in Appendix B.

CONTRIBUTION OF OTHERS

The author was awarded a JCU scholarship for the duration of their study. No other financial assistance was received. The author received no substantial research design, statistical, data analysis, or technical assistance apart from that duly provided by a PhD supervisor or associate supervisor. Ms Margaret Bowden provided professional editorial assistance which was limited to Standards D (language and illustrations) and E (completeness and consistency) of the *Australian Standards for Editing Practice* (Council of Australian Societies of Editors, 2001) in line with James Cook University's *Proof-Reading and Editing of Theses and Dissertations* (2010) policy.

INCORPORATED PUBLISHED WORKS

The following articles by the author have been published or are under review for publication, and form an integral part of this thesis. Chapters where these articles appear are indicated. Where articles were co-authored, this was in co-operation with the PhD candidate's supervisor or associate supervisor. Co-authorship contributions were towards the concept, drafting and final approval of the published work, as duly provided by a PhD supervisor or associate supervisor. Articles that were published are reproduced in Appendix C.

Chapter 2 – Crowe, M., & Sheppard, L. (2010). Qualitative and quantitative research designs are more similar than different [Invited editorial]. *Internet Journal of Allied Health Sciences and Practice*, *8*(4). Retrieved from http://ijahsp.nova.edu/

Chapter 3 – Crowe, M., & Sheppard, L. (2011). Mind mapping research methods. *Quality and Quantity*, (Online First). doi:10.1007/s11135-011-9463-8

Chapter 4 – Crowe, M., & Sheppard, L. (2010). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology*, *64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

Chapter 5 – Crowe, M., & Sheppard, L. (2011). A proposed critical appraisal tool shows good results compared to other tools: An evaluation of construct validity. *International Journal of Nursing Studies*, *14*(12). 1505-1516. doi:10.1016/j.ijnurstu.2011.06.004

Chapter 6 – Crowe, M., Sheppard, L. & Campbell, A. (2011). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology*, (Online). doi:10.1016/j.jclinepi.2011.08.006

Chapter 7 – Crowe, M., Sheppard, L. & Campbell, A. (2011). A comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: A randomised trial. *International Journal of Evidence-Based Healthcare*, *9*(4), 444-449. doi:10.1111/j.1744-1609.2011.00237.x

By signing these declarations, the author confirms to the best of their knowledge that the statements are accurate and true.

6 May 2011

Michael Crowe — Dated

Acknowledgements

Charlie and Dad, who were here for the beginning but not at the end.

Thanks to friends and colleagues at JCU for their encouragement throughout. Also, thanks to the participants in the research who volunteered their spare time to read and appraise papers.

My family and Mum, who think I'm a kind of 'eejit' for doing this in the first place. Even so, they have supported me all the way.

Lorraine Sheppard, generous, encouraging, invaluable, without your help I would still be thinking about doing a PhD rather than finishing. And Alistair Campbell, for helping me through decisions on design and analysis.

Most of all, my gratitude goes to Anne. Why we were doing PhDs at the same time is still a mystery. Your love, support, guidance, tenacity, and ability to cut through the bulldust are inspirational.

Abstract

OBJECTIVE

To design and evaluate a critical appraisal tool (CAT) that can assess the research methods used in a broad range of qualitative and quantitative health research papers; has the depth to fully assess these research papers; has an appropriate scoring system; and has validity and reliability data available to evaluate the scores obtained by the tool.

Critical appraisal is defined here as the impartial assessment of one or more research papers to determine their strengths, weaknesses and benefits.

STUDY DESIGN AND SETTING

The study was a sequential mixed methods research design where data collected in one phase informed the design and focus of the next. Data collection took place between July 2008 and September 2010 at James Cook University, Australia. There were two sections to the study: collection and synthesis of secondary data; and planning, collection and analysis of primary data.

The study began with an exploration of the divide between qualitative and quantitative research. This showed that the divide is more an historical distinction than a current one. As such, there are no theoretical impediments for a single

qualitative and quantitative research CAT. The scope of research methods was examined next through the use of mind maps. This exploration was required so that the design of a CAT could be situated within an overall understanding of research methods. A critical review of how CATs are designed was the final part of secondary data analysis. This review of 45 papers informed the design of the proposed critical appraisal tool, which was based on empirical evidence and the nature of research methods rather than subjective or biased assessments of what a critical appraisal tool could include.

The first part of the primary data collection was an exploratory study of the validity of the scores obtained by the proposed CAT. A random selection of 60 health research papers were analysed using the proposed CAT and five alternative CATs. Next was an exploratory study of reliability, where the proposed CAT was used by five raters, each of whom appraised 24 randomly selected research papers. The final part was to test whether using a CAT was an improvement over using no CAT to appraise research papers because there is little empirical evidence to show if this is true. A total of ten raters were randomly assigned to two groups and they appraised a random selection of five health research papers. One group used the proposed CAT, while the other group did not use any CAT.

RESULTS

Critical review – Explanations on how a critical appraisal tool was designed and guidelines on how to use the CAT were available in five (11%) out of 45 papers evaluated. Thirty-eight CATs (84%) reported little or no validity evaluation and 33 CATs (73%) had no reliability testing. The questions and statements which made up each CAT were coded into a proposed CAT with eight categories, 22 items, and 98 item descriptors, such that each category and item was distinct from every other.

Validity – In all research designs, the proposed CAT had significant ($p < 0.05$, 2-tailed) weak to moderate positive Kendall's tau correlations with the alternative CATs ($0.33 \leq \tau \leq 0.55$), except in the *Preamble* category. There were significant moderate to strong positive correlations in true experimental ($0.68 \leq \tau \leq 0.70$); quasi-experimental ($0.70 \leq \tau \leq 1.00$); descriptive, exploratory or observational ($0.72 \leq \tau \leq 1.00$); qualitative ($0.74 \leq \tau \leq 0.81$); and systematic review ($0.62 \leq \tau \leq 0.82$) research designs. There were no significant correlations in single system research designs.

Reliability – The intraclass correlation coefficient (ICC) for all research papers was 0.83 for consistency and 0.74 for absolute agreement using the proposed CAT. The G study showed a majority paper effect (53–70%) for each research design, with small to moderate rater effects or paper × rater interaction effects (0–27%).

Compare CAT with no CAT – The ICC for absolute agreement was 0.76 for the group not using a CAT and 0.88 for the proposed CAT group. A G study showed that the group not using a CAT had a total score variance of 24% attributable to either the rater or paper × rater interactions, whereas in the proposed CAT group this variance was 12%. Analysis of covariance (ANCOVA) showed that there were significant effects in the group not using a CAT 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 $\eta^2 = 0.50$).

DISCUSSION

Critical review – Many CATs have been developed based on a subjective view of research quality rather than on evidence for what elements should or should not be included in a critical appraisal of research. When choosing a CAT, researchers should: (1) take into account the context of the appraisal; (2) determine whether the

CAT was developed using the best evidence available; (3) ensure that the validity of the scores obtained from the CAT can be verified; and (4) analyse the scores obtained from the CAT for reliability.

Validity – The proposed CAT exhibited a good degree of validity based on the theory the CAT was built, the collection of empirical evidence, and the stated context for its use. Therefore, inferences made based on the scores obtained using the proposed CAT should reflect the value of the papers appraised.

Reliability – Given the assessment of validity and the reliability scores obtained, the proposed CAT appears to be a viable tool that can be used across a wide range of research designs and appraisal situations. Any variability in the scores obtained using the proposed CAT can be explained by the diverse subject matter of papers and participants' unfamiliarity with some research designs. Difficulties with subject matter and research designs are less likely in normal use of the proposed CAT where raters are more familiar with the subject matter and research designs used.

Compare CAT with no CAT – The proposed CAT was more reliable than not using a CAT when appraising research papers. In the group not using a CAT there were significant effects for rater and subject matter knowledge. In the proposed CAT group the rater effect was almost eliminated and there was no subject matter knowledge effect. There was no research design knowledge effect in either group.

CONCLUSION

A CAT was designed and evaluated, which met the aim and objectives of the study. The proposed CAT can be used across a broad range of qualitative and quantitative health research; has the depth to fully assess research papers; has an appropriate scoring system; and has validity and reliability data available. Further research can

extend the proposed CAT to determine whether it is useful in criterion-referencing health research and general research. Furthermore, the proposed CAT can be applied to the increased use of mixed and multiple research methods, and be used to assess, understand and communicate this research knowledge.

Table of contents

List of tables

List of figures

Symbols and abbreviations

$E\rho^2$ – Relative generalizability coefficient

Also: Relative G coefficient; Generalizability coefficient; G coefficient

Alternative: $E\rho^2_\delta$; ρ^2_δ.

Φ – Absolute generalizability coefficient

Also: Absolute G coefficient; Index of dependability

Alternative: ρ^2_Δ

AERA – American Educational Research Association

AMSTAR – Assessment of Multiple Systematic Reviews

APA – American Psychology Association

CASP – Critical Appraisal Skills Programme

CAT – Critical appraisal tool

CCAT – Crowe Critical Appraisal Tool

CEBM – Centre for Evidence-Based Medicine

CHE – Centre for Health Evidence

CONSORT – Consolidated Standards for Reporting of Trials

COREQ – Consolidated Criteria for Reporting Qualitative Research

CRD – Centre for Reviews and Dissemination

CTT – Classical Test Theory

D study – Decision study

DEO – Descriptive, exploratory or observational research designs

DOI – Digital object identifier

EBP – Evidence-based practice

EMS – Expected mean square

G coefficient – Generalizability coefficient

G study – Generalizability study

G theory – Generalizability theory

IA – Informal appraisal

ICC – Intraclass correlation coefficient

IRT – Item response theory

JCU – James Cook University

MOOSE – Meta-analysis of Observational Studies in Epidemiology

NCME – National Council on Measurement in Education

NCMUE – (OBSOLETE, SEE NCME) National Council on Measurements Used in Education

NHMRC – National Health and Medical Research Council

PCAT – Proposed critical appraisal tool

PEDro – Physiotherapy Evidence Database

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-analyses

QUOROM – (OBSOLETE, SEE PRISMA) Quality of Reporting of Meta-analyses

RCT – Randomised controlled trial

RSS – Really Simple Syndication

SQUIRE – Standards for Quality Improvement Reporting Excellence

STROBE – Strengthening the Reporting of Observational Studies in Epidemiology

Definitions

Critical appraisal

The impartial assessment of one or more research papers to determine their strengths, weaknesses, and benefits. Where,

1. Strengths – Suitability of research methods to answer the research question.
2. Weaknesses – Identification and, where possible, reduction of limitations due to research methods.
3. Benefits – Implications based on sound conclusions drawn from the research methods used, results obtained, and current evidence.

Critical appraisal tool

A structured approach to critical appraisal.

Research design

The basic approach or approaches used to answer a research question, such as true experimental or phenomenological designs. Research design is one element of research methods.

Research methodology

The philosophical (ontological) and theoretical (epistemological) basis for research designs.

Research methods

The overall process of initiating, implementing, analysing, and reporting research. The term is always used in the plural. Elements of research methods are research question, research design, sampling techniques, ethical matters, data collection, data analysis, and report findings.

RSS (Really Simple Syndication)

A standardised method to periodically and automatically download frequently updated information from a source connected to the internet. Also known as a feed, web feed, or channel.

Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality.

Edward Thorndike (1918)

Chapter 1 – Introduction

In an ideal world, every research paper would be entirely relevant to the issue being explored and the research methods used would have no flaws or limitations. However, it is unlikely that a research paper will address exactly the problem being investigated and all research has flaws [1]. Therefore, the ability to sift through and critically appraise research papers are essential skills.

1.1 Critical appraisal

Critical appraisal is used in systematic reviews, teaching, and journal clubs [2, 3]. Two basic questions need to be answered in order to critically appraise a research paper:

1. Is this paper relevant?
2. Is this paper any good?

Whether a research paper is relevant to a researcher's needs depends on the problem being explored. This issue is not discussed here. However, determining whether a research paper is any good is the core of critical appraisal. But what is meant by, 'Is this research paper any good?'

A good research paper could be interpreted as a research paper that has quality. But then, 'What is quality?' Quality tends to be a subjective concept. If you ask any group of people about the quality of a paper, film or apple sauce, you will get many of

different opinions. Thus, saying a good research paper is a quality research paper only substitutes one concept for another equally undefined concept.

The subjective aspect of critical appraisal can be removed by concentrating on objective concepts such as whether a research paper has internal and external research validity. Common definitions of these concepts are [4 (p. 130), 5 (pp. 176-185)]:

1. *Internal research validity* – The research methods used in experimental designs that are affected by the extent to which changes in the dependent variable (outcome, predictor) can be attributed to the independent variable (intervention, treatment, exposure).
2. *External research validity* – Whether the study results can be generalised to other environments or people outside those used in the study, and is affected by sample (size and method), research design, and measures used.

If internal research validity is limited to experimental research designs, this means that the vast majority of research designs such as quasi-experimental, single system, descriptive, exploratory or observational, qualitative, mixed methods, and systematic review cannot be assessed for internal validity because they do not compare dependent and independent variables. Meanwhile, external research validity is not possible for many descriptive, exploratory or observational designs, or qualitative research because, these research designs are dependent on their context and may not be generalisable [6].

Therefore, critical appraisal tends to be defined in a way that circumvents the use of restrictive terms, such as 'internal and external validity', and overly expansive terms, such as 'quality'. The definition of critical appraisal for the purposes of this thesis, based on the research undertaken by the author and other sources [7, 8] is:

The impartial assessment of one or more research papers to determine their strengths, weaknesses, and benefits.

Where,

1. *Strengths* – Suitability of research methods to answer the research question.
2. *Weaknesses* – Identification and, where possible, reduction of limitations due to research methods.
3. *Benefits* – Implications of the research based on sound conclusions drawn from the research methods used, results obtained, and current evidence.

For this thesis, research methods means a combination of research design, sampling techniques, ethical matters, data collection, and data analysis.

1.2 CRITICAL APPRAISAL TOOLS

The definition of critical appraisal does not indicate any practical method to critically appraise research papers. In essence, the practice of critical appraisal is achieved through the use of critical appraisal tools. Or put another way, a critical appraisal tool is a structured approach to critical appraisal.

Critical appraisal tools generally take the form of a series of questions or statements that a reader uses to assess research papers. There are hundreds of critical appraisal tools [9, 10, 11, 12]. However, many tools:

1. Were developed for one or a limited number of research design(s) [10, 13].
2. Lack the depth to properly assess research papers, in so far as many critical appraisal tools use statements or questions that are relatively easy to quantify (for example, internal research validity) and ignore concepts that are harder to quantify, such as whether the research design is appropriate for the research question [10, 14].

3. Have inappropriate scoring systems that may hide defects in the research paper being assessed [15, 16, 17, 18].
4. Have no information on the validity or reliability of data collected and, therefore, cannot claim to assess a research paper accurately [19, 20, 21].

These limitations highlight the need to develop and evaluate a more comprehensive and methodologically sound critical appraisal tool.

1.3 AIM AND OBJECTIVES

The overall aim of this research, based on the identified research problem, was:

> To design and evaluate a critical appraisal tool that can assess the research methods used in a broad range of qualitative and quantitative health research papers; has the depth to fully assess these research papers; has an appropriate scoring system; and has validity and reliability data available to evaluate the scores obtained by the tool.

Six objectives were identified to achieve this aim:

1. Determine whether a critical appraisal tool can be used for both qualitative and quantitative research papers.
2. Examine the features of research methods so that their variety was understood before developing a critical appraisal tool.
3. Critically review the literature on the design of existing critical appraisal tools and use this information to create a proposed critical appraisal tool.
4. Refine the initial draft of the proposed critical appraisal tool, develop a scoring system, and evaluate the validity of the scores obtained by the proposed critical appraisal tool.
5. Examine the reliability of the scores obtained by the proposed critical appraisal tool.

6. Compare structured critical appraisal, using the proposed critical appraisal tool, with informal appraisal (not using a critical appraisal tool) when appraising research papers.

Analysis of available literature was used to explore the first three objectives, which contributed primarily to the design of the critical appraisal tool. The remaining three objectives required the collection of primary data and contributed mainly to the evaluation of the critical appraisal tool through validity and reliability testing.

Expending this effort in design, validity and reliability may enable the development of a standardised critical appraisal tool. In other words, a tool that can be administered in a consistent manner and the scores given to health research papers by different raters can be interpreted in a consistent manner. In the short term, this would mean that researchers could be confident that the critical appraisal tool would act as an accurate measure of health research, which is the scope of this thesis. In the longer term, the critical appraisal tool could become a means for criterion-referencing health research papers [22 (pp. 49-52)]. This outcome would be the subject for further research.

1.4 LIMITS TO RESEARCH SCOPE

There were limits to the research scope that may be obvious from the aims and objectives but are highlighted here. Firstly, the research was limited to health research. This was because of the emphasis in health disciplines on the critical appraisal of research and the use of systematic reviews. Also, the author is situated in the School of Public Health, Tropical Medicine and Rehabilitation Science in James Cook University so limiting the scope to health research was a natural boundary.

Second, the research scope was limited to research papers. This means papers with a focus on exploring a research question, using stated research methods, and published in an academic journal. The research papers did not need to be peer reviewed. It also meant that some texts which may be used in a systematic review were excluded. These may include grey literature, published reports, books and book chapters, and magazine or paper articles.

Third, the development of a critical appraisal tool was not considered for a number of specific purposes within health. These areas require professional or specialist knowledge the author did not have and could not be reasonably expected to acquire over the course of a PhD project. The areas include: medical diagnostics/ prognostics; econometrics and health economics; clinical practice guidelines; health service delivery; and quality assurance or assessment of health programs.

1.5 KEY ASSUMPTIONS

OR: HOW I LEARNED TO STOP WORRYING AND LOVE THE QUESTION

[Apology to Kubrick, S. (1964). *Dr Strangelove or: How I learned to stop worrying and love the bomb*]

Taking a leaf from qualitative research [23 (pp. 63-66)], particularly self-reflection, the author's underlying beliefs on research methods should be made clear. These beliefs have influenced how this research was approached and place the author in the pragmatic school of research. This lean towards pragmatism becomes clear in Chapters 2 and 3, but it is best to declare it early.

When the author completed an undergraduate degree in marketing and management in the early 1990s there was a large emphasis on research methods, yet a divide between qualitative and quantitative research was never emphasised. It came as a surprise, in 2008, that an argument was taking place within health

research about a fundamental, unassailable difference between qualitative and quantitative research. As a result, it was necessary to show that disagreements regarding the nature of qualitative and quantitative research have been dispensed with by other academic fields. Therefore, the assessment of qualitative and quantitative research papers in a single critical appraisal tool was not only possible but permissible.

Similarly, if there is no divide between qualitative and quantitative research, there can be no divide within quantitative research. How, then, can a single hierarchy of evidence and a gold standard for research design exist [24]? To assert that randomised controlled trials (RCTs) are the top of a hierarchy and a gold standard of research is to declare that RCTs are always possible, always ethical, and should be chosen ahead of any other research design (see also Chapters 2 and 3). This stance is patently unsustainable. Where stringent ethical procedures exist, for example, approval could not be given to an RCT on the effects of smoking tobacco. Or, more immediately, since there are no RCTs showing that a parachute will help prevent death or serious injury, who will be the first person to give their fully informed consent and volunteer for such a study [25]? The pragmatic view is to let the research question determine the research design rather than allowing a research design to dictate which research questions can be asked or answered.

1.6 THESIS STRUCTURE

Each chapter, except for this introduction and the final chapter, takes the form of a journal article that has been published or is in the publication process. Differences between articles as they appear in this thesis and published articles include:

1. The thesis uses a single, consistent referencing style whereas the published papers may have many different referencing styles.

2. Spelling, grammar, or other errors that were missed in published papers have been corrected, although no change has been made to the underlying ideas or structure.
3. Where a paper is yet to be published, the requirements of the publisher may require the format to be different to that included here.

There is some repetition between chapters as a consequence of submitting the thesis by publication. However, the repetition is limited and was left in place so that a reader interested in a particular chapter could go directly to that chapter without knowledge of previous chapters. Similarly, an *Additional material* section is presented after the *References* in Chapters 4–7. This material is integral to the thesis because it contains items such as search strategies, alternative CATs, and versions of the user guide written for the developed CAT. Score validity cannot be assessed fully without including these items in the main body of the thesis (see Chapter 5). This may be tedious for the reader but your understanding of this assessment requirement is sought.

Chapter 2 is an overview of the common philosophical arguments often used to divide qualitative from quantitative research. The chapter outlines why many of these arguments are misleading, and why qualitative and quantitative research are more similar than different. The result is that qualitative and quantitative research can be considered together in a single critical appraisal tool.

Chapter 3 focuses on identifying the breadth of research methods. It explores the different aspects of research methods, including the research problem, research designs, sampling techniques, ethical matters, data collection, data analysis, and reporting findings. This understanding of research methods can then be used to design a better critical appraisal tool.

A critical review of critical appraisal tools is described in Chapter 4. The review targeted papers where the primary aim was to explain or evaluate the design of a critical appraisal tool. From this, the building blocks of a new critical appraisal tool are developed.

Chapter 5 is the first of the primary research papers. The chapter evaluates score validity for the proposed critical appraisal tool. The chapter describes what is meant by validity, a concept that is often misunderstood in the literature. Extensive use of the *Standards for educational and psychological testing* is made from a theoretical and practical point of view.

Chapter 6 builds on the results from the previous chapter and examines score reliability for the proposed critical appraisal tool. Instead of limiting analysis to classical test theory, generalizability theory is used in this and the next chapter to better understand where errors occur in the appraisal of research papers. Generalizability theory was developed in the United States of America and consequently the method is spelt with a 'z'. Otherwise, Australian spelling standards are followed.

Chapter 7 investigates whether a structured approach is an improvement over an informal approach to critical appraisal because there are few studies which show whether this is the case. The chapter also investigates whether subject matter knowledge or research design knowledge affect critical appraisal.

Chapter 8 explores the conclusions that can be drawn from the research, acknowledges any limitations of the research, and indicates potential future studies.

1.7 REFERENCES

1. Simon, S. D. (2001). Is the randomized clinical trial the gold standard of research? *Journal of Andrology, 22*(6), 938-943.

2. D'Auria, J. P. (2007). Using an evidence-based approach to critical appraisal. *Journal of Pediatric Health Care, 21*(5), 343-346. doi:10.1016/j.pedhc.2007.06.002

3. Kastelic, J. P. (2006). Critical evaluation of scientific articles and other sources of information: An introduction to evidence-based veterinary medicine. *Theriogenology, 66*(3), 534-542. doi:10.1016/j.theriogenology.2006.04.017

4. Polgar, S., & Thomas, S. A. (2007). *Introduction to research in the health sciences* (5th ed.). Edinburgh: Churchill Livingstone.

5. Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

6. Trochim, W. M. (2006). The research methods knowledge base. Retrieved 29 January 2011, from http://www.socialresearchmethods.net/kb/

7. Avis, M. (1994). Reading research critically. I. An introduction to appraisal: Designs and objectives. *Journal of Clinical Nursing, 3*(4), 227-234. doi:10.1111/j.1365-2702.1994.tb00393.x

8. Earl-Slater, A. (2001). Critical appraisal and hierarchies of the evidence. *British Journal of Clinical Governance, 6*(1), 59-63. doi:10.1108/14664100110385154

9. Armijo Olivo, S., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156-175. doi:10.2522/ptj.20070147

10. Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitc, C., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27). doi:10.3310/hta7270

11. Katrak, P., Bialocerkowski, A., Massy-Westropp, N., Kumar, V. S. S., & Grimmer, K. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology, 4*(1). doi:10.1186/1471-2288-4-22

12. Sanderson, S., Tatt, I. D., & Higgins, J. P. T. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36*(3), 666-676. doi:10.1093/ije/dym018

13. Khan, K. S., ter Riet, G., Glanville, J., Sowden, A. J., & Kleijen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews (CRD Report 4). York, England: University of York.

14. Moyer, A., & Finney, J. W. (2005). Rating methodological quality: Toward improved assessment and investigation. *Accountability in Research, 12*(4), 299-313. doi:10.1080/08989620500440287

15. Heller, R. F., Verma, A., Gemmell, I., Harrison, R., Hart, J., & Edwards, R. (2008). Critical appraisal for public health: A new checklist. *Public Health, 122*(1), 92-98. doi:10.1016/j.puhe.2007.04.012

16. Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1060. doi:10.1001/jama.282.11.1054

17. Kuper, A., Lingard, L., & Levinson, W. (2008). Critically appraising qualitative research. *BMJ, 337*(7671), 687-689. doi:10.1136/bmj.a1035

18. Walsh, D., & Downe, S. (2006). Appraising the quality of qualitative research. *Midwifery, 22*(2), 108-119. doi:10.1016/j.midw.2005.05.004

19. Bialocerkowski, A. E., Grimmer, K. A., Milanese, S. F., & Kumar, S. (2004). Application of current research evidence to clinical physiotherapy practice. *Journal of Allied Health, 33*(4), 230-237.

20. Burnett, J., Kumar, S., & Grimmer, K. (2005). Development of a generic critical appraisal tool by consensus: Presentation of first round Delphi survey results. *Internet Journal of Allied Health Sciences and Practice, 3*(1), 22. Retrieved from http://ijahsp.nova.edu/

21. Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

22. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

23. Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage.

24. Guyatt, G. H., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J., & Cook, R. J. (1995). Users' guides to the medical literature: IX. A method for grading health care recommendations. *JAMA, 274*(22), 1800-1804. doi:10.1001/jama.1995.03530220066035

25. Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ, 327*(7429), 1459-1461. doi:10.1136/bmj.327.7429.1459

Chapter 2 – Qualitative and quantitative research

This chapter explores the similarities and differences between qualitative and quantitative research. It shows that the differences are minor, and therefore qualitative and quantitative research can be assessed in the same critical appraisal tool, thereby meeting Objective 1 of the study.

The chapter consists of an article accepted for publication on 1 September 2010 and available online 1 October 2010 (Appendix C.1):

> Crowe, M., & Sheppard, L. (2010). Qualitative and quantitative research designs are more similar than different [Invited editorial]. *Internet Journal of Allied Health Sciences and Practice*, *8*(4). Retrieved from http://ijahsp.nova.edu/

Changes have been made to the published article to ensure thesis consistency. Copyright permission, which allows this paper to be reproduced, can be found in Appendix A.1.

Qualitative and quantitative research designs are more similar than different

2.1 ABSTRACT

The qualitative/quantitative divide has been extensively debated in social science and educational research. However, health researchers are still bound by traditional distinctions between qualitative and quantitative research. This paper argues that although these distinctions were valid at the turn of the 20th century, they no-longer hold true. Advances in both qualitative and quantitative methods, and the need to explore increasingly complex situations, mean it is more important to concentrate on how best to answer the research question rather than focusing on the research design being used.

2.2 INTRODUCTION

There has been a traditional divide between qualitative and quantitative research, and nothing can start, continue or inflame an argument among research theorists than saying, with fundamentalist glee and certitude that, 'My research design is better than yours'. However, this chapter is not an exercise in fundamentalism. Nor is it meant as an exhaustive discussion of research methodology. Instead, this is a brief look at the topic where the arguments and content are kept purposely simple because this type of discussion can quickly become a morass of jargon.

The argument put forward here is that the distinction between qualitative and quantitative research may have had validity at the turn of the 20th century, but as ideas about research have continued to evolve and develop the distinction has become more historical than actual [1]. Whether research is qualitative or

quantitative, the techniques are far more similar than they are different and, by maintaining the myth of incompatibility, researchers may miss important ways of finding answers to their research questions [2].

The reasons often forwarded for why qualitative and quantitative research are fundamentally different generally reduce to four areas: (1) Research methodology; (2) Context, values, and involvement; (3) Data, analysis, and participants; and (4) A common error. Each of these areas is taken in turn and the assumptions exposed.

2.3 RESEARCH METHODOLOGY

Research methodology is most often described as the overall philosophy underpinning research, whereas research methods are the practical guidelines or techniques used to produce research [3]. Research methodology is covered here in just enough depth to debunk the differences between qualitative and quantitative research that are commonly stated. Those differences can be described as realism versus idealism; causality versus interpretation; and hypotheses versus description.

A very basic definition of realism is, 'Things exist only in the real world' and, therefore, anything that cannot be observed through the senses is of no consequence. On the other hand, a basic definition of idealism is, 'Things exist only within the mind' and, therefore, are open to interpretation [4, 5]. Realism is stated as the concept underpinning quantitative research, while idealism is the concept that is said to underpin qualitative research [5]. But is this correct?

A brief thought about both definitions shows that a reasonable person can come up with examples whereby the basic definitions do not hold. For example, 'Do thoughts exist?' – Yes – Therefore the realism definition lacks completeness. 'If humans

disappeared in the morning, would the world still exist?' – Yes, it existed before humans and it would be conceited to think it would end just because humanity ends – Therefore the idealism definition lacks completeness.

There are, of course, a multitude of definitions for realism and idealism [6]. Why, then, is quantitative research said to be realist, and why is qualitative research said to be idealist? It basically comes down to the assumptions made about the nature of reality, and since philosophers have been arguing about this for thousands of years without coming to a conclusion, it is unlikely that researchers will arrive at a conclusion any time soon. So the same argument of idealist versus realist is rolled out based on no real evidence [5, 6, 7]. There is an alternate position, however, where some authors have suggested that whether a researcher uses qualitative or quantitative techniques, they are in fact most likely to be critical realists, meaning that some of our perceptions accurately represent the world as it is and some of our perceptions do not represent the world as it is [8].

The second part on the philosophy of research is causality versus interpretation. Until the early 20th century, many scientists searched for causality, also known as cause and effect. However, in the 1930s with the ideas of quantum mechanics and relativity gaining more ground, the Newtonian, mechanistic view of the world changed [1, 2, 9]. Suddenly, scientists could no longer be sure of causality because observation of a phenomenon could change the nature of that phenomenon. It became "...questionable to what extent causality is of scientific interest" [10 (p. 308)]. And so the quantitative side no longer look for causality but deal in probabilities and correlations, and the predictive value of these. Even so, causality, as the purpose of quantitative research, is still put forward as a stumbling block between qualitative and quantitative research [11].

What of qualitative research? The emphasis is on the interpretation of how social reality is constructed, or the cultural or other meaning of phenomena experienced by those who are in a study [12, 13]. However, the assumption that quantitative data do not need interpretation but simply manifest meaning through mathematical means while only qualitative research requires interpretation is not true. Quantitative and qualitative researchers need to make judgements about their data in order to elicit new meaning, extract alternative meanings, and interpret results based on previous research. Any research without interpretation is simply disaggregate [14, 15].

This leads directly into the area of theory and, by extension, hypotheses and deductive reasoning versus description, and by further extension, inductive reasoning. The qualitative camp states that data collected can only describe the situation as it is and that no theories can be developed. Where then does this leave the branch of qualitative research called Grounded Theory, whereby theories are developed based on the data collected [16]? The quantitative camp argues that first you need to have a theory and from that you develop a hypothesis to be tested. Yet there are examples where this is not so. If you look at surveys, a quantitative technique, there is no need for a hypothesis or a theory – the point of a survey is to find out information and not to test a hypothesis [17].

Both sides use a combination of inductive and deductive reasoning. In fact, it is not too difficult to see that induction and deduction are parts of the same process (Figure 2.1) [18]. For example, in epidemiology, diseases are observed, patterns of disease are detected, tentative hypotheses are postulated about the underlying cause or causes, and theories of the disease are formulated – all inductive reasoning. From there, the theory is tested based on exposure and non-exposure, results are observed, and the theory of the disease is rejected or not rejected – all this is deductive. So why should there be a limit to our understanding based on one type of reasoning over the other?

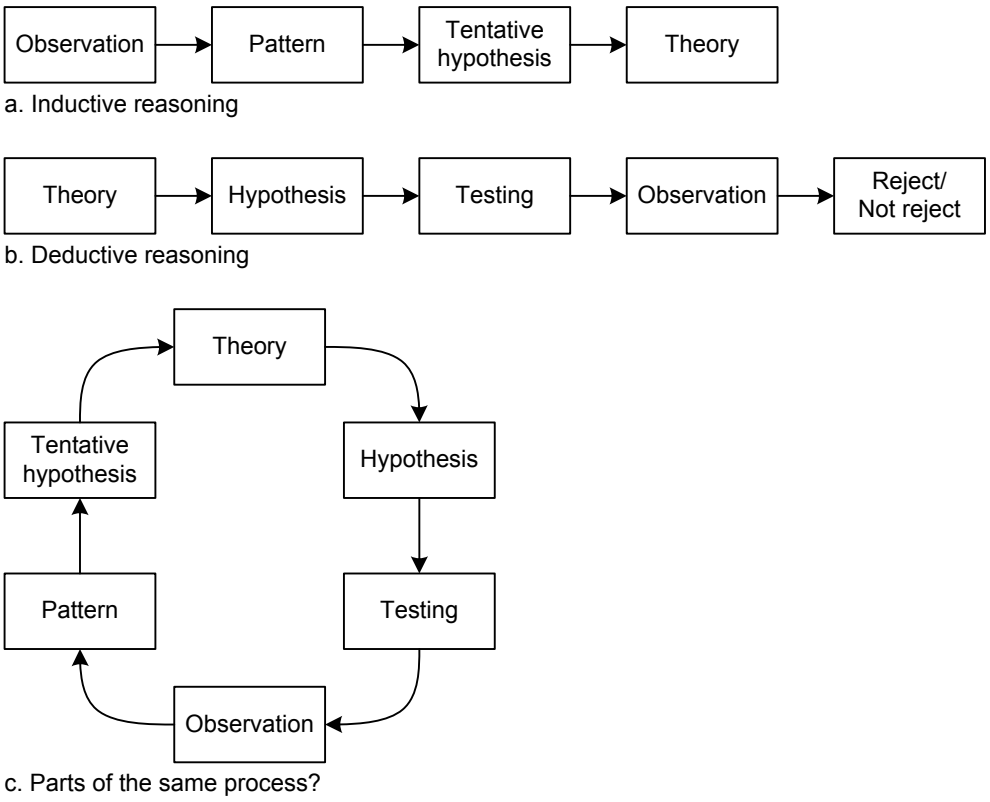

Figure 2.1 Inductive and deductive reasoning

2.4 CONTEXT, VALUES, AND INVOLVEMENT

The second set of differences between qualitative and quantitative research can be summed up as context, values, and involvement. In the quantitative camp, research is supposed to be conducted independent of context, free of societal or cultural values, and the researcher is detached from or not involved in the process. In the qualitative camp, the research is said to be context dependent, societal and cultural values are present and explicitly stated, and the researcher is involved in the process [5, 11].

However, all research has a context. Quantitative research can attempt to control for this by limiting the context through controlling variables, but in some quantitative techniques, such as developmental studies, this is not possible. Thus, the context of the research becomes more important. Qualitative research does not attempt to

control for context, and it is through the context of the research that the research gains value. However, qualitative research cannot always be said to happen in a naturalistic setting, for example focus groups [2, 19]. Therefore, whether a researcher has a qualitative or quantitative focus, they approach the problem by creating a controlled environment, either purposely or as a natural consequence of their actions, in order to accomplish their research, which is then extrapolated to a more complex environment or real world situation [8].

On the topic of societal and cultural values, no research is value free [11]. Researchers bring their own ideas, influences, and personalities into the research project: after all, researchers are human. Quantitative research has learned from qualitative research in recognising that these things are present and need to be accounted for in the way research is conducted. In fact, there has been a movement in health research to publicly register randomised controlled trials before they begin so that the procedure, influences, and the veracity of the results can be publicly determined [20].

Equally, qualitative and quantitative researchers are deeply involved in their own research. It is more a matter of when this involvement occurs. The quantitative researcher's involvement is notionally suspended as the data are collected: the researcher does not influence or attempt to interact with participants in any way that may affect the results. However, this has more to do with introducing as few biases as possible into the research than a lack of wanting to be involved. Meanwhile, the qualitative researcher is notionally involved the whole way through their research, although this is not necessarily the case in large projects where the researcher may not be involved in all or any data collection [2, 5].

2.5 DATA, ANALYSIS, AND PARTICIPANTS

The third group of differences put forward are that qualitative research uses words as the data, thematic analysis of the data, and has few participants, whereas quantitative research uses numbers as the data, statistical analysis, and has many participants. Again, on the surface, this appears true but with a little digging the distinction is difficult to maintain.

Quantitative studies can have one participant, for example a case study or single-subject design [17]. Similarly, the reason qualitative studies can have fewer numbers of participants may be a matter of thematic saturation: participants are recruited until themes presented by participants have been voiced or exhibited on at least two occasions [12]. In this way, recruitment to qualitative studies is based on the information retrieved from interviews and observation rather than on a predetermined sample size.

In reference to data and analysis, the distinction is a matter of precision rather than use of different data and analysis. If the research requires a high degree of precision, then numbers and statistical analysis may be the requirement. However, if the degree of precision is not as important and the participants' views are of more value, then the use of words and thematic analysis, or another qualitative analysis technique, are more useful. Also, where the topic being studied is too complex to reduce to quantitative data, it is better to allow that complexity to stand and to analyse the data in a qualitative manner. This is true whether the research technique employed is primarily qualitative or quantitative [2, 15].

2.6 A COMMON ERROR

Both camps make a common error that has been alluded to in the previous sections: they assume that there is only one facet to the other. In other words, the qualitative

camp appears to assume that there is one stereotypical quantitative design, and the quantitative side appears to assume that there is one stereotypical qualitative design. However, both camps have a number of research designs at their disposal (Figure 2.2). In the qualitative camp research designs include narrative, phenomenology, grounded theory, and ethnographic designs. In the quantitative camp there are different types of true experimental, quasi-experimental, descriptive, and observational research designs. Furthermore, in some cases qualitative and quantitative designs are used together, such as in mixed methods.

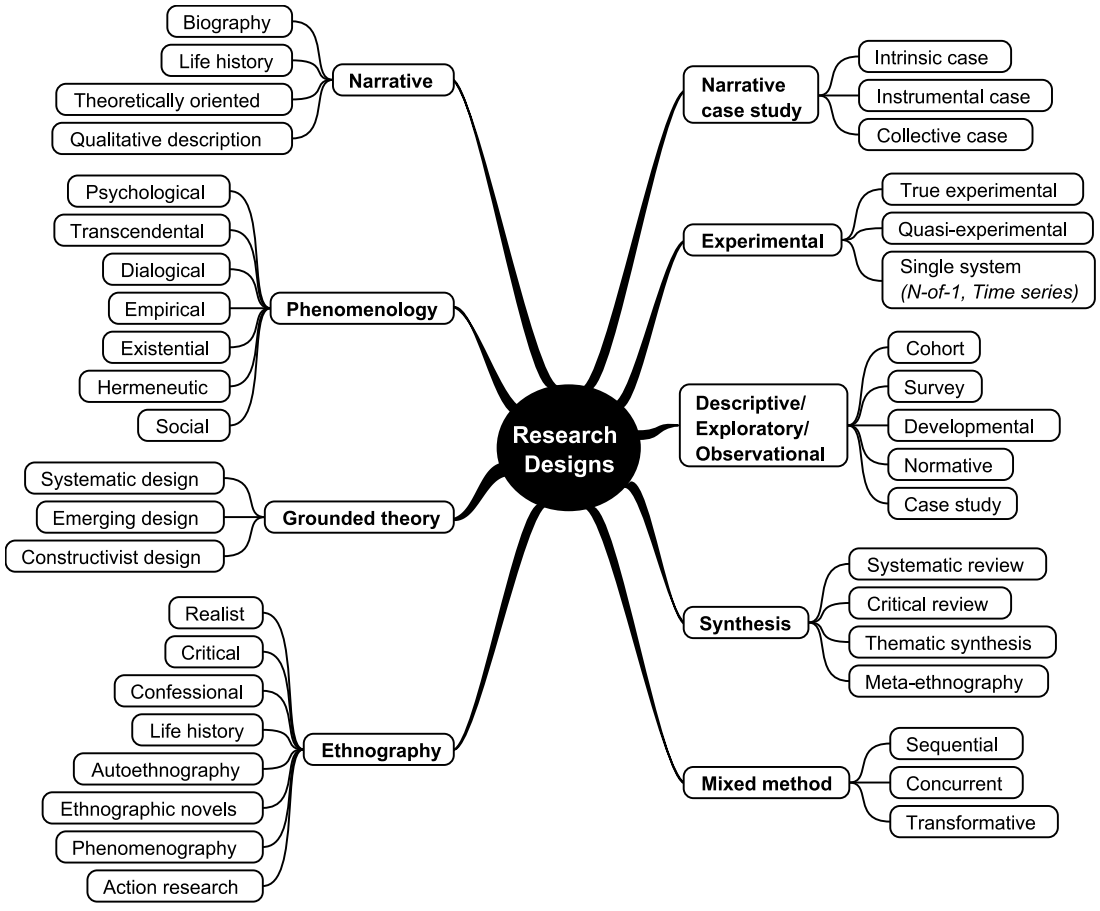

Figure 2.2 Research designs

When there is such an array of research designs to use, it seems strange to begin a research project by stating that design 'X' is the one to use before fully determining the research question. Furthermore, a research objective, purpose, or question is

normally stated in a way that is independent of the research method employed [15]. Therefore, a better strategy is to concentrate on the question being asked and use that to determine the best research design or designs to answer it [21, 22]. Choosing a research design first, then working on your question, is like choosing a car as your form of transportation and then deciding that you would like to go on an overseas trip. Surely the better strategy is to decide where you want to go and then decide which is the best way to get there?

This is not to say that a single person should be, or can be, an expert in all areas of research, or will not have a preference for certain types of research [2]. However, a researcher should know that different research designs exist and what each can achieve. Also, a research project may require more than one research design to satisfy the research question or questions. Such a realisation could lead to greater co-operation between researchers and more comprehensive results.

This common error is also present in the assumption, largely in the health research arena, that there is one hierarchy of evidence that will satisfy all research questions [23]. Normally the hierarchy is stated as meta-analysis at the top, followed by randomised controlled trails, and so on down to case reports (Table 2.1). The assumption is that methods further up the hierarchy are better and produce better results than those at the bottom. But is this correct?

Table 2.1 Hierarchy of evidence for quantitative studies

1. Systematic reviews and meta-analyses
2. Randomised controlled trials with definitive results
3. Randomised controlled trials with non-definitive results
4. Cohort studies
5. Case-control studies
6. Cross sectional surveys
7. Case reports

If the example is a quantitative-type question, such as, 'What evidence exists for ultra-sound to be used to speed up the recovery from soft tissue injuries?', then there is an argument that the hierarchy is appropriate. The researcher tries to find meta-analysis and randomised controlled trial papers to answer the question. What if a qualitative type question is asked such as, 'How do primary carers from different cultural backgrounds cope long term with family members who have an acquired brain injury?' In this case, the researcher's hierarchy may consist of generalisable studies and conceptual studies [24]. Thus, the hierarchy of evidence depends on the research question.

Researchers and funding bodies need to be more flexible in their understanding of the appropriateness of a research design for a particular research question, and whether that question is worth asking and investigating. After all, evidence based practice, and the call for medicine and other health professionals to base their practice on the best evidence available, includes the best diagnosis and treatment currently available as well as the "...thoughtful identification and compassionate use of individual patients' predicaments, rights, and preferences in making clinical decisions about their care" [25].

2.7 Conclusion

Although this chapter argues that qualitative and quantitative research are far more similar than they are different, that is not to say that there are no differences. All research designs have their strengths and weaknesses, and it is up to the researcher to be aware of those strengths and weaknesses.

One of the major reasons for the continued divide between qualitative and quantitative researchers is that qualitative and quantitative research are still taught as being fundamentally different [22]. However, since they are not fundamentally

different, it is time to teach the variety of research designs from which researchers can choose and to base that choice on the research question.

Finally, based on these arguments, a critical appraisal tool which incorporates both qualitative and quantitative research designs should be possible. The main objection, the belief that qualitative and quantitative research are fundamentally different, has been shown to have little foundation in modern research methodology.

2.8 IN SUMMARY

- Qualitative and quantitative research designs are more similar than different.
- The most suitable research design to use should be based on the research question.
- For critical appraisal, qualitative and quantitative research can be considered together in one tool.
- The next chapter explores research methods using mind maps (Objective 2).

2.9 REFERENCES

1. Hunt, S. D. (1991). Positivism and paradigm dominance in consumer research: Toward critical pluralism and rapprochement. *Journal of Consumer Research, 18*(1), 32-44.

2. Howe, K. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10-16.

3. Alasuutari, P., Bickman, L., & Brannen, J. (2008). Introduction: Social research in changing social conditions. In P. Alasuutari, L. Bickman & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 1-8). London: Sage.

4. Miller, A. (Fall 2008). Realism. In E. N. Zalta (Ed.), Stanford encyclopedia of philosophy. Stanford, CA: Stanford University. Retrieved from http://plato.stanford.edu/archives/fall2008/entries/realism/

5. Hammersley, M. (1992). Deconstructing the qualitative-quantitative divide. In J. Brannen (Ed.), *Mixing methods: Qualitative and quantitative research* (pp. 39-55). Aldershot: Avebury.

6. Hudson, L. A., & Ozanne, J. L. (1988). Alternative ways of seeking knowledge in consumer research. *Journal of Consumer Research, 14*(4), 508-521.

7. Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher, 21*(6), 13-17. doi:10.3102/0013189x021006013

8. Steinmetz, G. (1998). Critical realism and historical sociology. A review article. *Comparattive Studies in Society and History, 40*(1), 170-186. doi:10.1017/s0010417598980069

9. Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational inquirers. *Educational Researcher, 15*(1), 4-12.

10. Braithwaite, R. B. (1968). *Scientific explanation: A study of the function of theory, probability, and law in science*. Cambridge, MA: Cambridge University Press.

11. Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Pearson.

12. Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage.

13. Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

14. Brannen, J. (2004). Working qualitatively and quantitatively. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (pp. 312-326). London: Sage.

15. Onwuegbuzie, A. J., & Leech, N. L. (2005). Taking the "q" out of research: Teaching research methodology courses without the divide between quantitative and qualitative paradigms. *Quality & Quantity, 39*(3), 267-295. doi:10.1007/s11135-004-1670-0

16. Miller, D. C., & Salkind, N. J. (2002). *Handbook of research design and social measurement* (6th ed.). Thousand Oaks, CA: Sage.

17. Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

18. Trochim, W. M. (2006). The research methods knowledge base. Retrieved 29 January 2011, from http://www.socialresearchmethods.net/kb/

19. Howe, K., & Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher, 19*(4), 2-9.

20. World Medical Association. (2008). *Declaration of Helsinki: Ethical principles for medical research involving human subjects.* Paper presented at the 59th World Medical Association General Assembly, Seoul, South Korea. Retrieved from http://www.wma.net/e/policy/b3.htm

21. Luttrell, W. (2005). Crossing anxious borders: Teaching across the quantitative-qualitative 'divide'. *International Journal of Research & Method in Education, 28*(2), 183-195. doi:10.1080/01406720500256251

22. Silverman, D., & Marvasti, A. B. (2008). *Doing qualitative research: A comprehensive guide.* Los Angeles, CA: Sage.

23. Guyatt, G. H., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J., & Cook, R. J. (1995). Users' guides to the medical literature: IX. A method for grading

health care recommendations. *JAMA, 274*(22), 1800-1804. doi:10.1001/jama.1995.03530220066035

24. Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M., & Hughes, E. (2007). A hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology, 60*(1), 43-49. doi:10.1016/j.jclinepi.2006.03.014

25. Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ, 312*(7023), 71-72.

Chapter 3 – Research methods

The purpose of this chapter is to understand research methods before beginning the development of a critical appraisal tool (CAT). This was to make sure that the scope of research methods were well understood before beginning development of a CAT, thereby meeting Objective 2 of the study. Mind maps were used to outline the main features of research methods.

The chapter consists of an article accepted for publication on 3 March 2011 and available online 18 March 2011 (Appendix C.2):

> Crowe, M., & Sheppard, L. (in press). Mind mapping research methods. *Quality and Quantity*, (Online First). doi:10.1007/s11135-011-9463-8

Changes have been made to the published article to ensure thesis consistency. Copyright permission, which allows this paper to be reproduced, can be found in Appendix A.2.

Mind mapping research methods

3.1 ABSTRACT

The objective is to conceptualise research methods using mind maps. The major aspects, rather than a complete picture, of research methods are illustrated in seven distinct areas: research problem; research design; sampling techniques; ethical matters; data collection; data analysis; and report findings. Brief descriptions explain the mind maps and why items were placed in certain areas when they might have been traditionally placed elsewhere. The mind maps show that although decisions made in one area of research methods may affect decisions made in another, there is no pre-determined connection between each area and the research design chosen. The mind maps can be used as a guide to teach, supervise, and chart a way though the concepts of research methods, and may help to produce more robust research.

3.2 INTRODUCTION

It can be difficult to conceptualise the entire topic of research methods. Information on research methods is readily available in text books, journals and websites, but nothing could be found that brought this information into a coherent, easy-to-manage whole for teaching, research development, or to aid the design of a critical appraisal tool. Therefore, the objective of this paper was to use mind maps to visualise the complexities and extent of research methods.

Mind maps were chosen because they can represent ideas that are linked around a central theme and there are very few rules for creating them. It has been said that the main rule is simply to bring your brain and imagination [1]. The lack of rules makes creating mind maps an easy and natural method of organising and visualising

complex data, such as research methods, and the interactions among the data. Furthermore, mind maps can help people learn concepts better than traditional linear formats and note taking [2, 3].

The mind maps of research methods presented here were developed over two years, and have been through at least nine major and numerous minor revisions. The mind maps represent the author's understanding of research methods at this time and, even after two years, they are updated when new information is found or a deeper understanding of research methods is realised. However, changes do not happen as often now as when the mind maps were first being developed. Therefore, the mind maps in their current form should represent a reasonable and stable analysis of current research methods.

It should be noted, however, that not every aspect of research methods was included in the mind maps because this would make them very large and unwieldy, and reduce their effectiveness as a research and teaching tool. The aim was, and still is, to show the major parts of research methods so that the mind maps act as a visual guide to the topic rather than a comprehensive reference. This is a similar concept to the map of a city, where major places of interest are included but not every detail because this would make the map too difficult to read and use.

Mind maps are also intended to be self-contained accounts of the idea they represent. They are built in a free-form manner rather than as a stepwise process [1]. However, the decisions made in creating the mind maps need to be described briefly for these particular mind maps to be useful to others. The descriptions do not give a full account of everything contained within each branch of the mind maps because this information can be found in good text books. Instead, the descriptions state why specific items have been placed in their current location whereas traditionally they may be placed elsewhere, and were intended to draw attention to particular items.

Throughout this thesis, the phrase *research methods* is used in the plural and indicates the overall process of initiating, implementing, analysing, and reporting research. The phrase *research design* refers to the overarching approach or approaches used to answer the research question, such as true experimental or phenomenological designs. *Research methodology* refers to the philosophical and theoretical (or, if you prefer, the ontological and epistemological [4]) basis for research designs. Certain aspects of research methodology were explored in the previous chapter [5].

The mind maps can be shown as one large mind map, which is included as a fold out insert (the very last page of the thesis). A summary of the full mind map is shown in Figure 3.1. Each branch represents a distinct aspect of research methods. These seven branches are shown as individual mind maps in the following sections (Figures 3.2–3.8).

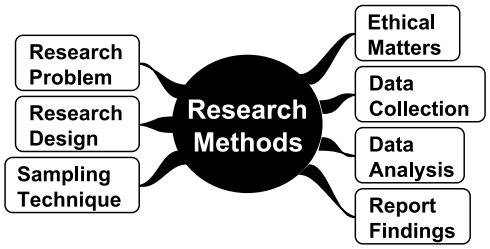

Figure 3.1 Research methods

Although the branches are independent of each other, decisions made in one branch may influence or dictate decisions made in other branches. Furthermore, it is recognised that research is not a linear process and that decisions are not necessarily made in the sequence outlined below. The sequence was chosen based on narrative rather than research considerations.

3.3 RESEARCH PROBLEM

The research problem (Figure 3.2) is one of the most straightforward mind maps. Each piece of research normally begins with the definition of the research problem. The research problem itself usually has two aspects: Purpose (why is the research question important?); and one or more objectives (for example, how will I know if I have satisfactorily addressed the research problem?) [6 (pp. 14-16), 7].

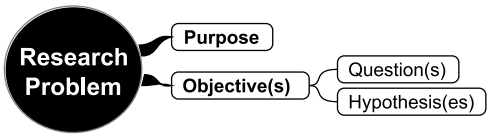

Figure 3.2 Research problem

3.4 RESEARCH DESIGN

There is no agreement within or between disciplines on what different research designs are called. Therefore, alternative names for research designs are given in parentheses and italics under the research design name the author most often uses (Figure 3.3). Experimental research designs are divided into the traditional true experimental and quasi-experimental designs [8]. However, a single system design branch was created for research designs where the intervention group acts as its own control. This contrasts with true experimental and quasi-experimental designs which have separate intervention and control groups. In other texts, single system designs can be scattered among quasi-experimental research designs, listed as pre-experimental, or some other combinations [8, 9 (pp. 55-82)]. Since this can be confusing for researchers, and especially for students, it appeared reasonable that a single system branch should be created which contained all single system designs, whether quasi-experiment, pre-experimental, or some other term.

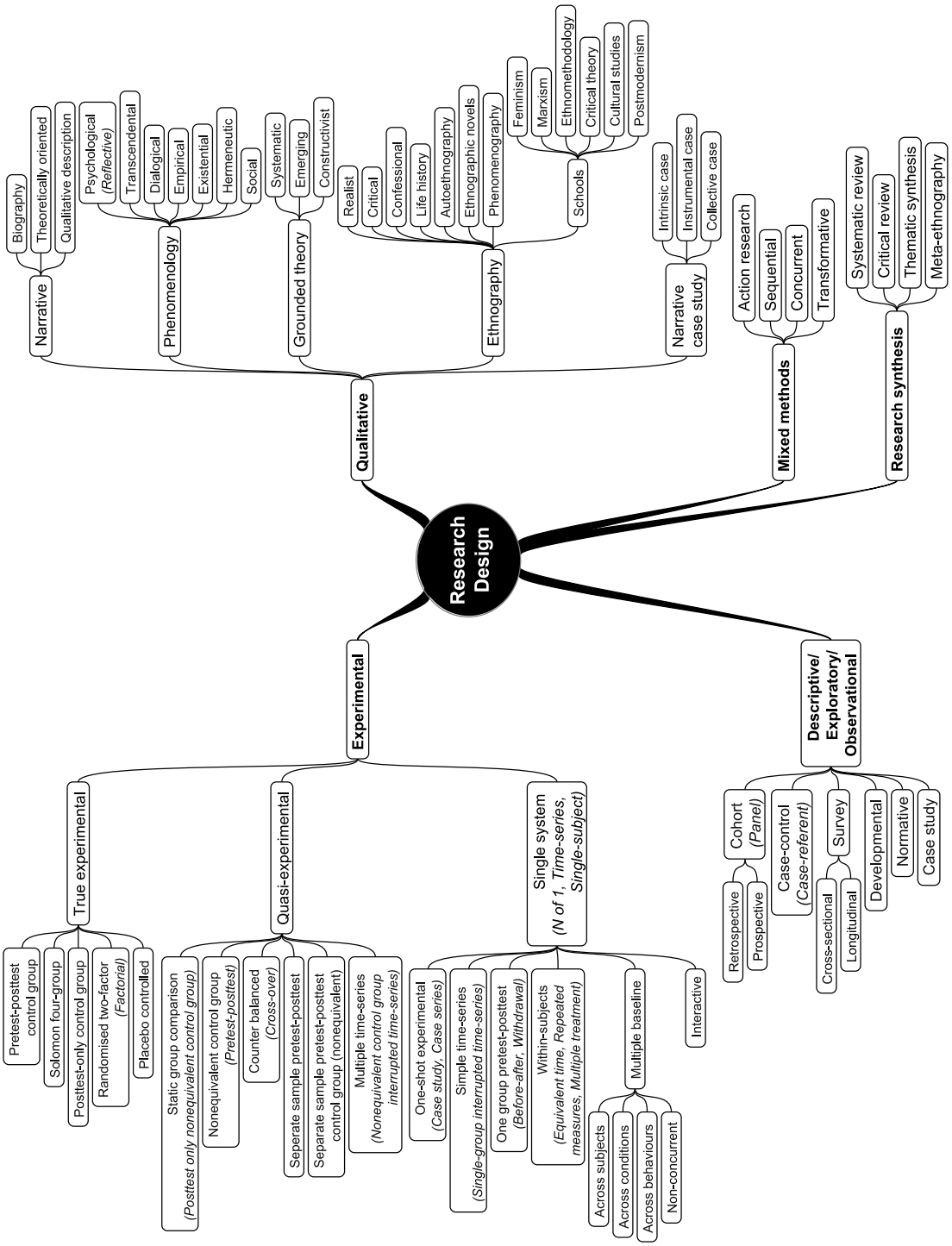

Figure 3.3 Research design

Descriptive, exploratory or observational (DEO) designs are quantitative research designs where there is no experimentation, intervention, or treatment [10]. This branch of research designs has an awkward name because, depending on the discipline (for example, health, sociology, business), it is called descriptive,

exploratory, observational, or some combination of these terms. However, shortening the title to DEO designs helps reduce the name's bulkiness.

Qualitative research designs are divided into five areas for simplicity, although other authors have argued there are as many as 16 [11 (pp. 81-135), 12]. Also, instead of putting the schools (also known as the orientations or ideologies) of qualitative research directly under qualitative design, they have been placed as a subset ethnographic research. This decision was made because a Marxist, feminist, or critical theory school or orientation, for example, can be seen as a social construct. Therefore, each has a cultural or ethnographic context rather than being an inherent characteristic of societies.

Mixed methods designs refer to using multiple qualitative, quantitative, or a mix of both approaches within the same research study [7, 12]. Mixed methods are in a separate branch so that their value is not overlooked. It should be noted that action research has been placed in mixed methods rather than in qualitative research. While it is true that action research can be purely qualitative, other authors have argued that a more comprehensive form of action research can be achieved by collecting qualitative and quantitative data [7].

The final branch in research designs is research synthesis, which has not traditionally been viewed as a research design. However, research synthesis has become an important technique for gathering secondary sources of data and pooling them to gain a better understanding of a topic [13 (pp. 4-7)]. Research synthesis can be a valuable and additive contribution to knowledge when completed in a systematic and thorough manner. Therefore, it should be considered a research design. Looking at the types of research synthesis, the main difference between a systematic review and a critical review is that a systematic review requires at least two reviewers whereas a critical review does not. They follow the same process in all

other aspects [13 (pp. 182-184)]. The two most widely reported qualitative research synthesis methods are thematic synthesis and meta-ethnography. However, other techniques have been developed [14, 15]. It is important to note that meta-analysis, which is often incorrectly used as a synonym for systematic review, is not listed here because it is a statistical technique that may be used in the data analysis part of a research synthesis.

3.5 SAMPLING TECHNIQUE

Sampling technique (Figure 3.4) can be treated as part of research design or sometimes data collection. However, good sampling technique is vital for good quality research, and sampling deserves to be seen as a unique and separate part of research methods [10 (pp. 143-158)].

Sampling technique is divided into three branches: method is the decision about whether to use probability or nonprobability sampling methods; size refers to calculating the sample size; and process refers to any procedures used in selecting or grouping individuals from the population of interest. It should be noted that this is the first of two situations in the mind maps where a decision is required from each of the branches rather than following a single branch to its conclusion. In other words, in defining a sample, information is required from the method, size and process branches.

When a research design is quantitative, probability does not need to be a characteristic of the sampling technique [16 (pp. 16-19)]. Similarly, when using qualitative research designs, the sampling technique does not have to be non-probability in nature. Most randomised controlled trials in health research use a non-probability convenience or purposive sampling method, even though the research would arguably be better if a random sample were used. In most cases,

however, random samples in health research are unethical, unfeasible, or uneconomical [10 (pp. 143-158)]. Furthermore, there is nothing within the theory or philosophy of qualitative research that says all qualitative research must use non-probability samples [17, 18]. If a qualitative study wanted to examine a specific issue across the population, for example, there is nothing to stop researchers from using a probabilistic, stratified sampling method along with a non-probability sample size method.

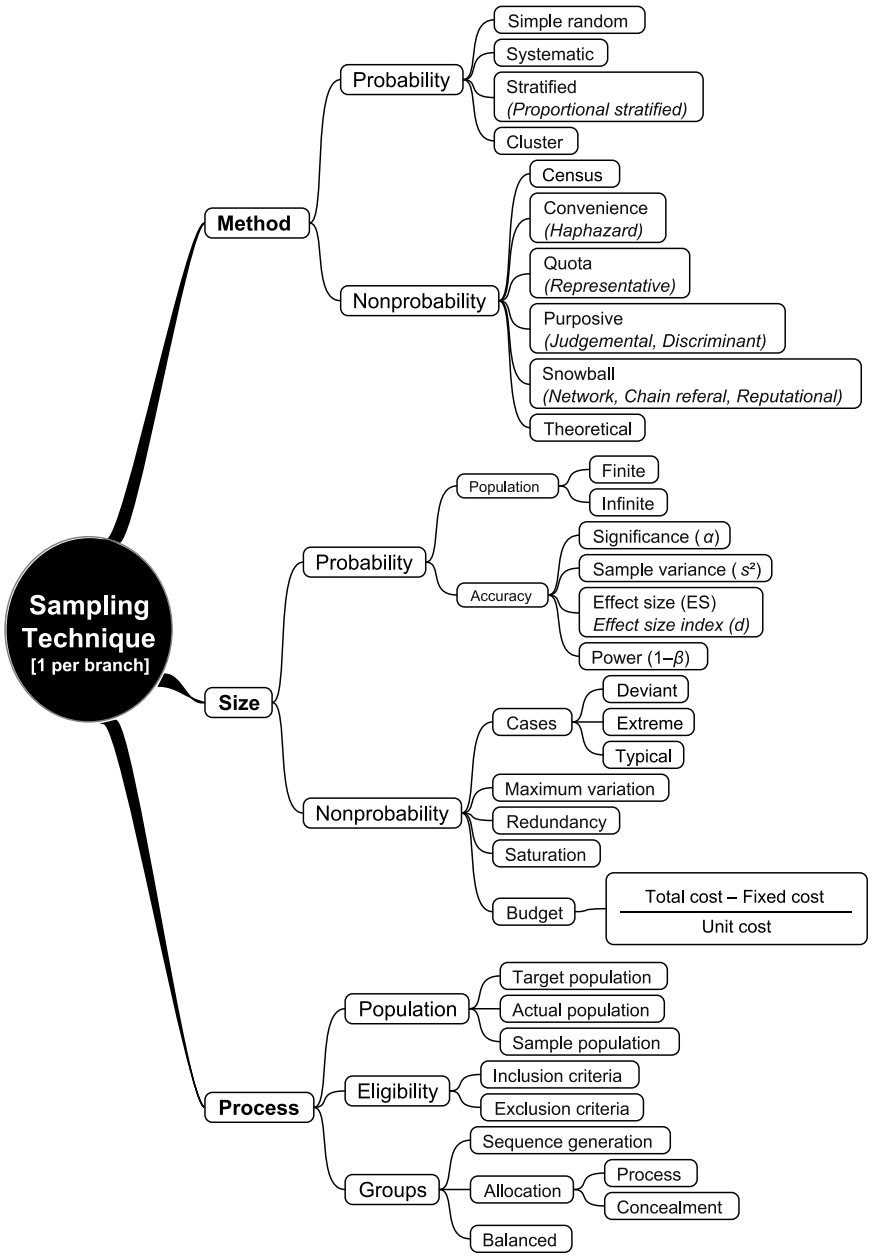

Figure 3.4 Sampling technique

It may also be noticed that researcher, participant, and other forms of blinding are not included in sampling. This is because they are considered data collection elements, and in particular, ensure the quality of the data collected. On the other hand, group allocation is seen as part of sampling because this decision affects sample size considerations [10 (pp. 170-171)].

Sampling also needs to be considered for research synthesis designs. Although a sample size is not required, how search strategies were determined (method), how the population of papers was defined (process), and the eligibility of papers (process) are all important sampling considerations for research synthesis [19 (pp. 1-108)].

3.6 ETHICAL MATTERS

Ethics is not just receiving ethical approval from an ethics board or committee. Ethical matters (Figure 3.5) should be incorporated throughout the research right from decisions on the research problem. Ethical matters encompass ethical behaviour towards participants (whether they are an eco-system, animal, or human) and ethical behaviour by the researcher. They are applicable for every research design [20].

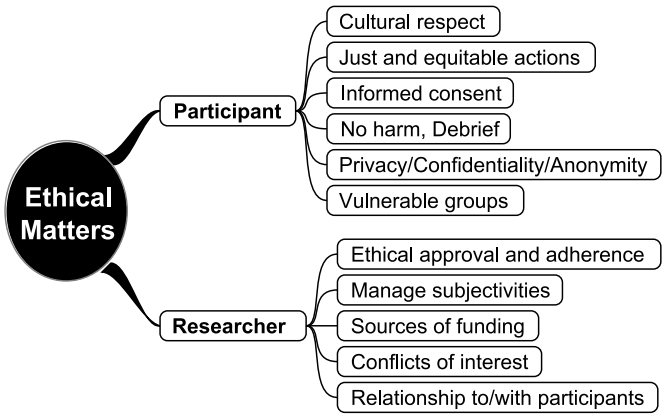

Figure 3.5 Ethical matters

In the example of research synthesis, a recent paper (see Chapter 6) showed that researchers may state that there are no ethical requirements [21]. However, when challenged about this, the researchers realised they had confused formal ethical approval for a study (which a systematic review does not normally need) with the wider concept of ethical behaviour by researchers (such as conflict of interest and funding disclosures). As a result, the researchers changed their minds and stated that ethical matters were a requirement for systematic reviews.

3.7 DATA COLLECTION

Data collection is an area of research methods that receives very little attention. This is bizarre given that data collection is often the most time consuming part of any research project and the quality of data collected is the basis on which conclusions are drawn. Poor data collection methods at best mean poor research and at worst the instigation of harmful actions. In Figure 3.6, data collection is divided into two branches: the method branch and the procedure branch.

The method branch describes the systems used to gather data [9 (pp. 99-139)]. The most common are: audit/review; observation; interview (which includes questionnaires); testing; or any combination of these. As with the sampling technique mind map, the method part of the data collection mind map requires the selection of one element from type, structure, and process for each data collection method used.

The procedure branch focuses on the processes used to gather data [12]. Under organisation, the researcher arranges and notes what is required to collect data, when, where, and by whom. Under participant/cases, the researcher decides on how to deal with problems gathering data from participants or cases, such as non-

participation, incomplete data, or when to stop an intervention if the results are either harmful or so good that all participants should be included.

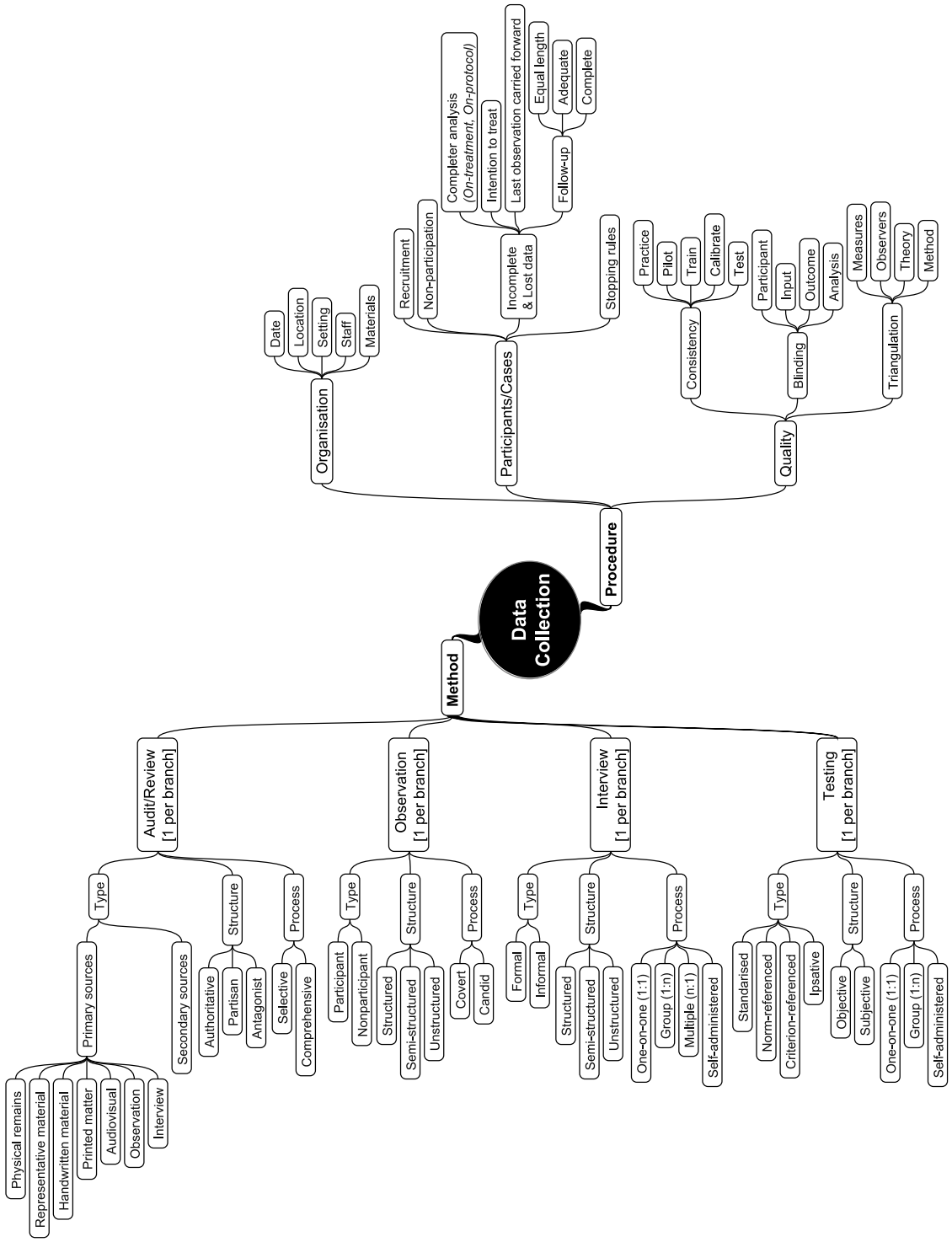

Figure 3.6 Data collection

Under quality, the researcher ensures that the data collected have consistency, which includes practicing data collection, using pilot tests, training research assistants, and calibrating instruments. Blinding is included here because it is more appropriately a data quality issue, where participants, researchers, and data analysts may be kept unaware of certain aspects of the research project. Lastly, there is triangulation, which is sometimes put together with data analysis techniques. However, triangulation starts by gathering data using multiple measures, observers, methods, or theoretical approaches [9 (pp. 312-312)]. After the for triangulation data are collected, they are analysed using appropriate data analysis methods.

3.8 DATA ANALYSIS

There are three branches within data analysis: essential; method; and interpret (Figure 3.7). The essential branch relates to undertaking a basic analysis of the data rather than, for example, jumping straight into inferential statistics. Examining data flows, raw data, and baseline data may help researchers to become aware of patterns within the data. It is also at this stage that incomplete and lost data should be analysed. This analysis is used to determine whether there are any reasons why data may be incomplete or lost, and how this may be analysed with respect to other data [7, 16].

The method branch concerns which method or methods of analysis are used. Since there are numerous materials available on both statistical and non-statistical methods of data analysis, the mind map does not go into detail. It should be noted, however, that meta-analysis is placed in the method branch under inferential statistics [10, 13, 16].

The interpret branch shows that all results are interpreted no matter which data analysis methods are used. Interpretation can be as simple as constructing a

narrative of what the results mean, based on the research objectives. More formal ways to interpret the data include examining data validity and reliability, converting data between qualitative and quantitative forms, and comparing data from different data collection methods, such as when using triangulation [7, 12].

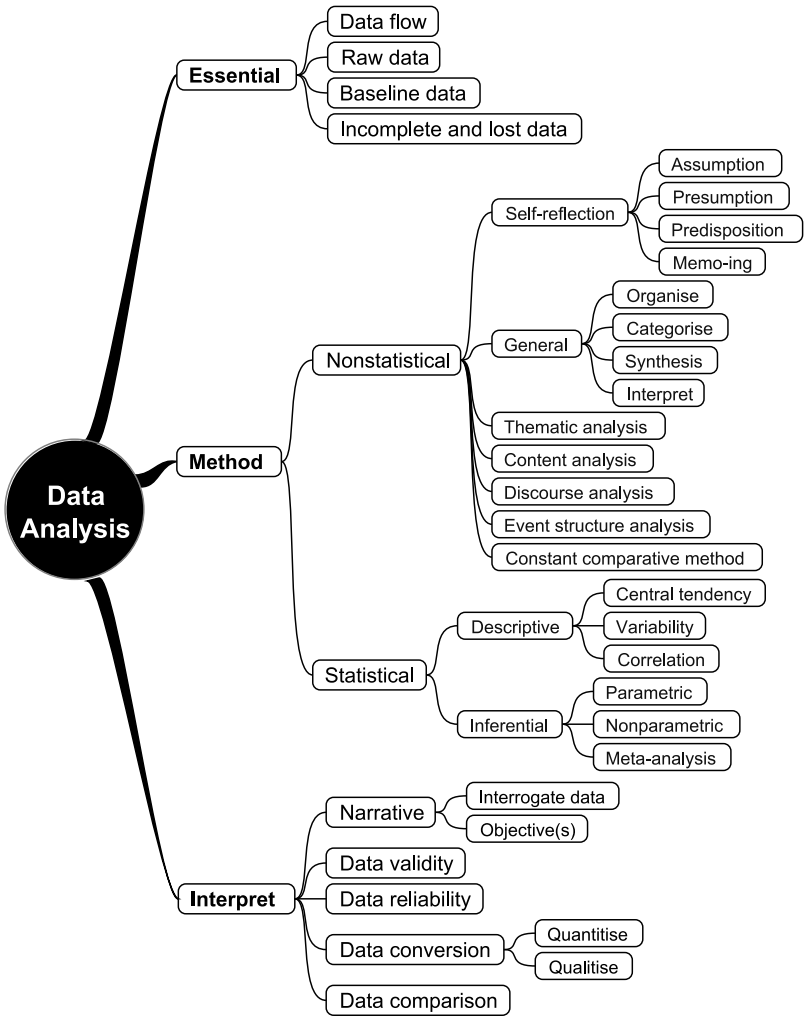

Figure 3.7 Data analysis

3.9 Report findings

There is little point in undertaking a study unless the research is reported. Reporting findings (Figure 3.8) has two branches: describe and contextualise. Describe reminds the researcher to report what they did during the research project.

Contextualise means that the researcher should place the results in context based on the findings in the current research, findings from previous research, and whether the results can be generalised [7].

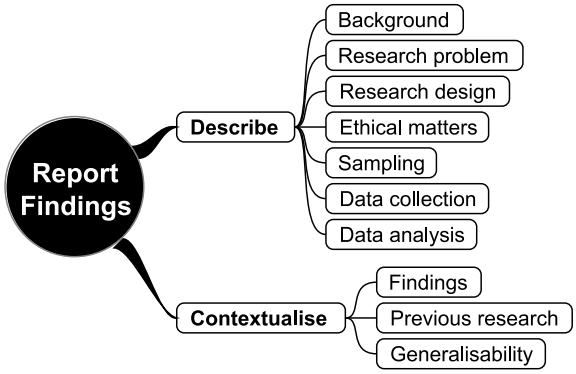

Figure 3.8 Report findings

3.10 Conclusion

The most important aspect of the research methods mind maps is that none of the seven main branches are connected directly to each other. Although decisions made in one branch may affect decisions made in another, there is no pre-determined connection between them, for example, defining the research problem does not automatically identify the research design to be used. Similarly, whether the underlying research design is qualitative or quantitative in nature, all the techniques shown in defining a research problem, creating a sample, exploring ethical matters, collecting data, analysing data, and reporting findings can be used.

Remember also that not every aspect of research methods is illustrated in these mind maps. Some may disagree with where certain items are placed or how they are described, but this is the nature of mind maps. These mind maps can be considered a work in progress and, as pointed out previously, they may change and develop as greater understanding of research methods is achieved.

The mind maps, then, can be seen as a good starting point for displaying the breadth and major components of research methods. This visual and non-traditional method of presentation can be used to learn, understand, and teach research methods. The mind maps show the variety of research methods available for the researcher, can be used by the research supervisor to guide students through the research process, and can be used by the research methods teacher to explain research methods from start to finish. These mind maps, when used as a guide through the research process, may assist researchers to produce more robust, higher quality research.

3.11 IN SUMMARY

- This exploration was undertaken to appreciate of the scope and complexity of research methods and provide context for the design of a critical appraisal tool.
- Research methods are illustrated in seven areas: research problem; research design; sampling techniques; ethical matters; data collection; data analysis; and report findings.
- There is no pre-determined connection between each area and the research design chosen.
- The next chapter reviews the design of critical appraisal tools and proposes an alternative tool structure (Objective 3).

3.12 REFERENCES

1. Buzan, T., & Abbott, S. (2005). *The ultimate book of mind maps: Unlock your creativity, boost your memory, change your life*. London: Thorsons.

2. Farrand, P., Hussain, F., & Hennessy, E. (2002). The efficacy of the 'mind map' study technique. *Medical Education, 36*(5), 426-431. doi:10.1046/j.1365-2923.2002.01205.x

3. Williams, C., Williams, S., & Appleton, K. (1997). Mind maps: An aid to effective formulation. *Behavioural and Cognitive Psychotherapy, 25*(3), 261-267. doi:10.1017/s1352465800018555

4. Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. St Leonards, NSW: Allen & Unwin.

5. Crowe, M., & Sheppard, L. (2010). Qualitative and quantitative research designs are more similar than different. *Internet Journal of Allied Health Sciences and Practice, 8*(4). Retrieved from http://ijahsp.nova.edu/

6. Miller, D. C., & Salkind, N. J. (2002). *Handbook of research design and social measurement* (6th ed.). Thousand Oaks, CA: Sage.

7. Onwuegbuzie, A. J., & Leech, N. L. (2005). Taking the "q" out of research: Teaching research methodology courses without the divide between quantitative and qualitative paradigms. *Quality & Quantity, 39*(3), 267-295. doi:10.1007/s11135-004-1670-0

8. Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

9. Polgar, S., & Thomas, S. A. (2007). *Introduction to research in the health sciences* (5th ed.). Edinburgh: Churchill Livingstone.

10. Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

11. Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage.

12. Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

13. Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

14. Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: A critical review. *BMC Medical Research Methodology, 9*(59). doi:10.1186/1471-2288-9-59

15. Dixon-Woods, M., Booth, A., & Sutton, A. J. (2007). Synthesizing qualitative research: A review of published reports. *Qualitative Research, 7*(3), 375-422. doi:10.1177/1468794107078517

16. Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

17. DePaulo, P. (2000). Sample size for qualitative research. *Quirk's Marketing Research Review,* (Article ID: 20001202). Retrieved from http://www.quirks.com/articles/a2000/20001202.aspx

18. Onwuegbuzie, A. J., & Leech, N. L. (2007). A call for qualitative power analyses. *Quality and Quantity, 41*(1), 105-121. doi:10.1007/s11135-005-1098-1

19. Centre for Reviews and Dissemination. (2009). *Systematic reviews: CRD's guidance for undertaking reviews in health care*. York: University of York.

20. National Health and Medical Research Council, Australian Research Council, & Universities Australia. (2007). *Australian code for the responsible conduct of research*. Canberra, ACT: NHMRC.

21. Crowe, M., Sheppard, L., & Campbell, A. (under review). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology*.

Chapter 4 – Review of critical appraisal tool design

This chapter investigates the design of critical appraisal tools (CAT) through a critical review of the literature. Information drawn from Chapters 2 and 3 was also incorporated. Based on the findings of this review, an outline of the proposed CAT was developed. This realised Objective 3 of the six research objectives.

The chapter consists of an article accepted for publication on 5 February 2010, available online 21 June 2010, and published in January 2011 (Appendix C.3):

> Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology*, *64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

Since the paper was published, one extra article was found:

> Stige, B., Malterud, K., & Midtgarden, T. (2009). Toward an agenda for evaluation of qualitative research. *Qualitative Health Research*, *19*(10), 1504-1516. doi:10.1177/1049732309348501

Therefore, the original article was updated to incorporate the additional data. Other changes have also been made to the published article to ensure thesis consistency. Copyright permission, which allows this paper to be reproduced, can be found in Appendix A.3.

A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed

4.1 ABSTRACT

Objective – To evaluate critical appraisal tools (CATs) that have been through a peer-reviewed development process with the aim of analysing well designed, documented, and researched CATs which could be used to develop a comprehensive CAT.

Study design and setting – A critical review of the development of CATs was undertaken.

Results – Of the 45 CATs reviewed, 26 (58%) were applicable to more than one research design, 11 (24%) to true experimental studies, and the remaining eight (18%) to individual research designs. Comprehensive explanations of how a CAT was developed and guidelines to use the CAT were available in five (11%) instances. There was no validation process reported in 12 CATs (27%) and 34 CATs (76%) had not been tested for reliability. The questions and statements that made up each CAT were coded into eight categories and 22 items such that each item was distinct from every other.

Conclusions – CATs are being developed while ignoring basic research techniques, the evidence available for design, and comprehensive validation and reliability testing. The basic structure for a comprehensive CAT is suggested, which requires further study to verify its overall usefulness. Meanwhile, users of CATs should be careful about which CAT they use and how they use it.

4.2 BACKGROUND

Critical appraisal, a key component of systematic reviews, is the thorough evaluation of research to identify the best papers on any given topic [1]. Many people perceive systematic reviews as synonymous with meta-analysis – the synthesis of experimental studies and the statistical analysis of data in these studies to establish the best treatment for a condition [2]. However, systematic reviews are not limited to gathering data from experimental studies. They can contain information from a wide variety of sources, including other types of quantitative research, qualitative research, and grey literature [3, 4]. Since systematic reviews can contain information from this variety of sources, it can be difficult to incorporate these data into a coherent whole using CATs [1]. This is because many of these CATs are useful only for a limited number of research designs and in many cases a particular CAT is designed for one specific research project [2]. In this way, scores cannot be directly compared when the research is appraised using two or more different CATs because the evidence uses two or more research designs.

Furthermore, the majority of CATs lack the depth necessary to comprehensively assess the research being analysed and few were designed to assess the quality of the research [2]. Most CATs cover basic items that are simple enough to quantify, such as internal validity, while many ignore vital information such as the suitability of the research design, which can be more difficult to interpret. Some authors go further and suggest that to properly appraise research, CATs need to focus on individual aspects of quality such as the completeness of the research report, adherence to ethical practices, and other empirically verified criteria [5].

Another issue with CATs is whether the appraisal of research should be based on what is reported in a journal or other publication rather than on what actually happened. It cannot be determined solely from what is reported whether

explanations are missing due to space pressure, a lack of understanding, or that the work was not done. The alternative suggested is that authors of systematic reviews should contact the original authors of research papers to clarify matters [1, 6, 7, 8]. However, the arguments about appraising a published paper based on what was reported or what really happened has an air of speciousness: after all, academics mark students' papers all the time based on what was written rather than what the student meant to write. It is up to the author to ensure that important information is not missing from a paper before it is published and not to transfer blame to the publisher afterwards.

The construction of items in CATs generally takes three different forms: open-ended questions; closed questions; and statements. There is greater support for closed questions over open-ended questions because they are easier to analyse, especially electronically [1, 9]. The disadvantage is that open-ended questions can help the appraiser interrogate the research more thoroughly, gaining a better understanding in the process, than they would with closed questions.

When appraising items, CATs use either summary scores (including scales and weighting schemes) or component scoring. In using a summary score, all items are added up to produce a single overall mark for a paper and then papers can be ranked based on this summary score. The alternative is to use a component measure where each component of the CAT is compared across all papers included in a review.

The problem with a summary score is that a single score can hide serious defects in a paper if it scores high in other areas [10, 11, 12, 13, 14]. This is also true where the summary score of a CAT is converted into a scale, for example where papers having a summary score in certain ranges are designated criteria such as 'good', 'fair', 'poor', or something similar. The added difficulty with scales over a summary score is that the definition of the scale tends to be arbitrary, with no objective reason why a

certain range of scores means that the research is in one category rather than another. Also, the summary score is further diluted and may increasingly hide defective studies [2, 11, 15].

A way to alleviate the problem of hiding defective studies, when using a summary score, is to implement weighting schemes. It is argued that some items in an appraisal are more important than others and a superior overall score can be reached by increasing the weightings towards the more important parts of the research. However, the weighting schemes proposed so far are based mostly on the opinion of the authors rather than on evidence from the literature. As research into weighting schemes progressed, it was discovered that within a CAT there was little or no difference in the ranking of papers with or without weightings in place [11, 13, 16, 17].

Finally, perhaps most importantly, is that although CATs are used to assess research validity and reliability, many of the tools themselves have not gone through a validation or reliability process [18, 19, 20]. If CATs are to be an effective tool in evaluating research, they should be subjected to the same standards as the research they are used to appraise.

There have been a number of reviews of CATs [2, 11, 15, 21, 22, 23, 24]. However, these studies have focused on tools used in research rather than on the development of the CAT in the first place. As a result, many of the CATs reviewed were one-off tools developed for a specific purpose or modifications of previous tools that display many of the problems outlined above.

The aim of this review was to investigate CATs where the paper predominantly focused on the development of the CAT. The premise was that by limiting the scope to CATs where the authors have specifically written about the design of the tool,

many of the design flaws would have been worked through. In this way, it was expected that this would be a review of well-designed, well-documented, and well-researched CATs, with the prospect of developing a comprehensive CAT based on the best tools in the field of critical appraisal.

4.3 METHODS

The research design was a critical review of the literature, where the author searched, categorised, and analysed the literature [1]. The sampling method used was a non-probability sample to saturation based on the following *a priori* inclusion criteria, exclusion criteria, and search strategy.

4.3.1 Inclusion criteria

1. The paper must have substantially focused on the development of one or more CATs because the main aim of the critical review was the development of CATs.
2. The CAT must have been developed for the appraisal of primary research designs or systematic reviews, and have a structure that was general in nature so that the CAT could be used for other research.
3. Where more than one tool was published within the same paper, they should be included as separate CATs.
4. The paper was published in a peer reviewed journal.
5. The paper was published from 1980 onwards. This time period was chosen because it was from the early 1980s that the topic of quality in research and critical appraisal came to the fore [25].
6. The paper was in English.

4.3.2 Exclusion criteria

1. As a result of inclusion criterion no. 2, the following types of tools were excluded:
    a. Diagnostic/prognostic studies
    b. Economic evaluations
    c. Clinical guidelines
    d. Metrics
    e. Quality assurance or service delivery.
2. Reporting guidelines. These focus primarily on how a paper should be written rather than how it should be appraised.
3. Papers that explain how to critically appraise a paper. CATs mentioned in these papers were investigated further to ascertain if the CATs were suitable for inclusion.
4. A CAT developed by another person or group. Where a paper explained a CAT by another person or group, the original CAT was investigated to determine if that tool could be included but the second order explanation was not included because this could lead to double counting.
5. Systematic reviews of CATs. To prevent double counting, systematic reviews of tools were excluded but individual tools that were part of a systematic review were investigated further.

4.3.3 Search strategy

The search terms used and databases searched are listed in Table 4.1. The search terms were used to search only the title and abstract of papers listed in each database. The body of papers was not included because a pilot search showed that too many false positive results were retrieved. The search was undertaken in October 2008 and, where possible, the search criteria used were saved in the databases so

that notifications of new papers relating to the search would be available via RSS or email. A full exploration of the search strategy is in section 4.9 (p. 78).

Table 4.1 Search terms and databases

Search terms	Databases
(critical appraisal OR critical assessment OR critical evaluation OR critical review OR quality appraisal OR quality assessment OR quality evaluation OR quality review OR research appraisal OR research assessment OR research evaluation OR research quality OR research review) AND (checklist OR scale OR tool) AND Language = English AND Year ≥ 1980	CSA Illumina EBSCOhost Gale InfoTrac Informit ISI Web of Knowledge JStore OvidSP ProQuest Scopus The Cochrane Library

A consequence of the search strategy was that a number of well known and often cited CATs were excluded from the review. Examples of these include the tools from McMaster University School of Rehabilitation Science [26], the Cochrane Collaboration [9], Crombie [27], and Cooper [28], as these tools were not found in the peer reviewed literature. This further highlights the situation where tools have not been through a peer-reviewed validation or reliability checking process.

Once the strategy was in place, a large number of papers (6,255) were identified. If a paper did not substantially address the development of a CAT in the initial reading of its title and abstract, it was removed immediately from the list of papers. The list was de-duplicated and the remaining papers were assessed thoroughly to determine eligibility. This included searching through references in accepted papers to ensure that potential papers missed in the initial search were not overlooked. In total, 45 papers met the criteria for inclusion (see Figure 4.1 for a flow-diagram).

4.3.4 Ethical matters

This research did not require ethics approval because there were no human or animal interventions. There were no conflicts of interest or funding sources to declare.

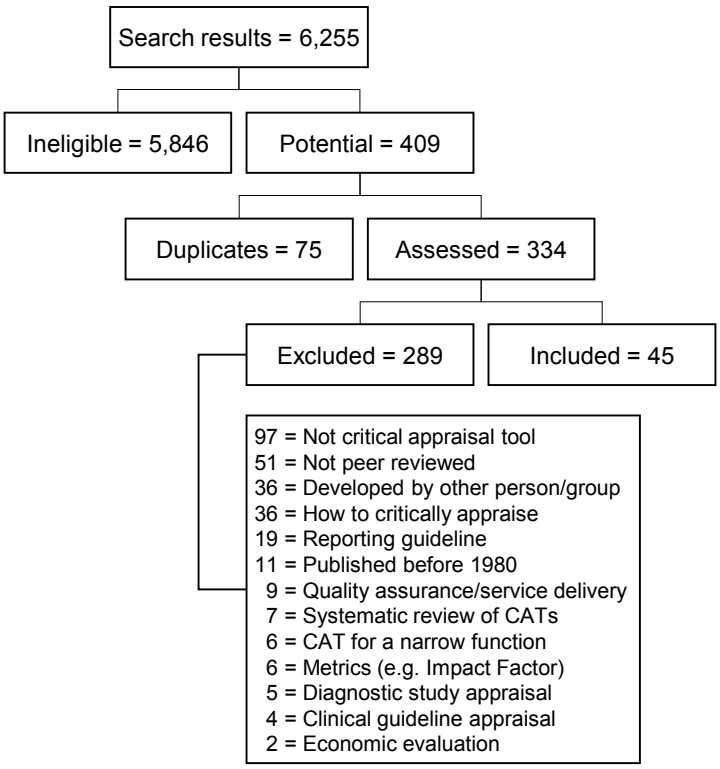

Figure 4.1 Flow-diagram of search results

4.4 RESULTS

On reading the 45 papers in the review, two different methods of analysis were undertaken. A descriptive quantitative analysis of the papers was completed first, which explored the structure, research methods, and analysis of the data used by the papers. Secondly, a qualitative analysis was completed, which explored the content of the questions or statements used within each CAT so that these questions or statements could be summarised and classified.

4.4.1 Quantitative analysis

The years in which CATs were published broke down into seven (16%) in the 1980s; 14 (31%) in the 1990s; and 24 (53%) in the 2000s. There were no criteria to restrict the search to the area of health research, however only one paper could possibly be categorised as being outside the health area and this paper was from an author based in a health sciences library [29].

The research design or designs in seven of the 45 CATs (16%) were not explicitly stated [10, 30, 31, 32, 33, 34, 35]. In these cases, the research design or designs were inferred from the contents of the paper. These papers are indicated with an asterisk (*) in Table 4.2.

Twenty-four of the CATs in the review (53%) were described as general in nature – they were self-described as being suitable to appraise a broad spectrum of research designs (Table 4.2). Six of these general CATs (13%) stated that they were applicable to all research designs [29, 36, 37, 38, 39, 40]; five (11%) stated that they were applicable to all quantitative research designs [10, 13, 16, 31, 32]; four (9%) stated that they were applicable to all experimental research designs [30, 34, 41, 42]; and nine (20%) stated that they were applicable to all qualitative designs [12, 14, 43, 44, 45, 46, 47, 48, 49]. Of these 24 general CATs, two (8%) were published in the 1980s, five (21%) in the 1990s, and 17 (71%) in the 2000s. Four of the six CATs that covered all research designs and all nine general qualitative CATs were published in the 2000s.

Table 4.2 Summary of critical appraisal tools

Source	Research design†	Explained	Guide	Validity§	Reliability
Glynn, 2006 [29]	All	Partial	Yes	C – –	No
Pluye et al., 2009 [39]	All (Mixed methods)	Yes	Yes	C – –	No
Hawker et al., 2002 [36]	All	Partial	Yes	C – –	No
MacAuley, 1994 [37]	All	No	Partial	C – –	No
Nielsen & Reilly, 1985 [38]	All	No	Partial	C – –	No
Rasmussen et al.,2000 [40]	All	Partial	No	– – –	No
Cho & Bero, 1994 [16]	Quantitative	Partial	No	C C –	Yes
Valentine & Cooper, 2008 [13]	Quantitative	Yes	Yes	C – –	Some
Heacock et al., 1997 [31]	*Quantitative	Partial	Yes	C – –	Some
Meijman & de Melker, 1995 [32]	*Quantitative	Partial	Partial	C – –	Some
Heller et al., 2008 [10]	*Quantitative	Partial	Partial	C – –	No
Moncrieff et al., 2001 [42]	Experimental	Partial	Partial	C C –	Yes
Downs & Black, 1998 [41]	Experimental	Partial	Yes	C C –	Yes
Duffy, 1985 [30]	*Experimental	No	No	C – –	No
Urschel, 2005 [34]	*Experimental	Partial	Partial	– – –	No
Reis et al., 2007 [47]	Qualitative	Partial	No	C – –	Yes
Walsh & Downe, 2006 [14]	Qualitative	Yes	Partial	C – –	No
Long & Godfrey, 2004 [46]	Qualitative	Yes	Partial	C – –	No
Cesario et al., 2002 [43]	Qualitative	Partial	No	C – –	No
Kuper et al., 2008 [12]	Qualitative	No	Partial	– – –	No
Côté & Turgeon, 2005 [44]	Qualitative	Yes	Yes	– – –	No
Dixon-Woods et al., 2004 [45]	Qualitative	Partial	No	– – –	No
Treloar et al., 2000 [48]	Qualitative	No	Partial	– – –	No
Stige et al., 2009 [49]	Qualitative	Yes	Partial	– – –	No
Genaidy et al., 2007 [50]	Epidemiology	Yes	Yes	C – –	Yes
DuRant, 1994 [51]	Epidemiology	No	No	– – –	No
Sindhu et al., 1997 [52]	True experimental	Yes	No	C C C	Yes
Jadad et al., 1996 [53]	True experimental	Yes	Yes	C – C	Yes
Maher et al., 2003 [20]	True experimental	Partial	Yes	C C –	Yes
Boutron et al., 2005 [54]	True experimental	Partial	Partial	C – –	No
Melynk & Fineout-Overholt, 2005 [55]	True experimental	Partial	Partial	C – –	No
Verhagen et al., 1998 [56]	True experimental	Yes	No	C – –	No
Reisch et al., 1989 [57]	True experimental	Partial	Yes	C – –	No
Evans & Pollock, 1985 [58]	True experimental	No	Yes	C – –	No
Chalmers et al., 1981 [59]	True experimental	No	Yes	C – –	No
de Vet et al., 1997 [17]	True experimental	Partial	Partial	– – –	No
Vickers, 1995 [35]	*True experimental	Partial	Yes	– – –	No
Rangel et al., 2003 [33]	*Cohort	Partial	Partial	C – –	Yes
Lichtenstein et al., 1987 [60]	Cohort	Partial	No	C – –	No
Shea et al., 2007 [61]	Systematic review	Partial	Partial	C – –	No
Oxman & Guyatt, 1988 [62]	Systematic review	No	Partial	C – –	No
Hunt & McKibbon, 1997 [63]	Systematic review	Partial	Partial	– – –	No
Wilson & Henry, 1992 [64]	Systematic review	Partial	Yes	– – –	No
Tate et al., 2008 [65]	Single system	Yes	Partial	C C –	Yes
Loney et al., 1998 [66]	Survey	Partial	Yes	C – –	No

* Research design inferred.
† Quantitative = Experimental, Descriptive/exploratory/observational (DEO);
Experimental = True experimental, Quasi-experimental, Single system;
DEO = Cohort, Survey, Other;
Epidemiology = True-experimental, Quasi-experimental, Cohort, Survey.
§ Validity key: C – – (Content validity); – C – (Concurrent validity); – – C (Construct validity).

The most numerous of the remaining 19 CATs that were specific in nature were for true experimental studies, with 11 (24%) [17, 20, 35, 52, 53, 54, 55, 56, 57, 58, 59]. The next most common CATs were for systematic reviews, with four tools (9%) [61, 62, 63, 64]. Two CATs, which were not included in the general research designs category, stated that they were specific to epidemiological research designs (self-described as being suitable for true experimental, quasi-experimental, cohort, and survey designs) [50, 51]. Finally, two CATs were described as relevant to cohort studies [33, 60], and one each for single subject designs [65] and surveys [66].

One CAT was designed specifically for mixed methods research [39]. However, it was also stated that the CAT could be used to appraise all research designs. Therefore, it was included in that category.

A reported expert or group of experts was responsible for the design of a CAT in 38 papers (84%). Three papers (7%) used the Delphi method or some modification of it. Three papers (7%) stated that they used evidence from the literature to develop their CAT and one paper used the Nominal Group Technique. Analysis of the methods used to develop CATs did not show any pattern based on the year the tool was published or any other factors.

The number of items in each CAT ranged from 1–10 items in 10 papers (22%), 11–20 in 11 papers (24%), and after that as the total number of items increased there was a general reduction in the number of tools. In the 1980s and 1990s, CATs were more likely to have 1–20 items (nine instances or 43% of papers published in that period), whereas in the 2000s CATs were more likely to have 11–30 items (14 instances or 58% of papers published in that period).

Items in CATs were answered through closed questions in 24 instances (53%), open-ended questions in 13 instances (29%), a combination of open and closed questions

in three instances (7%), and through statements in five instances (11%). The scoring system used was predominantly a summary score, 23 CATs (51%), or a component scoring system, 21 CATs (47%). One CAT used a combination of component and summary scores [57]. Four of the CATs that used a summary score converted that score into a scale, five used a weighting scheme, and one CAT used a combination of a scale and weighting scheme.

The time taken to complete an appraisal was mentioned in five CATs (11%). This ranged from 10–30 minutes. There were no trends apparent in answering, scoring, or time taken to complete appraisals.

Comprehensive explanations of how a CAT was developed and user guides to use a CAT were available in five instances (11%). Partial explanation or user guides were available in 23 CATs (51%), while 17 CATs (38%) had no explanation or user guide. Explanations for how CATs were developed improved over time, with 29% of CATs published having a comprehensive or partial explanation in the 1980s, 86% published in the 1990s, and 92% published in the 2000s. There was no apparent trend for CAT user guides, with an average of 36% of CATs having a comprehensive and 42% having a partial user guide across the three decades.

Considering validation of the CATs, two (4%) had undertaken content, concurrent and construct validation or, in the case of one CAT [53], had explained why it had not undergone concurrent validation. A further five CATs (11%) had undergone content and concurrent validation, 26 CATs (58%) had completed content validation, while for 12 CATs (27%) there was no mention of validation. In relation to reliability, 10 CATs (22%) had been reliability tested, three (7%) had undergone some reliability testing, and 32 (71%) had not been reliability tested.

Looking at validation and reliability with regard to the research designs covered by the CAT, in the group described as covering all research designs five out of six CATs (83%) had undertaken content validation only and none (0%) had tested for concurrent or construct validity, or reliability. In the general qualitative CATs, four out of nine CATs (44%) were tested for content validation only and one (11%) had been tested for reliability. In the area of CATs for systematic reviews, two out of four CATs (50%) were tested for content validation only, and none (0%) for concurrent or construct validity, or for reliability.

In the CATs developed for true experimental designs, three of the 11 CATs (27%) had undertaken reliability testing. One of these three had content, concurrent, and construct validation, the second had content and construct validation, while the third had content and concurrent validation. Six of the true experimental design CATs (55%) had undertaken content validation only but no reliability testing, and two of the 11 CATs (18%) had not been tested for validation or reliability.

In respect to limitations of the CATs developed, only 18 papers (40%) mentioned that there were limitations to the tool developed.

4.4.2 Qualitative analysis

The constant comparative method was used to qualitatively analyse items from the CATs so that a summary of items included in CATs could be derived from the evidence [67, 68]. In the context of this review, the constant comparative method involved: comparing and contrasting items within the CATs; establishing and refining categories from the items; setting boundaries on categories and items; finding and integrating evidence for and against categories and items; and summarising items from the CATs into the final categories and items. NVivo version

8 (QSR International, Doncaster, VIC, Australia), was used to assist with coding and categorising the items.

The first attempt at developing categories and items purely from the CATs themselves proved too difficult to complete. The CATs used such an array of structures, combinations, sort orders, and idiosyncrasies that it was impossible to cross compare them. Therefore, a base structure was sought to assist in the process. Two different methods of categorisation and itemisation were used within the CATs themselves: The first was research validity and the second was reporting of research.

The first method explored was to base the categories and items on research validity. This method was rejected for three reasons. First, critical appraisal and quality of research was often limited to internal (bias, confounds, interactions, effect modifiers, imprecision) and external (generalisation outside the study sample) research validity. Conclusion (the relationship between data and inferences made) and construct research validity (the relationship between what was researched and how it was operationalised) were seen as too difficult to appraise easily, as stated earlier [35, 41, 54]. Second, more issues were considered in the CATs than validity of the research (for example clear objectives/hypotheses or reasons certain research decisions were taken), which could not be readily incorporated into the research validity structure. Third, using research validity criteria seemed counter-intuitive when appraising research in comparison to criteria based on how research was reported. This is further backed by the layout of the CATs in this review, where only three (7%) were structured on a research validity basis.

Reporting guidelines already published were employed as the basis for the development of categories and items used in this review. The six reporting guidelines chosen were: CONSORT (Consolidated Standards for Reporting of Trials; true experimental studies) [69]; STROBE (Strengthening the Reporting of

Observational Studies in Epidemiology; cohort, case control, and cross-sectional studies) [70]; PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses, formerly QUOROM; meta-analysis of true experimental studies) [71]; MOOSE (Meta-analysis of Observational Studies in Epidemiology; meta-analysis of observational studies) [72]; COREQ (Consolidated Criteria for Reporting Qualitative Research; qualitative studies) [73]; and SQUIRE (Standards for Quality Improvement Reporting Excellence; formal, planned studies) [74]. These reporting guidelines represent a wide range of research designs and are not based on other guidelines [75]. This ensured less chance of bias towards any particular reporting method. In addition to the reporting guidelines, the basic structure of research methods (research design, sampling techniques, data collection, and data analysis) was also used to aid the creation of categories and items (see Chapter 3).

Seven iterations were required from first draft to generation of the final version of categories and items (Table 4.3). This involved combining items, re-coding data, and developing item descriptors so that each item was distinct from every other based on the evidence found within the CATs. In total, eight categories and 22 items were established, with each item comprised of a number of points to further describe it. Each category is described below.

The *Preamble* category had the least evidence for inclusion in a CAT. The requirement for *Title* and *Abstract* items mostly came from papers that were closer to reporting tools than critical appraisal tools [30, 37, 58]. The *Text* item had some support because it pertains to papers being clear and concise, with sufficient detail to enable other researchers to reproduce the research.

Table 4.3 Categories and items included in CATs

Category Item	**Item descriptor**	**Papers** (%)	**Source**
Preamble			
Text	1. Sufficient detail others could reproduce 2. Clear, concise writing/table(s)/diagram(s)/figure(s)	12 (27%)	[10, 12, 30, 37, 40, 43, 47, 49, 52, 57, 58, 59]
Title	1. Includes study design and aims	5 (11%)	[30, 36, 46, 57, 58]
Abstract	1. Key information 2. Balanced and informative	3 (7%)	[30, 36, 58]
Introduction			
Background	1. Summary of current knowledge 2. Specific problem addressed and reason(s) for addressing	13 (30%)	[14, 30, 31, 33, 34, 36, 38, 40, 43, 44, 46, 51, 58]
Objective	1. Primary objective(s), hypothesis(es), aim(s) 2. Secondary question(s)	27 (60%)	[10, 14, 16, 30, 31, 32, 34, 36, 38, 39, 40, 41, 42, 43, 44, 47, 48, 49, 50, 51, 52, 57, 60, 61, 62, 63, 64]
Research design			
Design type	1. Research design(s) chosen and why 2. Suitability of research design(s)	23 (52%)	[10, 14, 16, 29, 30, 31, 32, 36, 37, 38, 39, 40, 43, 44, 45, 46, 48, 50, 51, 52, 58, 59, 65]
Intervention, Input, Exposure	1. Precise details of the intervention(s)/input(s)/exposure(s) for each group 2. Main factors that contribute to choice of intervention(s)/input(s)/ exposure(s) 3. Intervention(s)/Input(s)/Exposure(s) valid and reliable	21 (48%)	[10, 13, 17, 33, 35, 36, 38, 39, 41, 42, 50, 51, 52, 54, 57, 58, 59, 60, 61, 62, 64]
Outcome, Output, Predictor	1. Clearly define outcome(s)/output(s)/predictor(s) 2. Main factors that contribute to choice of outcome(s)/output(s)/predictor(s) 3. Outcome(s)/Output(s)/Predictor(s) valid and reliable	23 (52%)	[10, 13, 16, 17, 29, 30, 31, 33, 34, 35, 38, 39, 41, 42, 46, 50, 51, 52, 55, 57, 58, 65, 66]
Bias and other	1. Potential sources of bias, confounding, interactions, effect modifiers, imprecision 2. Sequence generation, group allocation, group balance, and by whom 3. Equivalent treatment of participants/cases/groups	34 (76%)	[13, 16, 17, 20, 29, 31, 32, 33, 35, 36, 39, 40, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 59, 60, 61, 62, 64, 65]
Sampling			
Sampling method	1. Method(s) of selecting participants/cases/groups 2. Suitability of sampling method	15 (34%)	[10, 12, 13, 14, 20, 30, 36, 40, 43, 45, 46, 48, 51, 60, 66]
Sample size	1. Calculate sample size (statistical, theoretical, practical) 2. Suitability of sample size	21 (48%)	[13, 16, 17, 29, 30, 31, 32, 34, 35, 36, 40, 41, 42, 48, 50, 51, 52, 57, 58, 59, 66]
Sampling protocol	1. Description and suitability of target/actual/sample population 2. Inclusion and exclusion criteria for participants/ cases/groups 3. Recruitment of participants/cases/groups	36 (80%)	[10, 13, 14, 16, 17, 20, 29, 30, 31, 32, 33, 35, 36, 38, 39, 40, 41, 42, 44, 46, 47, 48, 49, 50, 51, 52, 53, 57, 58, 59, 60, 61, 62, 63, 64, 66]

Table 4.3 (continued)

Category Item	**Item descriptor**	**Papers** (%)	**Source**
Ethical matters			
Participant	1. Informed consent, equity 2. Privacy, confidentiality/anonymity	11 (24%)	[14, 29, 30, 40, 43, 46, 48, 49, 52, 57, 58]
Researcher	1. Ethical approval, funding, conflict(s) of interest 2. Subjectivities, relationship(s) with participants/ cases	14 (31%)	[10, 12, 14, 16, 29, 36, 42, 48, 49, 51, 57, 59, 61, 66]
Data collection			
Collection method	1. Method(s) used to collect the data 2. Suitability of collection method(s)	12 (27%)	[14, 29, 30, 40, 43, 44, 45, 46, 47, 57, 60, 65]
Collection protocol	1. Include date(s), location(s), setting(s), personnel, materials, processes 2. Method(s) to ensure/enhance quality of measurement/instrumentation 3. Manage non-participation, withdrawal, incomplete/ lost data	27 (60%)	[12, 13, 14, 29, 30, 32, 33, 35, 36, 38, 39, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 52, 54, 57, 59, 61, 64]
Results			
Analysis, Integration, Interpretation method	1. Method(s) used to analyse/integrate/interpret primary outcome(s)/output(s)/ predictor(s) 2. Methods for additional analysis/integration/ interpretation (e.g. subgroup analysis) 3. Suitability of analysis/integration/interpretation method(s)	27 (60%)	[10, 12, 14, 16, 29, 30, 31, 32, 34, 36, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 59, 60, 63]
Essential analysis	1. Flow of participants/cases/groups through each stage of research 2. Demographic and other characteristics of participants/cases/groups 3. Analyse raw data, response rate, non-participation, withdrawal, incomplete/lost data	32 (73%)	[10, 13, 14, 16, 17, 20, 29, 32, 33, 34, 35, 39, 41, 42, 43, 46, 48, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 64, 65, 66]
Outcome, output, predictor analysis	1. For each outcome/output/predictor, a summary of results and precision 2. Consider benefits, harms, unexpected results, problems, failures 3. Describe outlying data (e.g. diverse cases, adverse effects, minor themes)	31 (70%)	[10, 13, 14, 16, 17, 20, 29, 30, 33, 34, 35, 41, 42, 43, 44, 47, 48, 50, 51, 52, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66]
Discussion			
Interpret	1. Interpret results in the context of current evidence and objectives 2. Draw inferences consistent with the strength of the data 3. Consider alternative explanations for observed results 4. Account for bias, confounding, interactions, effect modifiers, imprecision	33 (73%)	[10, 14, 16, 29, 30, 31, 32, 33, 34, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 55, 57, 58, 59, 60, 62, 63, 64]
Generalise	1. Consider overall practical usefulness of the study 2. Discuss the generalisability (external validity) of the study results	25 (56%)	[10, 12, 14, 16, 29, 30, 31, 34, 35, 36, 37, 40, 43, 45, 46, 47, 48, 49, 50, 51, 52, 55, 63, 65, 66]
Concluding remarks	1. Highlight study's particular strengths 2. Suggest steps that may improve future results (i.e. limitations) 3. Suggest further studies	11 (25%)	[14, 29, 30, 31, 33, 36, 38, 40, 44, 46, 51]

In the *Introduction* category, the *Background* item also had low support for inclusion in a CAT compared to many other items. The need for an explicit statement of objectives, hypotheses, aims, and any secondary questions being explored, as shown in the *Objective* item, was seen to be important. This was true across all different research designs, except for CATs that dealt specifically with true experimental designs.

In the *Research design* category, all items were seen as relatively important across the CATs. The exception was for general qualitative design CATs, which were not strong with regard to the *Intervention, Input, Exposure* item. The *Design type* item refers to which research design was used, such as: true experimental; single system; descriptive, exploratory or observational; qualitative; or systematic review.

In the *Sampling* category, the importance of the *Sampling protocol* item for CATs was highest across all items described here. The *Sampling protocol* is defined here as the description and suitability of the sample, any inclusion or exclusion criteria used, and how recruitment was undertaken. The *Sample size* item also had some support among all the CATs except for systematic review CATs and general qualitative designs CATs. The need for a *Sampling method* item (which type of probability or non-probability sampling method was used) rated quite low among all research designs.

The *Ethical matters* category had the least overall support from the CATs, apart from the *Preamble* category. However, the general qualitative research CATs rated *Ethical matters* highly because ethics in research is not just about gaining approval from an Ethics Board. It also involves issues such as power relationships and reflexivity [12, 14, 46, 48]. There was slightly more support for the inclusion of the *Researcher* item (31%) than the *Participant* item (24%).

Under the *Data collection* category, there was good support for the inclusion of a *Collection protocol* item (protocols to administer the collection, quality control, and management of data). There was low support from the CATs for inclusion of the *Collection method* item (for example, whether the data were collected through an audit, observation, interview, or testing) except from the general qualitative research CATs.

There was very good agreement from the CATs for the inclusion of all the individual items in the *Results* category. This was perhaps not particularly surprising given that the description and analysis of study results traditionally receives most attention from systematic reviews and critical appraisal.

Finally, in the *Discussion* category, there was very good support from the CATs for the *Interpretation* item, perhaps not surprising given the requirement for the interpretation of results within studies. The *Generalisation* item had good support from the CATs and the *Concluding remarks* item had one of the lowest indications of support from the CATs.

4.5 DISCUSSION

The range of research designs covered by the papers in this critical review give some hope that the perception systematic reviews are limited to true experimental studies is being overcome. It is also interesting to note the number of CATs for appraising qualitative research designs, all of which were developed in the 2000s. A great deal of discussion has occurred about whether qualitative studies can, or should, be aggregated. However, given the importance of systematic reviews from an academic and policy viewpoint, it would appear that the inclusion of qualitative studies in systematic reviews is more likely to happen than not.

Leaving aside the qualitative CATs, almost half the remaining tools were described as being useful for all research designs, all quantitative designs, or all experimental designs. Furthermore, half of these general CATs were developed in the 2000s. It appears that the need for tools that can incorporate data from across research designs into a single systematic review has been taken seriously.

However, the trend to develop CATs for multiple purposes is strongly tempered by the concern that although CATs are used to appraise the evidence in the literature, they continue to be developed while ignoring the evidence for CAT development and basic research techniques. The first example of this is in the use of weighting scales. Although the use of weighting scales was shown to be unnecessary and subjective in the mid-1990s, three CATs were still developed using this method [17, 33, 52]. Even more worrying was that validation and reliability checks had not been undertaken in 12 CATs (27%), and, as for going beyond the basics, only six CATs (13%) tested concurrent validation and a miserly two CATs (4%) attempted construct validation, even though the concept of needing to appraise research for construct validation was mentioned in six CATs. Furthermore, 32 CATs (71%) did not test for reliability. This lack of use of the literature, validation, and reliability means that many CATs cannot be used with confidence.

Looking to the future, towards creating a CAT based on the evidence and good research practice, a list of categories and items have been developed from the evidence available. However, a number of things must be noted about these categories and items.

First, the items are based on the evidence collected in this review and, as such, the veracity of the items is yet to be tested. A number of items appear to have more to do with reporting than critical appraisal, such as *Text*, *Title*, *Background*, and *Concluding remark* items, and may not necessarily be required to critically appraise

a paper. It could also be argued that the *Text* item should be kept in mind when appraising any individual item because it promotes the inclusion of detail, and a clear and concise style.

This list of categories and items was developed based on reporting guidelines and critical appraisal criteria, even though some authors believe that the areas of critical appraisal and reporting are different and should not be mixed [69, 71]. However, an argument can be made that any list can be used in a number of ways depending on the context. Take for example a 'To do' list. It acts as a reminder of what needs to be achieved (equivalent to writing), what has been achieved (equivalent to reporting), and comparing the two (equivalent to appraising). Perhaps a reporting tool or critical appraisal tool could be used in different ways depending on the context? This argument remains to be tested with future investigation of the categories and items developed here.

Second, the items are not intended as a logical or stepwise sequence of research events. They are a guide to the areas of research that could be included in an academic research paper. In fact, not all items are applicable to all research designs. Surveys, for example, are not known for the introduction of an intervention and, except for action research, neither are qualitative techniques. Also, some items follow as a consequence of actively engaging in the research rather than being determined beforehand. In systematic reviews it is difficult, if not impossible, to determine a sample size before the research begins but the sampling protocol used should still be defended (for example, inclusion and exclusion criteria).

Finally, *Ethical matters* items have a lack of support, primarily due to two issues [53, 54, 56, 57]. First, some CATs have excluded ethical matters because older, important studies did not have requirements for the declaration of participant and research ethical standards. Therefore, imposing current ethical standards on these

studies could adversely affect their inclusion in future systematic reviews. Second, some CATs see ethical matters as a reporting requirement and not as an important component in assessing the quality of the research itself. These issues should not mean that they are summarily excluded from critical appraisal. However, their requirement needs further investigation.

The major disadvantage of the categories and items to be included in a future CAT based on the CATs reviewed here is that the CATs described have had limited validation or reliability testing. However, this review at least represents a start based on the evidence available, no matter what that evidence may be. Only further studies can show if this method holds up under investigation.

4.6 Conclusion

Perhaps the ultimate irony of critical appraisal is that as part of systemic reviews (arguably the pinnacle of scientific evidence), the tools used are based on each appraiser's concept of research quality. This dependence on a subjective measure may mean that a CAT cannot be developed which takes as its starting point a rational view of the research process described here. However, given that this rational view is what exists, anyone appraising research should ensure that:

1. The context of the review is taken into consideration before choosing a CAT.
2. The CAT chosen was developed using the best evidence available.
3. The scores obtained from using the CAT should have undergone and continue to undergo validity and reliability testing.

4.7 In summary

- Many CAT designs ignore basic research and testing protocols.
- The structure for a new CAT is outlined, based on the evidence available.

- When using a CAT, a reliability and validation process should be completed on the scores obtained even if similar data are available from other sources.
- The next chapter evaluates the construct validity of the proposed CAT (Objective 4).

4.8 REFERENCES

1. Khan, K. S., ter Riet, G., Glanville, J., Sowden, A. J., & Kleijen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews (CRD Report 4). York, England: University of York.

2. Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitc, C., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27). doi:10.3310/hta7270

3. Petticrew, M. (2001). Systematic reviews from astronomy to zoology: Myths and misconceptions. *BMJ, 322*(7278), 98-101. doi:10.1136/bmj.322.7278.98

4. Dixon-Woods, M., Bonas, S., Booth, A., Jones, D. R., Miller, T., Sutton, A. J., ... Young, B. (2006). How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research, 6*(1), 27-44. doi:10.1177/1468794106058867

5. Moyer, A., & Finney, J. W. (2005). Rating methodological quality: Toward improved assessment and investigation. *Accountability in Research, 12*(4), 299-313. doi:10.1080/08989620500440287

6. Jüni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ, 323*(7303), 42-46. doi:10.1136/bmj.323.7303.42

7. Devers, K. J. (1999). How will we know "good" qualitative research when we see it? Beginning the dialogue in health services research. *Health Services Research, 34*(5 Part II), 1153-1188.

8. Jadad, A. R., Moher, D., & Klassen, T. P. (1998). Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? *Archives of Pediatrics and Adolescent Medicine, 152*(8), 812-817. doi:10.1001/archpedi.152.8.812

9. Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions (Version 5.0.1)*. London: The Cochrane Collaboration. Retrieved from http://www.cochrane-handbook.org

10. Heller, R. F., Verma, A., Gemmell, I., Harrison, R., Hart, J., & Edwards, R. (2008). Critical appraisal for public health: A new checklist. *Public Health, 122*(1), 92-98. doi:10.1016/j.puhe.2007.04.012

11. Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1060. doi:10.1001/jama.282.11.1054

12. Kuper, A., Lingard, L., & Levinson, W. (2008). Critically appraising qualitative research. *BMJ, 337*(7671), 687-689. doi:10.1136/bmj.a1035

13. Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130-149. doi:10.1037/1082-989X.13.2.130

14. Walsh, D., & Downe, S. (2006). Appraising the quality of qualitative research. *Midwifery, 22*(2), 108-119. doi:10.1016/j.midw.2005.05.004

15. Armijo Olivo, S., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156-175. doi:10.2522/ptj.20070147

16. Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *JAMA, 272*(2), 101-104. doi:10.1001/jama.1994.03520020027007

17. de Vet, H. C. W., de Bie, R. A., van der Heijden, G. J. M. G., Verhagen, A. P., Sijpkes, P., & Knipschild, P. G. (1997). Systematic reviews on the basis of methodological criteria. *Physiotherapy, 83*(6), 284-289. doi:10.1016/S0031-9406(05)66175-5

18. Bialocerkowski, A. E., Grimmer, K. A., Milanese, S. F., & Kumar, S. (2004). Application of current research evidence to clinical physiotherapy practice. *Journal of Allied Health, 33*(4), 230-237.

19. Burnett, J., Kumar, S., & Grimmer, K. (2005). Development of a generic critical appraisal tool by consensus: Presentation of first round Delphi survey results. *Internet Journal of Allied Health Sciences and Practice, 3*(1), 22. Retrieved from http://ijahsp.nova.edu/

20. Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

21. Glenny, A.-M. (2005). No "gold standard" critical appraisal tool for allied health research. *Evidence-Based Dentistry, 6*(4), 100-101. doi:10.1038/sj.ebd.6400351

22. Katrak, P., Bialocerkowski, A., Massy-Westropp, N., Kumar, V. S. S., & Grimmer, K. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology, 4*(1). doi:10.1186/1471-2288-4-22

23. Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16*(1), 62-73. doi:10.1016/0197-2456(94)00031-W

24. Sanderson, S., Tatt, I. D., & Higgins, J. P. T. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36*(3), 666-676. doi:10.1093/ije/dym018

25. Sutherland, S. E. (2004). An introduction to systematic reviews. *Journal of Evidence Based Dental Practice, 4*(1), 47-51. doi:10.1016/j.jebdp.2004.02.021

26. Law, M., Stewart, D., Pollock, N., Letts, L., Bosch, J., Westmorland, M., & Philpot, A. (2008). Occupational Therapy Evidence-Based Practice Research Group. Retrieved 29 January 2011, from http://www.srs-mcmaster.ca/Default.aspx?tabid=630

27. Crombie, I. K. (1996). *The pocket guide to critical appraisal: A handbook for health care professionals*. London: BMJ.

28. Cooper, H. (1986). *The integrative research review: A systematic approach* (2nd ed.). Beverly Hills, CA: Sage.

29. Glynn, L. (2006). A critical appraisal tool for library and information research. *Library Hi Tech, 24*(3), 387-399. doi:10.1108/07378830610692154

30. Duffy, M. E. (1985). A research appraisal checklist for evaluating nursing research reports. *Nursing & Health Care, 6*(December), 539-547.

31. Heacock, H., Koehoorn, M., & Tan, J. (1997). Applying epidemiological principles to ergonomics: A checklist for incorporating sound design and interpretation of studies. *Applied Ergonomics, 28*(3), 165-172. doi:10.1016/S0003-6870(96)00066-X

32. Meijman, F. J., & de Melker, R. A. (1995). The extent of inter- and intra-reviewer agreement on the classification and assessment of designs of single-practice research. *Family Practice, 12*(1), 93-97. doi:10.1093/fampra/12.1.93

33. Rangel, S. J., Kelsey, J., Colby, C. E., Anderson, J., & Moss, R. L. (2003). Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *Journal of Pediatric Surgery, 38*(3), 390-396. doi:10.1053/jpsu.2003.50114

34. Urschel, J. D. (2005). How to analyze an article. *World Journal of Surgery, 29*(5), 557-560. doi:10.1007/s00268-005-7912-z

35. Vickers, A. (1995). Critical appraisal: How to read a clinical research paper. *Complementary Therapies in Medicine, 3*(3), 158-166. doi:10.1016/S0965-2299(95)80057-3

36. Hawker, S., Payne, S., Kerr, C., Hardey, M., & Powell, J. (2002). Appraising the evidence: Reviewing disparate data systematically. *Qualitative Health Research, 12*(9), 1284-1299. doi:10.1177/1049732302238251

37. MacAuley, D. (1994). READER: An acronym to aid critical reading by general practitioners. *British Journal of General Practice, 44*(379), 83–85.

38. Nielsen, M. E., & Reilly, P. L. (1985). A guide to understanding and evaluating research articles. *Gifted Child Quarterly, 29*(2), 90-92. doi:10.1177/001698628502900210

39. Pluye, P., Gagnon, M.-P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies, 46*(4), 529-546. doi:10.1016/j.ijnurstu.2009.01.009

40. Rasmussen, L., O'Conner, M., Shinkle, S., & Thomas, M. K. (2000). The basic research review checklist. *Journal of Continuing Education in Nursing, 31*(1), 13-17.

41. Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377-384.

42. Moncrieff, J., Churchill, R., Drummond, D. C., & McGuire, H. (2001). Development of quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods in Psychiatric Research, 10*(3), 126-133. doi:10.1002/mpr.108

43. Cesario, S., Morin, K., & Santa-Donato, A. (2002). Evaluating the level of evidence of qualitative research. *Journal of Obstetric, Gynecologic, and Neonatal Nursing, 31*(6), 708-714. doi:10.1177/0884217502239216

44. Côté, L., & Turgeon, J. (2005). Appraising qualitative research articles in medicine and medical education. *Medical Teacher, 27*(1), 71-75.

45. Dixon-Woods, M., Shaw, R. L., Agarwal, S., & Smith, J. A. (2004). The problem of appraising qualitative research. *Quality and Safety in Health Care, 13*(3), 223-225. doi:10.1136/qshc.2003.008714

46. Long, A. F., & Godfrey, M. (2004). An evaluation tool to assess the quality of qualitative research studies. *International Journal of Social Research Methodology, 7*(2), 181-196. doi:10.1080/1364557032000045302

47. Reis, S., Hermoni, D., Van-Raalte, R., Dahan, R., & Borkan, J. M. (2007). Aggregation of qualitative studies - From theory to practice: Patient priorities and family medicine/general practice evaluations. *Patient Education and Counseling, 65*(2), 214-222. doi:10.1016/j.pec.2006.07.011

48. Treloar, C., Champness, S., Simpson, P. L., & Higginbotham, N. (2000). Critical appraisal checklist for qualitative research studies. *Indian Journal of Pediatrics, 67*(5), 347-351.

49. Stige, B., Malterud, K., & Midtgarden, T. (2009). Toward an agenda for evaluation of qualitative research. *Qualitative Health Research, 19*(10), 1504-1516. doi:10.1177/1049732309348501

50. Genaidy, A. M., Lemasters, G. K., Lockey, J., Succop, P., Deddens, J., Sobeih, T., & Dunning, K. (2007). An epidemiological appraisal instrument - A tool for evaluation of epidemiological studies. *Ergonomics, 50*(6), 920-960. doi:10.1080/00140130701237667

51. DuRant, R. H. (1994). Checklist for the evaluation of research articles. *Journal of Adolescent Health, 15*(1), 4-8. doi:10.1016/1054-139X(94)90381-6

52. Sindhu, F., Carpenter, L., & Seers, K. (1997). Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. *Journal of Advanced Nursing, 25*(6), 1262-1268. doi:10.1046/j.1365-2648.1997.19970251262.x

53. Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12. doi:10.1016/0197-2456(95)00134-4

54. Boutron, I., Moher, D., Tugwell, P., Giraudeau, B., Poiraudeau, S., Nizard, R., & Ravaud, P. (2005). A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. *Journal of Clinical Epidemiology, 58*(12), 1233-1240. doi:10.1016/j.jclinepi.2005.05.004

55. Melnyk, B. M., & Fineout-Overholt, E. (2005). Rapid critical appraisal of randomized controlled trials (RCTs): An essential skill for evidence-based practice (EBP). *Pediatric Nursing, 31*(1), 50-52.

56. Verhagen, A. P., de Vet, H. C. W., de Bie, R. A., Kessels, A. G. H., Boers, M., Bouter, L. M., & Knipschild, P. G. (1998). The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of Clinical Epidemiology, 51*(12), 1235-1241. doi:10.1016/S0895-4356(98)00131-0

57. Reisch, J. S., Tyson, J. E., & Mize, S. G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics, 84*(5), 815-827.

58. Evans, M., & Pollock, A. V. (1985). A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *British Journal of Surgery, 72*(4), 256-260. doi:10.1002/bjs.1800720403

59. Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials, 2*(1), 31-49.

60. Lichtenstein, M. J., Mulrow, C. D., & Elwood, P. C. (1987). Guidelines for reading case-control studies. *Journal of Chronic Diseases, 40*(9), 893-903. doi:10.1016/0021-9681(87)90190-1

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

62. Oxman, A. D., & Guyatt, G. H. (1988). Guidelines for reading literature reviews. *Canadian Medical Association Journal, 138*(8), 697-703.

63. Hunt, D. L., & McKibbon, K. A. (1997). Locating and appraising systematic reviews. *Annals of Internal Medicine, 126*(7), 532-538.

64. Wilson, A., & Henry, D. A. (1992). Meta-analysis Part 2: Assessing the quality of published meta-analyses. *Medical Journal of Australia, 156*(3), 173-174,177-178,180,184-187.

65. Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18*(4), 385-401. doi:10.1080/09602010802009201

66. Loney, P. L., Chambers, L. W., Bennett, K. J., Roberts, J. G., & Stratford, P. W. (1998). Critical appraisal of the health research literature: Prevalence or incidence of a health problem. *Chronic Diseases in Canada, 19*(4), 170-176.

67. Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity, 36*(4), 391-409. doi:10.1023/A:1020909529486

68. Dye, J. F., Schatz, I. M., Rosenberg, B. A., & Coleman, S. T. (2000). Constant comparison method: A kaleidoscope of data. *The Qualitative Report, 4*(1/2). Retrieved from http://www.nova.edu/ssss/QR/

69. Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA, 285*(15), 1992-1995. doi:10.1001/jama.285.15.1992

70. von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine, 4*(10), e296. doi:10.1371/journal.pmed.0040296

71. Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet, 354*(9193), 1896-1900. doi:10.1016/S0140-6736(99)04149-5

72. Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., ... Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA, 283*(15), 2008-2012. doi:10.1001/jama.283.15.2008

73. 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*(6), 349-357. doi:10.1093/intqhc/mzm042

74. Ogrinc, G., Mooney, S. E., Estrada, C., Foster, T., Goldmann, D., Hall, L. W., ... Watts, B. (2008). The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: Explanation and elaboration. *Quality and Safety in Health Care, 17*(Supplement 1), i13-i32. doi:10.1136/qshc.2008.029058

75. The Equator Network. (n.d.). EQUATOR: Enhancing the QUAlity and Transparency Of health Research. Retrieved 29 January 2011, from http://www.equator-network.org/

4.9 Additional material – Search strategy

OvidSP

CINHAL and MEDLINE (1996–2009)

1	(("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") and (tool* OR checklist* OR scale*)).ti,ab	1,889
2	limit 2 to English	1,717
3	remove duplicates from 2	1,499
	Meet criteria	54

Informit

A+ Education, ACCOUNT, AED, AEI-ATSIS, AEM, AFPD , AGIS Plus Text, AHB , AHRR, AIATSIS, AIMMAT, ALISA, ANR-Index, ANR-Index Archive, ANR-Research , ANR-Research Archive, ANSTI, ANZBiP, APA-FT, APAIS, APECLIT, ARCH, ARLIT, ASIANRES, ATI, ATRI, ATSIhealth, AUSCHRON, AUSPORT, AUSTGUIDE, AusportMed, BERITA, BIPE, BUILD, Business Collection, CHERUB, CHRONICLES, CIA, CINCH, CSI, DELTAA, DRUG, EDGE, ENDANGER, ENGINE, EVA, Engineering Collection, FNQ, Family & Society Plus, GIBLIN, HIVA, Health & Society, Humanities & Social Sciences Collection, ILRS, INDBIO, INTAN MAS, INTAX, IREL, Indigenous Australia, MAIS, MEDGE, MIHILIST, Media Scan, Meditext, OMNDIST, OMNRES, PDIP, PDIR, PERIND, REEF, RURAL, SAGE, SCANfile, SIAL, SMC, SNIPER, TAXABS, THESES, VALISE, VPI&E Catalogue, WORKLIT

#1	AB=("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review")	1,730
#2	TI=("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review")	1,645
#3	#1 OR #2	3,233
#4	AB=(tool* OR checklist* OR scale*) OR TI=(tool* OR checklist* OR scale*)	59,975
#5	#3 AND #4	179
	Meet criteria	0

CSA Illumina

ARTbibliographies Modern, ASFA, BioOne, Criminal Justice Abstracts, CSA Linguistics and Language Behavior Abstracts, DAAI, ERIC, Oceanic Abstracts, PsycARTICLES, PsycINFO, Social Services Abstracts, Sociological Abstracts, Zoological Record Plus

#1	(AB=("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") or TI=("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review")) and (AB=(tool* OR checklist* OR scale*) OR TI=(tool* OR checklist* OR scale*)) LIMITED TO: English only	1,364
	Meet criteria	29

EBSCOhost

EconLit; Film & Television Literature Index; Hospitality & Tourism Complete; Library, Information Science & Technology Abstracts; SPORTDiscus

#1	(AB=("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") or TI=("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review")) and (AB=(tool* OR checklist* OR scale*) or TI=(tool* OR checklist* OR scale*))	295
	Meet criteria	19

Gale InfoTrac

Expanded Academic ASAP; Health Reference Center Academic

	("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") AND (tool* OR checklist* OR scale*)	72
	Meet criteria	1

ProQuest

ABI/INFORM Global; Academic Research Library; Accounting & Tax Periodicals; Banking Information Source; Career and Technical Education; CBCA (Business and Education); Pharmaceutical News Index; ProQuest (All)

	("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") AND (tool* OR checklist* OR scale*) LIMIT TO: Citation and Abstract; English EXCLUDE: Book Reviews; Newspapers	1,009
	Meet criteria	30

ISI Web of Knowledge

#1	Title=(("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") AND (tool* OR checklist* OR scale*)) TIMESPAN: All Years DOCUMENT TYPE: Article OR Review LANGUAGE: English	142
	Meet criteria	8

JStore

#1	ab:("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") AND (tool* OR checklist* OR scale*) AND la:(en))	48
#2	ti:("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") AND (tool* OR checklist* OR scale*) AND la:(en))	9
	Meet criteria	1

Scopus

#1	TITLE-ABS-KEY(("critical appraisal" OR "critical assessment" OR "critical evaluation" OR "critical review" OR "quality appraisal" OR "quality assessment" OR "quality evaluation" OR "quality review" OR "research appraisal" OR "research assessment" OR "research evaluation" OR "research quality" OR "research review") AND (tool* OR checklist* OR scale*)) AND LANGUAGE(english) AND DOCTYPE(ar OR re) AND (LIMIT-TO(EXACTKEYWORD, "Methodology") OR LIMIT-TO(EXACTKEYWORD, "Rating scale") OR LIMIT-TO(EXACTKEYWORD, "Evaluation") OR LIMIT-TO(EXACTKEYWORD, "Scoring system") OR LIMIT-TO(EXACTKEYWORD, "Quality assessment") OR LIMIT-TO(EXACTKEYWORD, "Research Design")) AND (LIMIT-TO(SRCTYPE, "j"))	789
	Meet criteria	47

The Cochrane Library

#1	(CMR: critical appraisal):kw OR (CMR: checklists and guidelines):kw OR (CMR: quality assessments):kw	966
#2	((“critical appraisal” OR “critical assessment” OR “critical evaluation” OR “critical review” OR “quality appraisal” OR “quality assessment” OR “quality evaluation” OR “quality review” OR “research appraisal” OR “research assessment” OR “research evaluation” OR “research quality” OR “research review”) AND (tool* OR checklist* OR scale*)):ti	73
#3	#1 OR #2	849
	Meet criteria	130

Other research and reference lists

	Meet criteria	47

Chapter 5 – Evaluation of validity

As stated previously, a large number of critical appraisal tools (CATs) have been developed with little or no evaluation of the validity of their scores. This chapter refines the outline of the proposed CAT from Chapter 4 and evaluates the validity of the scores obtained by its use, thereby satisfying Objective 4 of the study.

The chapter consists of an article accepted for publication on 18 June 2011 and available online 30 July 2011 (Appendix C.4):

> Crowe, M., & Sheppard, L. (2011). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies*, *14*(12). 1505-1516. doi:10.1016/j.ijnurstu.2011.06.004

Changes have been made to the published article to ensure thesis consistency. Copyright permission, which allows this paper to be reproduced, can be found in Appendix A.3.

Note

The treatment of validity in this and the previous chapter are different. In the previous chapter, validity was explored as having three types: content, concurrent, and construct validity. This approach was taken because the majority of papers that described validity in that chapter did so in those terms. In this chapter validity is described as the unitary concept of construct validity. Other validities (such as

content and concurrent) are seen as subsets of construct validity. The use of construct validity as a unitary concept follows current validity theory in the area of measurement and testing. These differences in how validity is conceptualised are further discussed below.

A general critical appraisal tool: An evaluation of construct validity

5.1 ABSTRACT

Background – Many critical appraisal tools (CATs) exist for which there is little or no information on development of the CAT, evaluation of validity, or testing reliability. The proposed CAT was developed based on a number of other CATs, general research methods theory, and reporting guidelines, but requires further study to determine its effectiveness.

Objectives – To establish a scoring system and to evaluate the construct validity of the proposed critical appraisal tool before undertaking reliability testing.

Methods – Data obtained from this exploratory study and information on the design of the proposed CAT were combined to evaluate construct validity using the *Standards for educational and psychological testing*, which consist of five types of evidence: test content, response process, internal structure, relations to other variables, and consequences of testing. The proposed CAT was analysed against five alternative CATs to obtain data for internal structure and relations to other variables. A random sample of 1o papers from each of six different research designs across the range of health-related research were selected, giving a total sample size of 60 papers.

Results – In all categories within the proposed CAT there were significant ($p < 0.05$, 2-tailed) weak to moderate positive correlations (Kendall's tau $0.33 \leq \tau \leq 0.55$) with the alternative CATs, except in the *Preamble* category. There were significant moderate to strong positive correlations in quasi-experimental ($0.70 \leq \tau \leq 1.00$), descriptive, exploratory or observational ($0.72 \leq \tau \leq 1.00$), qualitative

($0.74 \leq \tau \leq 0.81$), and systematic review ($0.62 \leq \tau \leq 0.82$) designs, and to a lesser extent in true experimental design ($0.68 \leq \tau \leq 0.70$). There were no significant correlations in the single system research designs.

Conclusions – Based on the results obtained, the theory on which the proposed CAT was designed, and the objective of the proposed CAT there was enough evidence to suggest that sound inferences can be made based on the scores obtained when using the proposed CAT.

5.2 INTRODUCTION

The purpose of a critical appraisal tool (CAT) is to assist readers to rate research papers based on the research methods used and the conclusions drawn. CATs are used in systematic reviews, literature reviews, and anywhere a reader wishes to remain objective when reading a research paper. However, there are a number of well documented problems with many existing CATs. These problems include:

1. Tools that are limited in the research designs they can evaluate. In many cases only one research design can be evaluated by a particular CAT and different papers with different research designs cannot be directly compared [1, 2].
2. Tools that lack the depth to comprehensively assess papers being analysed, so that not all aspects of the research undertaken are appraised [2, 3].
3. Tools that use inappropriate scoring systems, such as simplistic summary scores, in such a way that defects in studies may be hidden [2, 4, 5].
4. Tools developed with little regard for basic research techniques so that there is limited or no validity or reliability data. Therefore, these tools cannot claim to validly and reliably assess the research appraised [1, 6, 7].

The review of CAT design in Chapter 4 suggested a new structure for a CAT to overcome problems one and two [1]. The structure was based on a qualitative analysis of 45 CATs where information on the design of the CATs was available. The analysis used was the constant comparative method where each item from one CAT was compared with items from other CATs, so that distinct categories of items were created. A combination of research methods theory and standards for the reporting of research, such as CONSORT (Consolidated Standards for Reporting of Trials) [8], PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses, formerly QUOROM) [9] and COREQ (Consolidated Criteria for Reporting Qualitative Research) [10], were also used. The 45 CATs that were analysed could be used to appraise different research designs, so the proposed CAT could potentially be used across all those research designs. This process culminated in: a list of eight categories (*Preamble*, *Introduction*, *Design*, *Sampling*, *Data collection*, *Ethical matters*, *Results*, and *Discussion*), where each category included similar information and there was no overlap between categories; 22 items; and a large number of item descriptors upon which a research paper could be assessed.

This chapter builds on Chapter 4 by tackling the two outstanding details of the proposed CAT yet to be resolved: the scoring system (problem 3); and, validity and reliability testing (problem 4) [1]. Validity and scoring are the subject of this paper. They were considered independently from reliability because validity and scoring are heavily intertwined, and validity is "...the most fundamental consideration in developing and evaluating" a tool [11 (p. 9)]. It is irrelevant whether a tool has reliability if there is no proper and thorough validity testing [11 (pp. 9-11)]. In other words, if all raters independently agree on a score a paper should receive (reliability), this is immaterial if the score does not accurately reflect what is being measured (validity). Therefore, validation of the proposed CAT was required before reliability could be examined. Reliability is the subject the next chapter.

However, before exploring the methods used in this research, the exact nature of validity needs to be explained. This is necessary to counteract the persistent belief that: (1) the definition of validity is, 'does the test measure what it is supposed to measure?'; and, (2) that validity consists or is the sum of many different types of validity (for example content, criterion, construct, face, divergent, convergent, predictive, concurrent) [12 (pp. 249-252)].

5.2.1 Construct validity

The American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurements Used in Education (NCMUE) expanded the definition of validity to four types between 1952 and 1954 to overcome shortfalls in validity theory [13]. The four types of validity identified were content, predictive, concurrent, and construct validity. However, even at that time, predictive and concurrent validity were normally considered to be part of the same type of validity called criterion-related validity [14]. This view of validity lead directly to the threes C's of validity that is often used today and which consists of three separate types of validity (content, criterion, and construct) [12 (pp. 249-252)].

Almost as soon as the four-types model of validity was introduced, some authors were doubtful of its veracity. In 1955, Cronbach and Meehl stated that content validity could be considered as part of construct validity [14]. In 1957, Loevinger went even further to state that predictive, concurrent, and content validity were *ad hoc* hypotheses and, therefore, of no scientific importance. This meant that only construct validity was worthy of study, even if the other validities existed [15].

Further research into validity throughout the 1960s and 1970s led to the establishment of a unified theory of validity in the mid-1980s which stated that construct validity was the only validity [11 (pp. 9-11), 16]. This unitary view

maintained that validity refers to "...the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" [11 (p. 9)], where test refers to an evaluation method. In other words, validity referred to the interpretation of scores based on:

1. The theory upon which the test was built.
2. The evidence for how the scores can be interpreted.
3. The stated context for test use.

Therefore, it cannot be claimed that a test is valid. All that can be said is that under the assumptions around which a test was built, the evidence shows that scores can be interpreted in a certain way. If any of the theoretical, evidential, or contextual aspects of the test change, then validity must be re-examined and interpretation of the score may also change. In short, validity is an ongoing process where evidence for how test scores can be interpreted is required each time a test is administered [11 (p. 11), 16, 17].

The unified approach to validity was formalised in the *Standards for educational and psychological testing* in 1985 and further refined in 1999 [11, 18]. Given the poor track record in CAT validity [1], the *Standards* were used in validity testing the proposed CAT because: (1) they are a clear guide to validity; and (2) they can be applied to any evaluation method and, therefore, can be applied to a CAT (that is, a method for evaluating research) [11 (p. 3)].

5.2.2 Validity evaluation

Evaluating construct validity is considered a mixture of reasoned argument, theoretical foundations, and empirical evidence that together support the credibility of score interpretation [12 (pp. 249-252), 16, 17]. Five types of evidence are gathered to evaluate construct validity: *Test content*, *Internal structure*, *Response process*,

Relations to other variables, and *Consequences of testing* [11 (pp. 11-17)]. These types of evidence are described below in relation to the development of a CAT.

Test content

Test content explores the specification of the construct, analysis of test content against the construct (for example themes, words, formats, questions, procedures, guidelines), and threats to construct validity. Analysis is a mixture of logic, empirical evidence and expert judgement [11 (pp. 11-12)].

The major threats to construct validity are construct underrepresentation and construct-irrelevant variance, either or both of which may be present within a test [11 (p. 10), 16]. Construct underrepresentation is when a test is too narrowly focused and fails to include important aspects of a construct. In a CAT, this means that certain aspects of research evaluation may be omitted and the resulting score is not representative of the breadth of research methods used. This is an argument that could be used against the Jadad scale, a CAT commonly used to appraise true experimental research designs, which has only three criteria against which to appraise a research paper [19].

Construct-irrelevant variance is when a test is too broad and includes items that are not relevant to the construct being measured [11 (p. 10), 16]. In a CAT, this can mean three things: (1) the tool includes items which over-represent one aspect of research design (for example, multiple references to blinding); (2) the tool includes items that favour one research design over another (for example, true experimental designs over any other designs); or (3) the tool includes items that are not related to appraising research. In case 1 and 2, a positive or negative bias is introduced toward one or more aspects of research design, while in case 3 the items should be removed from the test.

In designing a CAT, or any test, there is a struggle between ensuring that representative items for a construct are included in the CAT and also ensuring that none of the items can lead to one or more aspects of research design having an advantage or disadvantage [16]. However, what constitutes construct underrepresentation or construct-irrelevant variance is open to interpretation, although methods, such as expert consensus, can be used to reduce subjectivity. For example, some authors argue that only items which refer to how a research design was implemented (also known as internal research validity) should be included in critical appraisal, while others argue that critical appraisal is a broader undertaking and should include ethics and the suitability of a research design [20, 21].

Internal structure

The relationship between test items should reflect the construct on which score interpretations are based. If a construct is unidimensional, for example, then the test items must be homogeneous within the dimension posited and interpretation of the score must be based on the assumed unidimensionality of the construct. Internal structure also asks whether items in the test function differently for different subgroups of test takers. Analysis is theoretical and empirical [11 (p. 13)].

Whether a construct is unidimensional or not, current construct validity theory favours measuring homogeneous constructs together with a single score [11 (p. 13), 17]. The advantage of this approach is that changes to the score reflect changes in measurement of a single underlying construct. Previously, scores could represent multiple (heterogeneous) constructs and changes in a score could not be attributed to any particular construct. However, multiple homogeneous constructs may be added together where there is enough theoretical and empirical evidence to show that a single score will enhance understanding without impairing precision of score interpretation [17].

There are four ways to envision the internal structure of research in relation to CATs [1, 2]. Research can be seen as:

1. Too complex to be reduced to numbers.
2. A unidimensional construct that can be allocated a single score.
3. Multiple constructs that should be scored for each individual construct without summing the scores.
4. Multiple constructs where summing individual construct scores into a single score does not affect the precision of the scores.

The first point has an air of intellectual laziness which supposes anything complex is too difficult to be understood in simple terms. This contradicts scientific reason and, therefore, the first point was rejected. The second point, explained in the previous chapter, was that treating research as a unidimensional construct may be deceptive. Therefore, a simple single score for appraising research may hide weak sections of the research and should not be encouraged. Consequently, points 3 and 4 were explored as methods for scoring research.

Response processes

Response process ensures there is a fit between the processes used by a test taker to deliver a response and the construct being tested. This is why test takers in mathematics, for example, are asked to show how they arrived at an answer and not just provide the final result. The response process also includes whether test scores can be interpreted in the same way across subgroups of test takers. Analysis is theoretical and empirical [11 (pp. 12-13)].

A CAT, therefore, should give a reader the ability to include where they found evidence for different aspects of the research and why they thought this constituted evidence for or against giving a particular score given to the research. Having

information about the reader, such as their research experience, enables researchers to gather evidence about potential differences between subgroups of readers.

Relations to other variables

Test scores can be analysed in relation to scores from other tests of the same or related constructs, criteria the test is meant to predict, and measures other than test scores that are hypothesised to be related to the construct in question. In terms of 'traditional' validity this encompasses such evidence as convergent, discriminant, predictive and concurrent validity, and validity generalisation. However, it must be remembered that these validities are subsets of the unitary concept of construct validity and not different types of validity [11 (pp. 13-16)]. Analysis is primarily empirical. In a new CAT, scores should be tested against existing CATs, where the existing CATs have validity and reliability data available, and are reported to test similar or the same constructs as the new CAT.

Consequences of testing

Test scores can be analysed in relation to intended and unintended consequences of score interpretation. Intended consequences of score interpretation occur when a benefit can be realised. However, that benefit must have the possibility of being realised and must not be overstated; that is, the claims must be backed by empirical evidence. Unintended consequences of score interpretation may occur when there are threats to construct validity (construct underrepresentation and construct-irrelevant variance), or when the test scores are misinterpreted or misused. It is generally up to the test developers (the authors of the test) and test users (administrators of the test) to ensure that misinterpretation and misuse does not occur [11 (pp. 16-17)].

A CAT should not overstate its usefulness in appraising a paper and should not be used outside the research method or methods it was designed to appraise. Scores obtained from using the Jadad scale [19] for any research except health-related true experimental research, of example, could not be considered valid until they have been subject to appropriate evaluation of validity. Furthermore, most CATs are designed to be used for a particular project. It would generally be inappropriate to use the scores obtained from a CAT in one project and use them for another project because the contexts would be different. There are exceptions, however, such as the PEDro scale [7], where a CAT has undergone extensive validity and reliability testing for this purpose.

5.2.3 Study outline

This chapter aims to evaluate the validity of the scores collected from the proposed CAT using the approach outlined above. Test content and internal structure evidence were primarily collected in the review of CATs (Chapter 4) [1]. Response processes and relations to other variables evidence were largely collected in this study, as outlined in the Methods and Results sections below. The Discussion section combines arguments from the previous chapter, the results in this chapter, and statements regarding consequences of testing to form a theoretical, logical, and empirical argument for the validity of the scores obtained by the proposed CAT.

5.3 METHODS

The following steps regarding the scoring system, response process, and relations to other CATs were undertaken to meet the aims of the research for the proposed CAT:

1. Develop the scoring system and a user guide for the proposed CAT.
2. Pre-test the proposed CAT and user guide, and make amendments where necessary.

3. Compare the scores achieved by the proposed CAT against the scores achieved by an alternative CAT or CATs, where the alternative CAT or CATs must have validity and reliability data available.

5.3.1 Scoring system and user guide

Streiner and Norman [12 (pp. 48-49)] showed that scales with 5–7 numbers are most appropriate for a measurement tool. Other CATs have used a variety of scales including: Yes/No/Unknown/Not applicable; scoring 0-1, 1–3, or 1–6 points; or a combination of these. It was decided to use a 6-point scale, from 0–5, to score each category in the proposed CAT because a smaller range would be too narrow to accurately score papers. Choosing a scale from 0–5 also allows an appraiser to rate a category as 0 if there is no evidence for that category in the paper, 1 if there is the least evidence, a middle score of 3, and a highest score of 5. It was decided that this would give an appraiser high enough fidelity to accurately appraise each category and subsequently each paper. Also, only integers (whole numbers) were to be used in scoring to force an appraiser to make a decision rather than choose a half score.

A user guide was written for the proposed CAT (section 5.9.1, p. 117). It contained guidelines and procedures for scoring a research paper. The user guide included what was meant by each category, item and item descriptor within the proposed CAT, and how the scoring system was implemented. However, the user guide was not prescriptive. Appraisers were encouraged to use their judgement by taking all aspects of a category into account before assigning a score. Therefore, scoring entailed an objective and subjective assessment of each category.

5.3.2 Research design

Due to the aims of this research, the overall design was exploratory in nature. Since the proposed CAT had the purpose of appraising different types of research, broad groupings of research designs were chosen. The research design types, based on common groupings of research designs in the literature [22, 23, 24] and Chapter 3, were:

1. True experimental (for example Pre-test/post-test control group, Solomon four-group, Post-test only control group, Randomised two-factor, Placebo controlled trial).
2. Quasi-experimental (for example Post-test only, Non-equivalent control group, Counter balanced (cross-over), Separate sample pre-test post-test, Multiple time series).
3. Single system (for example One-shot experimental (case study), Simple time series, One group pre-test/post-test, Within subjects, Multiple baseline).
4. Descriptive, exploratory or observational (for example Cross-sectional, Longitudinal, Retrospective, Prospective, Correlational, Predictive, Cohort, Case-control, Survey).
5. Qualitative (for example Phenomenology, Ethnography, Grounded theory, Narrative, Narrative case study).
6. Systematic reviews (not limited to meta-analysis).

Pre-testing used the structure as outlined in Chapter 4 for the proposed CAT [1]. It was not possible to compare the proposed CAT against a single alternative CAT because no such tool was found that covered all health research designs, and had validity and reliability data available. Instead, five comparison tools were chosen that had validity and reliability data available, based on the review of CATs (Chapter 4). The user guide or other publications for the alterative CATs were used to help score papers. The alternative CATs (section 5.9.2, p. 128) were:

1. Physiotherapy Evidence Database (PEDro) scale for true experimental designs [7].
2. Cho and Bero scale for quasi-experimental, and descriptive, exploratory or observational designs [25].
3. Single-Case Experimental Design (SCED) scale for single system designs [26].
4. Reis et al scale for qualitative designs [27].
5. Assessment of Multiple Systematic Reviews (AMSTAR) for systematic reviews [28].

The outcome to be assessed was the scores collected by each tool, that is the proposed CAT and each alternative CAT. This outcome was assessed by a single appraiser (the author) which was determined to be sufficient for a validation process because separate appraisals of each paper were made using two different tools. This is similar to two appraisers using one tool from a testing point of view [29]. Also, this research was the first step of validity testing, which is seen as a continuing process rather than a one off exercise [11 (p. 9), 16, 17]. As further research into the proposed CAT is undertaken, more evidence of validity can be obtained.

5.3.3 Sample of papers

The sample size was calculated based on formulae provided by Streiner and Norman [12 (pp. 198-202)] to show a large enough difference between the proposed tool and the alternative CATs. This method was developed for sample size calculations in testing theory and should provide an approximation for validity testing. The formulae used were:

$$n = 2 + \frac{k}{2(k-1)(z_R - z_{R^-})^2}$$

Where

$$z_R = \frac{1}{2}\ln\left[\frac{1+(k-1)r}{1-r}\right]$$

$$z_{R^-} = \frac{1}{2}\ln\left[\frac{1+(k-1)\left(r-\frac{\mathrm{CI}}{2}\right)}{1-\left(r-\frac{\mathrm{CI}}{2}\right)}\right]$$

n, number of raters
k, number of papers
ln, natural log
r, reliability coefficient
CI, confidence interval

A Microsoft Excel worksheet function and decision table were developed (section 5.9.3, p. 133) using these formulae to calculate sample size based on the known values: number of observations (n) was 2, the confidence interval (CI) was 0.19, and the reliability coefficient (r) was 0.90. Therefore, the sample size (k) was calculated as 10. However, the research designs were split into six groups, which meant that ten papers were required from each research design group. This gave a total sample size (N) of 60 papers.

Two different methods were trialled for selecting papers for the research. The first method was to use previous reviews of the literature to obtain a wide range of research papers which used various research designs. This method proved too difficult to acquire a reasonable variety of papers with a reasonable spread of research designs.

The second selection process, which took place in September 2009, used the full text of journals subscribed to by James Cook University (JCU) in OvidSP (Ovid, New York). This gave access to a total of 278 journals primarily from Medline (1948-2009) but also from Biological Abstracts (1969-2001), BIOSIS Previews (2002-2008), and International Pharmaceutical Abstracts (1970-2009). This guaranteed that the author could obtain papers without needing to use inter-library loans or spend an extended period of time finding the papers in printed or on-line journals. A

search strategy (Table 5.1) was put in place to find unique papers in each research design type (Table 5.1, Part 1). The search terms chosen were applied only to the abstract of full text journal papers to reduce the likelihood of false positives. Each design search was then limited to papers with references, and original or review papers (Table 5.1, Part 2). This limited the papers returned to substantial research papers rather than, for example, editorials or news items. Only papers that were in a specific research design, minus duplicates from other research designs, were allowed. This ensured that the papers in each research design were unique (Table 5.1, Part 3).

Table 5.1 Paper search strategy

	Part 1 Research design	**No.**
1	("Randomised controlled trial" or "Randomized controlled trial").ab	1,029
2	"Quasi-experimental".ab	155
3	("Single-subject" or "N of 1" or "N-of-1" or "Single system").ab	873
4	("Descriptive research" or "Exploratory research" or "Observational research" or "Cohort study" or "Survey research").ab	2,946
5	("Narrative research" or "Phenomenology" or "Phenomenological" or "Grounded theory" or "Ethnography" or "Ethnographical" or "Narrative case study").ab	387
6	"systematic review".ab	459
	Part 2 Limit papers	.
7	limit 1 to (articles with references and (original articles or "review articles"))	872
8	limit 2 to (articles with references and (original articles or "review articles"))	117
9	limit 3 to (articles with references and (original articles or "review articles"))	662
10	limit 4 to (articles with references and (original articles or "review articles"))	2,475
11	limit 5 to (articles with references and (original articles or "review articles"))	233
12	limit 6 to (articles with references and (original articles or "review articles"))	356
	Part 3 Unique papers	
13	7 not (8 or 9 or 10 or 11 or 12)	859
14	8 not (7 or 9 or 10 or 11 or 12)	116
15	9 not (7 or 8 or 10 or 11 or 12)	645
16	10 not (7 or 8 or 9 or 11 or 12)	2,446
17	11 not (7 or 8 or 9 or 10 or 12)	230
18	12 not (7 or 8 or 9 or 10 or 11)	332

OvidSP numbers each paper from 1 to n for each search. Using the random sequence generator from RANDOM.ORG [30], the papers 1 to n for each set of results were assigned a random number in case OvidSP placed the papers were in a specific sequence. A pool of 30 papers per research design was then randomly selected for

appraisal, again using the above random sequence generator, and the exact sequence of each paper was noted. This process ensured that the papers were randomised first and, in turn, the papers were randomly selected so that the author had no influence on the papers selected.

The first paper selected for each research design was used for pre-testing, while papers 2–11 were used in the main study. If any of the first 11 papers was unsuitable, then papers 12–30, in order, could be used to achieve a full sample size.

5.3.4 Data collection and analysis

Each paper was read in the period December 2009 to February 2010. The proposed CAT was used first in each case, followed immediately by the alternative tool. Although this could lead to order effects, it was decided that this sequence was less likely to cause confusion or inappropriate scoring due to the use of multiple alternative CATs. If only one alternative CAT was used then the order of CAT used would have been randomly allocated for each paper appraised. Data were collated only when all 60 papers had been read and scored. This was to reduce bias where the author could consciously or unconsciously observe patterns between the tools used.

An assumption was made that each category in the proposed CAT could be treated as a separate homogeneous construct in order to overcome the previously mentioned problems with summary scores. The assumption was made based on the methods used to create the proposed CAT that were outlined in Chapter 4 [1]. This enabled evaluation of the proposed CAT (a multi-dimensional construct) against each of the alternative CATs (notionally, unidimensional constructs) [17]. However, not all alternative CATs contained items for each of the categories in the proposed CAT. In those cases no direct comparison could be made between the proposed CAT and the alternative CAT for that category. Furthermore, the total scores for the proposed

CAT and the alternative CATs were compared because it could be possible that the proposed CAT total score was sufficient for score interpretation without impairing precision [17] .

Scores for each category collected by the alternative CATs needed to be converted into a format comparable with the proposed CAT. The process used was to total the scores in each category and to divide by the total possible score available in that category. This number was then converted to a score out of five. There was a concern whether to use the category scores rounded to two decimal places or to convert the score to the nearest integer. In the total percentage score, the decision was between the percentage based on the raw score or either of the converted scores. It was decided that all these scores would be used and checked to see if there were any differences in the results.

Scores obtained by the proposed CAT and the alternative CATs were correlated. The exact correlation method used was dependent on the nature of the data obtained.

5.3.5 Ethics

This research was part of a larger study which received authorisation from the James Cook University Human Ethics Committee (Approval No. H3415). There were no conflicts of interest or funding sources to declare.

5.4 RESULTS

A total of 36 papers were rejected from the papers randomly selected because they did not have the required research design. The rejected papers included six from true experimental, four from quasi-experimental, 12 from single system, one from

DEO, three from qualitative, and 10 from systematic review. A full list of papers used in pre-testing and the main study is available in section 5.9.2 (p. 133).

5.4.1 Pre-testing

Only minor changes were required to the proposed CAT (see Table 5.2 for a sample of the form used in the main study) and user guide (see section 5.9.3, p. 117) after both were pre-tested by the author. The changes made were to ensure that wording and the order of item descriptors were consistent throughout. Check boxes were introduced to help appraisers keep track of items that were present, absent but should have been present, or were not applicable to the paper being appraised. These changes made the tool easier to use but did not alter underlying assumptions or reduce the tool to a simple checklist.

5.4.2 Main study

The publication year for the sample of papers was 12 (20%) in the 1990s and 58 (80%) in the 2000s. The majority of papers were published in: *Critical Care Medicine*, 8 (13%); *JAMA*, 6 (10%); *Cancer Nursing*, 4 (7%); and 3 (5%) each in the journals *Neurology* and *Neurosurgery*. Seven journals published two papers each, 14 (23%), and 22 journals published the remaining 22 papers (37%). Overall, there were 30 different health related topics covered in the 60 papers. The most common topics were Intensive Care Unit (ICU) and surgery, with 5 papers (8%) each. Geriatrics and oncology had 4 papers (7%) each. Depression, drug-gene interaction, human immunodeficiency virus (HIV), neonates, neurology, and pain management had 3 papers (5%) each. The remaining 24 papers (40%) were spread among 19 separate topics.

Table 5.2 Proposed CAT structure after initial pilot

Category Item	**Item descriptor**	**Score** [0–5]
Preamble		**Preamble**
Text	1. Sufficient detail others could reproduce ☐ 2. Clear/concise writing ☐, table(s) ☐, diagram(s) ☐, figure(s) ☐	
Title	1. Includes study aims ☐ and design ☐	
Abstract	1. Key information ☐ 2. Balanced ☐ and informative ☐	
Introduction		**Introduction**
Background	1. Summary of current knowledge ☐ 2. Specific problem(s) addressed ☐ and reason(s) for addressing ☐	
Objective	1. Primary objective(s), hypothesis(es), or aim(s) ☐ 2. Secondary question(s) ☐	
Design		**Design**
Research design	1. Research design(s) chosen ☐ and why ☐ 2. Suitability of research design(s) ☐	
Intervention, Treatment, Exposure	1. Intervention(s)/treatment(s)/exposure(s) chosen ☐ and why ☐ 2. Precise details of the intervention(s)/treatment(s)/exposure(s) ☐ for each group ☐ 3. Intervention(s)/treatment(s)/exposure(s) valid ☐ and reliable ☐	
Outcome, Output, Predictor, Measure	1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen ☐ and why ☐ 2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s) ☐ 3. Outcome(s)/output(s)/predictor(s)/measure(s) valid ☐ and reliable ☐	
Bias, etc	1. Potential bias ☐, confounding variables ☐, effect modifiers ☐, interactions ☐ 2. Sequence generation ☐, group allocation ☐, group balance ☐, and by whom ☐ 3. Equivalent treatment of participants/cases/groups ☐	
Sampling		**Sampling**
Sampling method	1. Sampling method(s) chosen ☐ and why ☐ 2. Suitability of sampling method ☐	
Sample size	1. Sample size ☐, how chosen ☐, and why ☐ 2. Suitability of sample size ☐	
Sampling protocol	1. Target/actual/sample population(s): description ☐ and suitability ☐ 2. Participants/cases/groups: inclusion ☐ and exclusion ☐ criteria 3. Recruitment of participants/cases/groups ☐	
Data collection		**Data collection**
Collection method	1. Collection method(s) chosen ☐ and why ☐ 2. Suitability of collection method(s) ☐	
Collection protocol	1. Include date(s) ☐, location(s) ☐, setting(s) ☐, personnel ☐, materials ☐, processes ☐ 2. Method(s) to ensure/enhance quality of measurement/instrumentation ☐ 3. Manage non-participation ☐, withdrawal ☐, incomplete/lost data ☐	
Ethical matters		**Ethical matters**
Participant ethics	1. Informed consent ☐, equity ☐ 2. Privacy ☐, confidentiality/anonymity ☐	
Researcher ethics	1. Ethical approval ☐, funding ☐, conflict(s) of interest ☐ 2. Subjectivities ☐, relationship(s) with participants/cases ☐	
Results		**Results**
Analysis, Integration, Interpretation method	1. A.I.I. method(s) for primary outcome(s)/output(s)/predictor(s) chosen ☐ and why ☐ 2. Additional A.I.I. methods (e.g. subgroup analysis) chosen ☐ and why ☐ 3. Suitability of analysis/integration/interpretation method(s) ☐	
Essential analysis	1. Flow of participants/cases/groups through each stage of research ☐ 2. Demographic and other characteristics of participants/cases/groups ☐ 3. Analyse raw data ☐, response rate ☐, non-participation/withdrawal/incomplete/ lost data ☐	
Outcome, Output, Predictor analysis	1. Summary of results ☐ and precision ☐ for each outcome/output/predictor/measure 2. Consideration of benefits/harms ☐, unexpected results ☐, problems/failures ☐ 3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes) ☐	
Discussion		**Discuss**
Interpretation	1. Interpretation of results in the context of current evidence ☐ and objectives ☐ 2. Draw inferences consistent with the strength of the data ☐ 3. Consideration of alternative explanations for observed results ☐ 4. Account for bias ☐, confounding/effect modifiers/interactions/imprecision ☐	
Generalisation	1. Consideration of overall practical usefulness of the study ☐ 2. Description of generalisability (external validity) of the study ☐	
Concluding remarks	1. Highlight study's particular strengths ☐ 2. Suggest steps that may improve future results (e.g. limitations) ☐ 3. Suggest further studies ☐	

All data were entered into SPSS version 18 (IBM SPSS, Chicago IL), examined and were found not to be normally distributed. When the alternative CATs' scores were examined, there was no statistical difference between the scores rounded to two decimal places or to the nearest integer. As a result, all alternative CAT results were based on the integer values because the proposed CAT was restricted to integer values. There was no statistical difference between the raw score percentage and the converted score percentage. Consequently, the raw score percentage was used.

The average scores obtained are shown in Table 5.3. Due to the primary aim of this research, the relationship between scores for the proposed CAT and the alternative CATs was most important, while the raw score obtained by each paper was not. However, the proposed CAT's and the alternative CATs' scores were quite high. The lowest total score % for the proposed CAT was 64% and for the alternative CATs it was 56%.

Table 5.3 Average scores for proposed CAT vs alternative CATs

Category	Research designs												All research designs		
	TE (n=10)		**QE** (n=10)		**SS** (n=10)		**DEO** (n=10)		**QL** (n=10)		**SR** (n=10)				
	Pro	*Alt*	*Pro*	*Alt*	*Pro*	*Alt*	*Pro*	*Alt*	*Pro*	*Alt*	*Pro*	*Alt*	*N*	*Pro*	*Alt*
Preamble	4.6	.	4.0	.	3.8	.	3.8	.	3.7	3.4	4.2	.	10	4.0	3.4
Introduction	5.0	.	4.9	4.8	4.9	.	5.0	5.0	4.7	4.4	4.9	4.0	40	4.9	4.6
Design	3.2	3.2	3.5	3.3	3.3	2.9	4.0	3.0	4.2	3.9	3.4	2.5	60	3.6	3.1
Sampling	2.5	4.0	2.3	3.0	2.0	.	3.3	2.6	2.9	3.0	2.8	1.9	50	2.6	2.9
Data collection	4.0	4.4	3.1	.	3.4	3.6	3.8	.	3.6	4.1	2.7	3.4	40	3.4	3.9
Ethical matters	2.5	.	1.2	2.0	0.6	.	1.4	3.3	1.6	.	1.9	1.5	30	1.5	2.3
Results	4.1	4.5	2.9	3.5	3.7	2.6	3.6	3.9	3.2	3.8	3.3	3.8	60	3.5	3.7
Discussion	3.7	.	3.9	4.6	3.7	2.0	4.0	5.0	3.2	3.2	3.7	.	40	3.7	3.7
Total score %	74	81	65	71	64	56	72	76	68	74	67	57	60	68	69

Pro, proposed CAT; Alt, alternative CAT(s); n, papers per research design; N, total papers all designs; TE, true experimental; QE, quasi-experimental; SS, single system; QL, qualitative; DEO, descriptive, exploratory, and observational; SR, systematic review; ., No data.

The proposed CAT scored true experimental, quasi-experimental, DEO, and qualitative research designs lower than the alternative CATs. On the other hand, the proposed CAT scored single system and systematic review designs higher than the

alternative CATs. The proposed CAT scored higher for the *Preamble*, *Introduction*, and *Design* categories and lower for *Sampling*, *Data collection*, *Ethical matters*, *Results*, and *Discussion* in relation to the alternative CATs. The total score % for all research designs was approximately equal for the proposed CAT (68%) and the alternative CATs (69%). These comparisons were also apparent in the raw data.

Kendall's tau (τ) was the most appropriate correlation coefficient to use because the data were not normally distributed and there were a large number of tied ranks [31 (pp. 1371-1385)]. Raw data were analysed pairwise and broken down by research design, proposed CAT category, and total score %. A summary of the results is available in Table 5.4. It was decided to use the following as a guide to Kendall's tau before analysis began: strong, greater than or equal to 0.75, moderate, 0.50 to 0.74, weak, 0.25 to 0.49, and little or no relationship, below 0.25 [24 (p. 525)]. Significance for results below was $p < 0.05$, 2-tailed.

Table 5.4 Kendall's tau for proposed CAT vs alternative CATs

Category	Research designs												All research designs		
	TE (n=10)		**QE** (n=10)		**SS** (n=10)		**DEO** (n=10)		**QL** (n=10)		**SR** (n=10)				
	tau	*p*	*tau*	*p*	*tau*	*p*	*tau*	*p*	*tau*	*p*	*tau*	*p*	*N*	*tau*	*p*
Preamble	.	.	.	.	.	.	.	.	0.49	0.10	.	.	10	0.49	0.10
Introduction	.	.	1.00	0.00*	.	.	1.00	0.00*	0.74	0.02*	-0.17	0.62	40	0.52	0.00*
Design	0.68	0.02*	0.50	0.08	0.19	0.56	-0.25	0.41	0.43	0.16	0.53	0.08	60	0.43	0.00*
Sampling	0.70	0.03*	0.81	0.01*	.	.	0.28	0.37	0.76	0.01*	0.31	0.31	50	0.40	0.00*
Data collection	-0.33	0.30	.	.	-0.14	0.65	.	.	0.19	0.53	0.82	0.01*	40	0.34	0.02*
Ethical matters	.	.	0.45	0.12	.	.	0.72	0.02*	.	.	0.75	0.02*	30	0.55	0.00*
Results	0.52	0.10	0.70	0.01*	-0.17	0.60	0.79	0.01*	0.81	0.01*	0.62	0.04*	60	0.46	0.00*
Discussion	.	.	0.52	0.10	0.14	0.65	1.00	0.00*	0.43	0.15	.	.	40	0.33	0.02*
Total score %	0.31	0.23	0.78	0.00*	0.18	0.51	0.25	0.35	0.80	0.00*	0.64	0.01*	60	0.59	0.00*

TE, true experimental; QE, quasi-experimental; SS, single system; DEO, descriptive, exploratory, and observational; QL, qualitative; SR, systematic review; n, papers per research design; N, total papers all research designs; ., No data;

Tau (τ) value relationships: Strong ≥ 0.75; Moderate 0.50 to 0.74; Weak 0.25 to 0.49; Little/none < 0.25.

* $p < 0.05$ (2-tailed).

The PEDro scale [7] (true experimental) had gaps for the *Preamble*, *Introduction*, *Ethical matters*, and *Discussion* categories. There were two significant results: a moderate positive correlation for the *Design* ($\tau = 0.68$) and *Sampling* ($\tau = 0.70$)

categories. The Cho and Bero scale [25] had gaps for the *Preamble* and *Data collection* categories for quasi-experimental research. There were four significant results: a perfect positive correlation for the *Introduction* ($\tau = 1.00$) category, a strong positive correlation for the *Sampling* ($\tau = 0.81$) category and total score % ($\tau = 0.78$), and a moderate positive correlation for the *Results* ($\tau = 0.70$) category. In the SCED scale [26] (single system) there were four gaps for the *Preamble*, *Introduction*, *Sampling*, and *Ethical matters* categories. There were no significant correlations for any of the other categories, or for total score %.

The DEO designs also used the Cho and Bero scale and, therefore, had the same gaps for *Preamble* and *Data collection* categories as previously stated. There were four significant results: a perfect positive correlation in the *Introduction* ($\tau = 1.00$) and *Discussion* ($\tau = 1.00$) categories; a strong positive correlation for the *Results* ($\tau = 0.79$) category; and a moderate positive correlation for *Ethical matters* ($\tau = 0.72$) category. The Reis et al scale (qualitative) [27], only had one gap for the *Ethical matters* category. There were four significant strong positive correlations in the *Introduction* ($\tau = 0.74$), *Sampling* ($\tau = 0.76$) and *Results* ($\tau = 0.81$) categories, and for the total score % ($\tau = 0.80$). Finally, for systematic reviews, AMSTAR [28] had two gaps for the *Preamble* and *Discussion* categories. There were four significant results: strong positive correlations for the *Data collection* ($\tau = 0.82$) and *Ethical matters* ($\tau = 0.75$) categories; and moderate positive correlations for the *Results* ($\tau = 0.62$) category and total score % ($\tau = 0.64$).

When each proposed CAT category was examined across all research designs, seven of the eight categories showed significant correlations. Moderate positive correlations were observed in the *Introduction* ($\tau = 0.52$) and *Ethical matters* ($\tau = 0.55$) categories, and total score % ($\tau = 0.59$). There were weak positive correlations in the *Design* ($\tau = 0.43$), *Sampling* ($\tau = 0.40$), *Data collection* ($\tau = 0.34$), *Results* ($\tau = 0.46$), and *Discussion* ($\tau = 0.33$) categories. When all total

score % results were taken into account, there was a significant moderate positive ($\tau = 0.59$) correlation between the proposed CAT and the alternative CATs combined.

5.5 DISCUSSION

The discussion is based on the evaluation of construct validity, as outlined in section 5.2.2. All information was based on the results obtained from the study of CAT design, as described in Chapter 4, and the results of this study. It must be remembered that evaluation of construct validity is an ongoing process. This discussion represents a preliminary evaluation of construct validity based on existing data. Further evaluation of construct validity will occur as more data are gathered in future research, thereby filling gaps in the evidence.

5.5.1 Test content

The proposed CAT evolved from research into the design of critical appraisal tools, research methods, and guidelines for the reporting of research [1]. This approach ensured that important aspects of the construct being examined were included because the tool was based on an amalgam of previous work by experts in the field. Pre-testing of the proposed CAT, and the guide to use and apply the tool aided further refinement, and helped ensure that application of the tool was consistent.

Another important aspect of evidence for test content was whether the proposed CAT exhibited construct underrepresentation or construct-irrelevant variance. In relation to construct underrepresentation, given the inclusive nature of the proposed CAT's design that was based on 45 other CATs, and that each category was developed to be as complete as possible, the possibility of construct underrepresentation should have been reduced. Furthermore, no evidence of

construct underrepresentation was apparent in the current study. However, the proposed CAT may be subject to construct-irrelevant variance due to the same inclusive nature of the design. The *Ethical matters* and *Preamble* categories, for example, had the least amount of support for inclusion, based on the results described in the previous chapter. In this study, the *Ethical matters* category had an average score of 1.5 in all research designs. Given the low overall scores for *Ethical matters* in the papers, this category should remain to demonstrate which research papers adequately apply research ethics. On the other hand, the *Preamble* category had an overall average score of 4.0, so it could be argued that this category should be removed.

However, it may be premature to remove a category at this stage. The *Preamble* category has items that may be important such as *Sufficient detail others could reproduce* (the study) and an abstract that is *Balanced and informative*. Also, these are preliminary results and further research should be undertaken before a final decision is made.

5.5.2 Internal structure

The proposed CAT was designed so that each category consisted of a unidimensional construct and that the categories did not overlap [1]. Furthermore, each category is scored separately based on three principles.

First, scoring was not simply a check list but allowed for a combined objective (tick boxes) and subjective scoring of each category based on the user guide. A scoring system with an objective and subjective component was chosen because previous research, in Chapter 4, outlined that critical appraisal may have aspects of objective and subjective assessment that cannot be reduced to a simple check list [1].

Second, only items applicable to a research design are included in the appraiser's score. In other words, only items that are present and should be present, and items that are absent but should be present, contribute to a category score. There must also be explicit evidence in a paper that an item is present. An item cannot be scored as present based on an assumption. These criteria help to ensure that categories score similarly for different research designs. Other CATs do not have this issue because most are designed to appraise one or a limited number of research designs. The disadvantage of the approach used here is that it requires a reviewer to have a reasonably detailed knowledge of research designs so that they can distinguish which items are applicable.

Third, each category was designed as a separate construct and the categories could only be totalled if this did not impair the precision of score interpretation, especially in relation to obscuring weaknesses in otherwise high scoring papers. The results from the current study seem to suggest that a single total score could be beneficial in interpreting research papers. This is because there are positive correlations between the total scores for the proposed CAT and the alternative CATs in all research designs, except single system, and for all research designs taken together. However, individual category scores must be published along with the total score so that weak scores in a category are not hidden. The total score should be reported as a percentage without decimal places because there are 40 possible, distinct percentage scores, none of which overlap. Each category should be reported as a score out of five and no weighting should be given to any category. In this way, both a score for the unidimensional categories and for the multidimensional total score are available for interpretation.

The recommendation for using the scores is to rank papers based on the total score and then use the scores from the categories, with consideration for the objectives of the appraisal and the characteristics of the papers themselves. The inferences which

can be made are that the higher the total score, augmented by and including the category scores, the higher the credibility of the paper being appraised and the results obtained by that research.

5.5.3 Response process

The proposed CAT aides the response process by the inclusion of tick boxes so that the reviewer can account for elements of the construct that were present, absent but should be present, or not applicable in the papers being appraised. A need for tick boxes was shown in pre-testing the proposed CAT. The tick boxes also help a reviewer to keep track of their appraisal because papers do not necessarily follow the same layout as the proposed CAT. A full version of the proposed CAT should provide space for a reader to include where they found particular evidence within a paper.

Another aspect of response process is whether there were any differences in the interpretation of scores across different research designs. The proposed CAT user guide, attempting to ensure that scores are interpreted similarly in all research designs, explicitly stated that each research design must be scored on its own merits rather than against a predetermined standard against which all research designs should be judged. Furthermore, the *Design*, *Sampling*, *Data collection*, and *Results* categories have at least one item that asks an appraiser whether the paper under review used suitable methods, based on the research question being pursued.

Due to limited data, other features of response process cannot be answered at this stage. These include an overall analysis of individual responses to each category, and whether appraisers are consistent with respect to their application and interpretation of scores. These aspects of the response process will be investigated in the following chapter.

5.5.4 Relations to other variables

Looking at all research designs, there were significant ($p < 0.05$, 2-tailed) weak to moderate correlations in all proposed CAT categories except the *Preamble* (only the qualitative research design had items in the *Preamble* category). There was some degree of heterogeneity across the categories in that no one category showed significant correlations for each research design. This was not surprising given the different approach each alternative CAT had toward scoring papers compared with each other and the proposed CAT. However, given that the scores in all research designs and across all categories (where there was more than one alternative CAT) showed a significant weak to moderate correlation, it can be concluded that the proposed CAT was measuring the same or similar construct in a way that was different to the alternate tools combined.

Single system was the only research design that had no significant correlations. The SCED scale was used because it was the only single system scale available, given the criteria for this study. In retrospect, the result is not surprising given that the SCED scale was developed specifically in the context of papers from the Psychological Database of Brain Impairment Treatment Efficacy (PsycBITE) [26], whereas the proposed CAT was designed for all health research. Therefore, further research on whether score validity for the proposed CAT is applicable to single system research, in particular, should be undertaken where possible.

5.5.5 Consequences of testing

The proposed CAT has only been tested in relation to health research and a sample of papers from 1994–2009, even though no time limit was imposed. Therefore, caution should be taken if the proposed CAT is used outside health research or the stated time period because scores obtained may be significantly biased (either

positively or negatively). Furthermore, the user guide should be followed to ensure that the proposed CAT is applied correctly.

5.5.6 Limitations

One reviewer (the author) was used to collect data for this study. If more reviewers were used, the study may have had greater strength. However, since the purpose was begin an exploration of construct validity and two observations were made for each paper, which is adequate from a testing theory view, this was a good start; given that many other CATs have no validity data available.

Not all aspects of construct validity evaluation were addressed by this or the previous study (Chapter 4). Evidence for a difference between subgroups of users of the proposed CAT needs to be explored in future research in the areas of internal structure and response processes. Also, further evidence on relations to other variables is required for reliability and validity generalisation; that is, to determine whether scores can be generalised to other situations beyond this study.

A possible reason for the bias in publication dates of the sample of papers, despite using a random selection method, may be that from the mid-1990s more journals requested structured abstracts in which authors must include specific information such as the research design [32]. This may also explain why the scores obtained by papers were relatively high: by searching for specific keywords in abstracts certain types of papers were more likely to be returned in the search.

Another area of potential bias was that *Critical Care Medicine* and *JAMA* accounted for 23% of papers. Data were not recorded on the number of papers published by each journal overall or in different years, before the sample was extracted. Also, this cannot be explained by publication frequency because other journals with the same

publication frequencies did not appear as often in the selection. Since the main aim of this study was to begin a process of validity testing and the same paper was appraised by the same reviewer using two different tools, the results should be able to stand. However, the ability to apply results to pre-1990s papers requires further investigation.

A better method of selection to counteract these biases could be to select papers from the whole database, and then fill a quota of papers based on research design and number of papers published in a particular year or range of years compared to the total number of papers published in that year or range of years. However, this process would be very time consuming and many more papers would need to be appraised to achieve the quotas set.

5.6 Conclusion

The benefits of the proposed CAT are that it is relatively simple to implement, can be used in all research designs in health research, and these scores can be compared directly. Other tools which are said to have this capacity have not undergone a validity or reliability testing process.

This preliminary step in the process of validity testing will continue into subsequent studies. Meanwhile, based on the aims of this study and the previous chapter, the proposed CAT has exhibited a good degree of construct validity. This was illustrated through a description and test of the proposed CAT against the theory on which it was built, the collection of empirical evidence for its score validity in relation to alternative CATs, and the stated context for its use. Therefore, sound inferences about research should be possible based on the scores obtained from the proposed CAT. Further research investigating the reliability of the proposed CAT was

undertaken to determine the consistency of the scores obtained. This is the subject of a separate paper [33] and constitutes the next chapter.

5.7 IN SUMMARY

- The unified theory of validity states that construct validity is the only validity.
- The validation process outlined for the proposed CAT uses current techniques.
- Based on theory and objectives, scores obtained using the proposed CAT should be valid.
- The next chapter examines the reliability of scores from the proposed CAT (Objective 5).

5.8 REFERENCES

1. Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

2. Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitc, C., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27). doi:10.3310/hta7270

3. Moyer, A., & Finney, J. W. (2005). Rating methodological quality: Toward improved assessment and investigation. *Accountability in Research, 12*(4), 299-313. doi:10.1080/08989620500440287

4. Armijo Olivo, S., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156-175. doi:10.2522/ptj.20070147

5. Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1060. doi:10.1001/jama.282.11.1054

6. Bialocerkowski, A. E., Grimmer, K. A., Milanese, S. F., & Kumar, S. (2004). Application of current research evidence to clinical physiotherapy practice. *Journal of Allied Health, 33*(4), 230-237.

7. Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

8. Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA, 285*(15), 1992-1995. doi:10.1001/jama.285.15.1992

9. Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., ... Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology, 62*(10), e1-e34. doi:10.1016/j.jclinepi.2009.06.006

10. 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*(6), 349-357. doi:10.1093/intqhc/mzm042

11. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

12. Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford: Oxford University Press.

13. American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), 1-38.

14. Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

15. Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(Monograph Supplement 9), 635-694.

16. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

17. Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25. doi:10.1146/annurev.clinpsy.032408.153639

18. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

19. Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of

randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12. doi:10.1016/0197-2456(95)00134-4

20. Dixon-Woods, M., Bonas, S., Booth, A., Jones, D. R., Miller, T., Sutton, A. J., ... Young, B. (2006). How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research, 6*(1), 27-44. doi:10.1177/1468794106058867

21. Petticrew, M. (2001). Systematic reviews from astronomy to zoology: Myths and misconceptions. *BMJ, 322*(7278), 98-101. doi:10.1136/bmj.322.7278.98

22. Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

23. Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Pearson.

24. Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

25. Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *JAMA, 272*(2), 101-104. doi:10.1001/jama.1994.03520020027007

26. Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18*(4), 385-401. doi:10.1080/09602010802009201

27. Reis, S., Hermoni, D., Van-Raalte, R., Dahan, R., & Borkan, J. M. (2007). Aggregation of qualitative studies - From theory to practice: Patient priorities and family medicine/general practice evaluations. *Patient Education and Counseling, 65*(2), 214-222. doi:10.1016/j.pec.2006.07.011

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

29. Uebersax, J. (2010). Statistical methods for rater and diagnostic agreement. Retrieved 7 September 2010, from http://www.john-uebersax.com/

30. Haadr, M. (2009). Random.org: Random sequence generator. Retrieved 29 January 2011, from http://www.random.org/sequences/

31. Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

32. Ad Hoc Working Group for Critical Appraisal of the Medical Literature. (1987). A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine, 106*(4), 598-604. doi:10.1059/0003-4819-106-4-598

33. Crowe, M., Sheppard, L., & Campbell, A. (under review). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology*.

5.9 ADDITIONAL MATERIAL

5.9.1 User guide for the proposed CAT (evaluation of validity)

Introduction

The critical appraisal tool assumes an awful lot. It assumes that the individual using the tool is familiar with research designs, sampling techniques, ethics, data collection methods, and statistical and non-statistical data analysis techniques. It may be helpful to refer to a general research methods text when appraising papers.

Papers being appraised are unlikely to have the information sought in the sequence outlined in the critical appraisal form. Therefore, it is suggested to read each paper quickly from start to finish to get an overall sense of what is being discussed. Then re-read the paper and fill in the scores.

Scoring

The appraisal form is divided into eight categories and 22 items. An item has multiple parts which describe the item and make it easier to appraise and score a category. Each category receives its own score on a 6 point scale from 0–5. A score of 0 is the lowest score a category can achieve, while a score of 5 is the highest. Only full number (integer) scores are to be awarded, i.e. no fractions.

In the appraisal form, there are tick boxes (☐) beside item descriptors. The tick box is useful to indicate if the item descriptor is:

- Present (☑) – For an item descriptor to be marked as present, there should be evidence of it being present rather than an assumption of presence.
- Absent (☒) – For an item descriptor to be marked as absent, it is implied that it should be present in the first place.
- Not applicable (■) – For an item descriptor to be marked as not applicable, the item descriptor must not be relevant given the characteristics of the paper being appraised and is, therefore, not considered when assigning a score to a category.

Whether an item descriptor is present, absent, or not applicable is further explored in the *Categories and Items* section.

While it may be tempting to add up all the present marks (☑) and all the absent marks (☒) in each category and to use the proportion of one to the other to calculate the score for the category, this is strongly discouraged. It is strongly discouraged because not all item descriptors in a category are of equal importance. For example, in the *Introduction* category there are two items (*Background* and *Objective*) and a total of five tick boxes. If a paper being appraised has all boxes marked as present (☑) except for *Primary objective(s), hypothesis(es), or aim(s)*, should the paper be scored 4/5 for that category? It could be argued that a research paper without a primary objective, hypothesis or aim is fundamentally flawed and, as a result, should be scored 0/5 even though the other four tick boxes were marked as present.

Therefore, the tick marks for present, absent, or not applicable are to be used as a guide to scoring a category rather than as a simple check list. It is up to the appraiser to take into consideration all aspects of each category and, based on both the tick marks and judgement assign a score to the category.

Similarly, the research design used in each paper should be appraised on its own merits and not relative to some preconceived notion of a hierarchy of research designs. What is most important is that the paper used an appropriate research design based on the research question it was addressing, rather than what research design in itself was used.

Finally, it is not the purpose of this tool to present a single score upon which an overall assessment of a paper can be made. Just like not all item descriptors are equal, neither are all categories the same. Categories, and as an extension all scores, are dissimilar, not equivalent, and cannot be added:

1. Each category is designed to be separate from every other category, while items within each category are as similar as possible. As a result, scores from each category are dissimilar.
2. The scores are ordinal or rank-order scales and because categories are dissimilar, a specific category scoring X is not necessarily the same as another category scoring X. That is, scores are not equivalent.
3. As a result of scores being dissimilar and not equivalent, scores cannot be added. For example, if you collected information on a person, such as how they rate a book, a movie, and a night club on a 5-star rating system, it would not make much sense to add these data together. However, the data can still be used to build a picture of the individual. In the same way, it does not make sense to add together the scores for the *Introduction* and *Discussion* categories

or any other combination of categories. However, the data can be used to build up a picture of the paper being appraised.

Categories and items

Preamble

Text

1. Sufficient detail others could reproduce
2. Clear, concise writing/table(s)/diagram(s)/figure(s)
    - These are over-arching concepts and should be present throughout the study.

Title

1. Includes study aims and design
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.

Abstract

1. Key information
2. Balanced and informative
    - This section cannot be completed until the article has been read in full.
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.

Introduction

Background

1. Summary of current knowledge
    - Current and applicable knowledge provides a context for the study.
2. Specific problem(s) addressed and reason(s) for addressing
    - Description of why the study was undertaken.
    - Links current knowledge and stated objective(s), hypothesis(es), or aim(s).

Objective

1. Primary objective(s), hypothesis(es), aim(s)
    - The study must have at least one stated objective, hypothesis, or aim.

2. Secondary question(s)
    - Secondary question(s) may sometimes arise based on the primary objective(s), hypothesis(es), or aim(s).
    - Since this is not always the case, a study without secondary questions should not be penalised.

Design

Research design

1. Research design(s) chosen and why
    - Description of the research design chosen and why it was chosen.
2. Suitability of research design(s)
    - The research design should be congruent with *Background*, *Objective*, *Intervention(s)/treatment(s)/exposure(s)*, and *Outcome(s)/output(s)/ predictor(s)*.

Intervention, Treatment, Exposure

1. Intervention(s)/treatment(s)/exposure(s) chosen and why
    - Where a study does not normally have an intervention/treatment/exposure, it should not be penalised when none is present.
    - Statement for every intervention/treatment/exposure chosen and why it was chosen.
    - Each intervention/treatment/exposure must be congruent with *Background*, *Objective*, and *Research design*.
2. Precise details of the intervention(s)/treatment(s)/exposure(s) for each group
    - Full details are presented for every intervention/treatment/exposure for every participant/case/group so that other studies could duplicate.
3. Intervention(s)/treatment(s)/exposure(s) valid and reliable
    - A statement of reliability/validation or why there is no validation/reliability for each intervention/treatment/exposure.

Outcome, Output, Predictor, Measure

1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen and why
    - All research has at least one expected outcome/output/predictor/measure.
    - Statement for each outcome/output/predictor/measure chosen and why it was chosen.
    - Each outcome/output/predictor/measure must be congruent with *Background*, *Objective*, *Research design*, and *Intervention/treatment/ exposure*.

2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s)
    - Full details are presented of every expected outcome/output/predictor/measure for every participant/case/group so that other studies could duplicate.
3. Outcome(s)/output(s)/predictor(s)/measure(s) valid and reliable
    - A statement of reliability/validation or why there is no validation/reliability for each outcome/output/predictor/measure.

Note In some cases the *Outcome(s)/output(s)/predictor(s)/measure(s)* may be similar to or the same as the *Objective(s), hypothesis(es), aim(s)*. However, in most cases to achieve the *Objective(s), hypothesis(es), aim(s)* a series of *Outcome(s)/output(s)/predictor(s)/measure(s)* are required.

Bias, etc.

1. Potential sources of bias, confounding variables, effect modifiers, interactions
    - Identification of potential sources of:
        - Bias – e.g. attrition, detection, experimental, information, interview, observation, performance, rater, recall, selection.
        - Confounding variables or factors – A variable which interferes between the intervention/treatment/exposure and the outcome/output/predictor/measure.
        - Effect modification – A variable which modifies the association between the intervention/treatment/exposure and the outcome/output/predictor/measure.
        - Interaction effects – When various combinations of intervention(s)/treatment(s)/exposure(s) cause different outcome(s)/output(s)/predictor(s)/measure(s).
    - Should be identified, as far as possible, within the *Research design* before data collection begins in order to minimise their effect.
    - See also *Sampling* and *Data collection*.
2. Sequence generation, group allocation, group balance, and by whom
    - In studies where participants/cases are allocated to groups, the methods used should be stated and procedures established before recruitment or data collection begins (e.g. blinding, method used to randomise, allocate to or balance groups).
3. Equivalent treatment of participants/cases/groups
    - Each participant/case/group must be treated equivalently apart from any intervention/treatment/exposure.

- If participants/cases/groups are not treated equivalently a statement regarding why this was not possible, how this may affect results, and procedures in place for managing participants/cases/groups.
- See also *Sampling protocol*, *Collection protocol*, and *Participant ethics*.

Sampling

Sampling method

1. Sampling method(s) chosen and why
   - Description of the sampling method chosen and why it was chosen.
   - Sampling methods are normally probability or non-probability based.
   - Examples include: Simple random, systematic, stratified, cluster, convenience, representative, purposive, snowball, and theoretical.
   - Also included here is the search strategy used for a systematic review (e.g. databases searched, search terms).
2. Suitability of sampling method
   - The sampling method should be decided and in place before recruitment or data collection begins.
   - The sampling method should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc.*

Sample size

1. Sample size, how chosen, and why
   - Description of the sample size, the method of sample size calculation, and why that method was chosen.
   - Sample size calculations are normally probability or non-probability based.
   - Examples of how calculations can be made include: Accuracy [e.g. confidence interval (α), population or sample variance (s^2, σ^2), effect size or index (ES, d), power (1-β)], analysis, population, redundancy, saturation, and budget.
2. Suitability of sample size
   - The sample size or estimate of sample size, with contingencies, should be described and calculated before recruitment/data collection begins.
   - The sample size should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc.*

Note Sample size calculations are not required for systematic reviews, because it is not possible to know the number of papers that will meet the selection criteria, or for *some* single system designs.

Sampling protocol

1. Description and suitability of target/actual/sample population(s)
   - The target/actual/sample population(s) should be described.
   - The target/actual/sample population(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc.*
2. Inclusion and exclusion criteria for participants/cases/groups
   - Inclusion and exclusion criteria should be explicitly stated and established before recruitment/data collection begins.
   - The use of inclusion and exclusion criteria (especially exclusion criteria) should not be used in such a way as to bias the sample.
3. Recruitment of participants/cases/groups
   - Description of procedures for recruitment and contingencies put in place.
   - Recruitment should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Bias etc.*, and other aspects of *Sampling*.
   - See also *Participant ethics*, *Researcher ethics*, and *Collection protocol*.

Note For systematic reviews inclusion and exclusion criteria *only* need to be appraised, because they refer to the parameters used to select papers.

Data collection

Collection method

1. Collection method(s) chosen and why
   - Description of the method(s) used to collect data and why each was chosen.
   - In systematic reviews, this refers to how information was extracted from papers, because these are the data collected.
2. Suitability of collection method(s)
   - The data collection method(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, *Bias etc.*, and *Sampling*.

Collection protocol

1. Include date(s), location(s), setting(s), personnel, materials, processes
    - Description of and details regarding exactly how data were collected, especially any factor(s) which may affect *Outcome/output/predictor/measure* or *Bias etc.*
2. Method(s) to ensure/enhance quality of measurement/instrumentation
    - Description of any method(s) used to enhance or ensure the quality of data collected (e.g. pilot study, instrument calibration, standardised test(s), independent/multiple measurement, valid/reliable tools).
    - Also includes any method(s) which reduce or eliminate bias, confounding variables, effect modifiers, interactions which are not an integral part of the *Design* category (e.g. blinding of participants, intervention(s), outcome(s), analysis; protocols and procedures implemented).
    - In qualitative studies, this relates to concepts such as trustworthiness, authenticity, and credibility.
    - See also *Bias etc.*
3. Manage non-participation, withdrawal, incomplete/lost data
    - Description of any method(s) used to manage or prevent non-participation, withdrawal, or incomplete/lost data.
    - These include but are not limited to: Intention to treat analysis (ITT); last observation carried forward (LOCF); follow up (FU), e.g. equal length, adequate or complete; and completer analysis, e.g. on-treatment, on-protocol.

Ethical matters

Note Some studies may have been conducted before *Ethical matters* were a major point of concern. The research ethics standards of the time may need to be taken into consideration rather than the prevailing standards.

Participant ethics

1. Informed consent, equity
    - All participants must have provided their informed consent.
    - Equity includes, but is not limited to, cultural respect, just and equitable actions, no harm to participants, debriefing, and consideration for vulnerable individuals or groups.
2. Privacy, confidentiality/anonymity
    - The privacy and confidentiality and/or anonymity of participants must be catered for.

- If this is not possible, the informed and written consent of individuals affected must be obtained.

Researcher ethics

1. Ethical approval, funding, conflict(s) of interest
   - A statement of ethical approval from recognised Ethics Committee(s) or Board(s) suitable for the study being undertaken.
   - Any real, perceived, or potential conflict(s) of interest should be stated.
   - All sources of funding should be stated.
2. Subjectivities, relationship(s) with participants/cases
   - Description of how the researcher(s) could have potentially or did affect the outcomes of the study through their presence or behaviour.
   - Includes a description of procedures used to minimise this occurring.
   - See also *Bias etc.*

Results

Analysis, Integration, Interpretation method

1. A.I.I. (Analysis/Integration/Interpretation) method(s) for primary outcome(s)/output(s)/predictor(s) chosen and why
   - Description of statistical and non-statistical method(s) used to analyse/integrate/interpret Outcome(s)/output(s)/predictor(s)/measure(s) and why each was chosen.
2. Additional A.I.I. methods (e.g. subgroup analysis) chosen and why
   - Description of additional statistical and non-statistical method(s) used to analyse/integrate/interpret Outcome(s)/output(s)/predictor(s)/measure(s) and why each was chosen.
3. Suitability of analysis/integration/interpretation method(s)
   - The analysis/integration/interpretation method(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor*, *Bias etc.*, *Sampling*, and *Data collection*.

Essential analysis

1. Flow of participants/cases/groups through each stage of research
   - Description of how participants/cases/groups advanced through the study.
   - Explanation of course of intervention/treatment/exposure.
2. Demographic and other characteristics of participants/cases/groups
   - Description of baseline characteristics of participants/cases/groups so this can be integrated into the analysis.

3. Analyse raw data, response rate, non-participation, withdrawal, incomplete/ lost data
    - Unadjusted data should be analysed.
    - There may be differences between those that completed and those that did not complete the study.

Outcome, Output, Predictor analysis

1. Summary of results and precision for each outcome/output/predictor/measure
    - Results summarised with, where possible, an indicator of the precision and effect size of each result for each outcome/output/predictor/measure.
    - Where data are adjusted, make clear what was adjusted and why.
    - Where data are categorised, report of internal and external boundaries.
    - Use of quotations to illustrate themes/findings, privileging of subject meaning, adequate description of findings, evidence of reflexivity.
2. Consideration of benefits/harms, unexpected results, problems/failures
    - Description of all outcomes, not just ones being looked for.
    - Description of differences between planned and actual implementation, and the potential effect on results.
3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes)
    - Exploration of outliners because they may not be anomalous.

Discussion

Interpretation

1. Interpretation of results in the context of current evidence and objectives
    - Summarises key results in relation to *Background* and *Objective*.
    - Compare and contrast other research findings.
2. Draw inferences consistent with the strength of the data
    - Do not over or under represent data.
    - Draw inferences based on the entirety of available evidence.
    - See also *Sampling* and *Data collection*.
3. Consideration of alternative explanations for observed results
    - Exploration of reasons for differences between observed and expected.
    - Determines if other factors may lead to similar results.
4. Account for bias, confounding, interactions, effect modifiers, imprecision
    - Discussion on magnitude and direction of *Bias etc.* and how this may have affected the results.
    - See also *Essential analysis*.

Generalisation

1. Consideration of overall practical usefulness of the study
   - Discussion on practical vs. theoretical usefulness.
2. Description of generalisability (external validity) of the study
   - Dependent on *Design*, *Sampling*, and *Data collection*.

Concluding remarks

1. Highlight study's particular strengths
   - What did the study do well?
2. Suggest steps that may improve future results (e.g. limitations)
   - How could the study have been better?
3. Suggest further studies
   - Where should the next study begin?

5.9.2 Alternative critical appraisal tools

PEDro (1999) True experimental

Design /6			**Where?**
2. Subjects were randomly allocated to groups (in a crossover study, subjects were randomly allocated an order in which treatments were received)	1	0	
3. Allocation was concealed	1	0	
4. The groups were similar at baseline regarding the most important prognostic indicators	1	0	
5. There was blinding of all subjects	1	0	
6. There was blinding of all therapists who administered the therapy	1	0	
7. There was blinding of all assessors who measured at least one key outcome	1	0	
Sampling /1			**Where?**
1. Eligibility criteria were specified.	1	0	
Data collection /2			
8. Measures of at least one key outcome were obtained from more than 85% of the subjects initially allocated to groups	1	0	
9. All subjects for whom outcome measures were available received the treatment or control condition as allocated or, where this was not the case, data for at least one key outcome was analysed by "intention to treat"	1	0	
Results /2			
10. The results of between-group statistical comparisons are reported for at least one key outcome	1	0	
11. The study provides both point measures and measures of variability for at least one key outcome	1	0	

Total	**/11**	**%**	

Scoring: 1 – Criterion clearly satisfied; 0 – Not clearly satisfied
Missing: Preamble; Introduction; Method *Ethical matters*; Discussion

Cho and Bero (1994) Quasi-experimental and DEO

Introduction / [0-2]				
2. What was the study question?				
3. Was the study questions sufficiently described?	2	1	0	n/a
Design / [9-25]				
1. Study design (choose 1 only) *Experimental, randomised:* Placebo-controlled trial | Comparative trial, no placebo | Time series trial ..... (5 points) *Experimental, un-randomised:* Placebo-controlled trial | Comparative trial, no placebo | Time series trial | ..... (4 points) Crossover trial | Natural experiment *Non-experimental:* Cohort, prospective | Cohort, retrospective | Case-control ..... (3 points) Cross-sectional ..... (2 points) Case reports or case series ..... (1 point) *None of the above (describe below):* ..... (0 points)				
4. Was the study design appropriate to answer the study question?	2	1	0	n/a
1c. Were the ~~therapeutic~~ outcomes measured in the study important?	2	1	0	
3c. Was the comparison group clinically meaningful?	2	1	0	
8. Were control subjects appropriate? (If no controls were used, check No)	2	1	0	n/a
11. If subjects were randomly allocated to treatment groups, was the method of random allocation sufficiently described? (If not randomly allocated, check n/a)	2	1	0	n/a
12. If blinding of investigators to intervention was possible, was it reported? (If not possible, check n/a)	2	1	0	n/a
13. If blinding of subjects to intervention was possible, was it reported? (If not possible, n/a)	2	1	0	n/a
14. Was measurement bias accounted for by methods other than blinding?	2	1	0	n/a
15. Were known confounders accounted for by study design? (If no known, n/a)	2	1	0	n/a
16. Were known confounders accounted for by analysis? (If no known, n/a)	2	1	0	n/a
Sampling / [0-10]				
17. Was there a sample size justification before the study?	2	1	0	n/a
5. Were both inclusion and exclusion criteria specified? (If case study, n/a)	2	1	0	n/a
7. Were subjects appropriate to the study question?	2	1	0	n/a
2c. Were the subjects of the study representative ~~of patients who would actually use the drug~~?	2	1	0	
9. Were subjects randomly selected from the target population?	2	1	0	n/a
10. If subjects randomly selected, was the method sufficiently well described? (If subjects were not randomly selected, n/a)	2	1	0	n/a
Ethical matters /4				
6c. Was approval from an institutional review board explicitly reported?	2	1	0	
7c. As far as could be determined from the article, was the study ethical?	2	1	0	
Results / [4-18]				
20. Were the statistical tests stated?	2	1	0	n/a
19. Were statistical analyses appropriate?	2	1	0	n/a
6. For case studies only: Were patient characteristics adequately reported? (If not case study, check n/a)	2	1	0	n/a
22. Were attrition of subjects and reason for attrition recorded?	2	1	0	n/a
18. Were post hoc power calculations or confidence intervals reported for statistically non-significant results?	2	1	0	n/a
21. Were exact P values or confidence intervals reported for each test?	2	1	0	n/a
23. For those subjects who completed the study, were results completely reported?	2	1	0	n/a
4c. Was the treatment effect clinically meaningful?	2	1	0	n/a
5c. Were side effects adequately measured?	2	1	0	n/a
Discussion /2				
24. Do the findings support the conclusions?	2	1	0	n/a

Total	/	**%**

Scoring: 2 – Yes; 1 – Partial; 0 – No
Missing: Preamble; Method *Data collection*

Tate et al (2008) Single system

Design /3			**Where?**
2. Target outcome The paper identifies a precise, repeatable, and operationally defined target outcome that can be used to measure treatment success.	1	0	
3. Design The study design allows for the examination of cause and effect relationships to demonstrate treatment efficacy.	1	0	
8. Independence of assessors To reduce assessment bias by employing a person who is otherwise uninvolved in the study, to provide an evaluation of the patients.	1	0	
Data collection /3			
4. Baseline To establish that sufficient observations have occurred during the pre-treatment period to provide an adequate baseline measure.	1	0	
5. Sampling outcome during treatment To establish that sufficient observation during the treatment phase has occurred to differentiate a treatment response from fluctuations that may occur at baseline.	1	0	
7. Inter-rater reliability To determine if the target outcome measure is reliable and collected in a consistent manner.	1	0	
Results /3			
1. Clinical history The study provides critical information regarding demographic and other characteristics of the research subject that allows the reader to determine the applicability of the treatment to another individual.	1	0	
6. Raw data record To provide an accurate representation of the variability of the target outcome.	1	0	
9. Statistical analysis To demonstrate the effectiveness of the treatment of interest by statistically comparing the results over the study phases.	1	0	
Discussion /2			
10. Replication To demonstrate that the application and results of the therapy are not limited to a specific individual or situation (i.e. that the results are reproduced in other circumstances – replicated across subjects, clinicians, or settings).	1	0	
11. Generalisation To demonstrate the functional utility of the treatment in extending beyond the target outcome or clinical environment into other areas of the individual's life.	1	0	

Total	**/11**	**%**	

Scoring: 1 – Explicit evidence; 0 – Not clearly satisfied
Missing: Preamble; Introduction; Method *Sampling*; Method *Ethical matters*

Reis et al (2007) Qualitative

Preamble **/3**					
15. Global *rating*: (Do this last)	3	2	1	0	u/r
Introduction **/3**					
1. Clarity and transparency: Are the *goals* clearly *described*?	3	2	1	0	u/r
Design **/3**					
11. *Rigor*: Does the research appear to have undertaken sufficient care, depth, and meticulousness in the *research process*?	3	2	1	0	u/r
Sampling **/3**					
7. Evidence of *relevance* of sampling method: Does the sample produce the type of knowledge necessary to understand the structures and processes within which the individuals or situations are located?	3	2	1	0	u/r
Data collection **/12**					
2. Clarity and transparency: Is the *data collection technique* clearly *described*?	3	2	1	0	u/r
3. Appropriateness of *data collection* method: Are the data collection methods used *appropriate* to the subject matter?	3	2	1	0	u/r
9. *Data adequacy*: Does the time, extent, and nature, of the researcher's involvement appear to be adequate to the subject studied?	3	2	1	0	u/r
13. *Relevance*: To the subject matter of the (systematic) review.	3	2	1	0	u/r
Results **/12**					
4. Clarity and transparency: Is the *data analysis technique* clearly *described*?	3	2	1	0	u/r
6. Privileging of *subject meaning*: Does the study illuminate the meanings, actions and context of those researched and illustrate a sense that the investigator successfully resonated with the subject matter?	3	2	1	0	u/r
8. Adequate *description of finding*: Is the description provided in enough detail and depth to allow interpretation of the meanings and connect of what is being studied?	3	2	1	0	u/r
12. *Reflexivity*: Evidence of reflexiveness in the process	3	2	1	0	u/r
Discussion **/6**					
10. *Theoretical and conceptual coherence and plausibility*: Does the research move logically from description of the data, through quotations or examples, to an analysis and interpretation of the meanings and their significance?	3	2	1	0	u/r
14. *Generalisability*: To the ~~European~~ context.	3	2	1	0	u/r

Total	**/45**	**%**

Scoring: 3 – High (Clearly); 2 – Moderate (Moderately); 1 – Low (Barely); 0 – No (None); u/r – Unable to rate
Missing: Method *Ethical matters*

Shea (2007) Systematic review

Introduction **/1**				
1. Was an 'a priori' design provided? The research question and inclusion criteria should be established before the conduct of the review.	Y	N	CA	NA
Design **/1**				
10. Was the likelihood of publication bias assessed? An assessment of publication bias should include a combination of graphical aids (e.g. funnel plot, other available tests) and/or statistical tests (e.g. Egger regression test).	Y	N	CA	NA
Sampling **/2**				
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion? The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.	Y	N	CA	NA
5. Was a list of studies (included and excluded) provided? A list of included and excluded studies should be provided.	Y	N	CA	NA
Data collection **/4**				
2. Was there duplicate study selection and data extraction? There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.	Y	N	CA	NA
3. Was a comprehensive literature search performed? At least two electronic sources should be searched. The report must include years and databases used (e.g. Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialised registers, or experts in the particular field of study, and by reviewing the references in the studies found.	Y	N	CA	NA
7. Was the scientific quality of the included studies assessed and documented? 'A priori' methods of assessment should be provided (e.g. for effectiveness studies if the author(s) chose to include only randomised, double-blind, placebo controlled studies, or allocation concealment as inclusion criteria); for other types of studies alternative items will be relevant.	Y	N	CA	NA
8. Was the scientific quality of the included studies used appropriately in formulating conclusions? The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations.	Y	N	CA	NA
Ethical matters **/1**				
11. Was the conflict of interest stated? Potential sources of support should be clearly acknowledged in both the systematic review and the included studies.	Y	N	CA	NA
Results **/2**				
6. Were the characteristics of the included studies provided? In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions and outcomes. The ranges of characteristics in all the studies analysed e.g. age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.	Y	N	CA	NA
9. Were the methods used to combine the findings of studies appropriate? For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e. Chi-squared test for homogeneity, I^2). If heterogeneity exists a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e. is it sensible to combine?).	Y	N	CA	NA

Total	**/**	**%**

Scoring: Y – Yes; N – No; CA – Can't Answer; NA – Not Applicable
Missing: Preamble; Discussion

5.9.3 Worksheet function and decision table

Microsoft Excel worksheet function

```
'Streiner, Norman (2008) Health measurement scales: a practical
guide to their development and use. Oxford: Oxford University Press
(pp. 198-202). Sample size based on standard error.
Function rSampleSizeSE (Observations As Integer, Reliability As
     Single, CIwidth As Single)

   Dim Zr, Zr1, SE, Top, Bottom As Double

   'Make sure inputs are in the correct range
   If Observations <= 1 Then
     rSampleSizeSE = "Observations must be integer, >=2"
   ElseIf Reliability >= 1 Or Reliability <= 0 Then
     rSampleSizeSE = "Est reliability must be >0 and <1"
   ElseIf CIwidth >= 1 Or CIwidth <= 0 Then
     rSampleSizeSE = "Confidence Interval must be >0 and <1"
   Else
     'Calculation
     Zr = 0.5 * (Log(((1 + ((Observations - 1) * Reliability))) / (1
          - Reliability)))
     Zr1 = 0.5 * (Log(((1 + ((Observations - 1) * (Reliability -
          (CIwidth / 2))))) / (1 - (Reliability - (CIwidth / 2)))))
     SE = Zr - Zr1
     Top = Observations
     Bottom = 2 * (Observations - 1) * ((SE) ^ 2)
     'Round up [WorksheetFunction.RoundUp()] final answer
          [2+(Top/Bottom] to the nearest integer [,0]
     rSampleSizeSE = WorksheetFunction.RoundUp(2 + (Top / Bottom), 0)
   End If
End Function
```

Decision table

	CI around *r*	0.19				
n	**Estimate of *r***					
	0.95	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**
2	6	10	17	24	32	40
3	5	8	12	17	22	27
4	5	7	11	15	19	23
5	4	7	10	14	17	20
6	4	7	10	13	16	19
7	4	7	9	13	16	18
8	4	7	9	12	15	18
9	4	6	9	12	15	17
10	4	6	9	12	15	17

n, number of observations

5.9.4 List of papers used for evaluation of validity

Pilot Study

Arthur, H., Smith, K., Kodis, J., & McKelvie, R. (2002). A controlled trial of hospital versus home-based exercise in cardiac patients. *Medicine & Science in Sports & Exercise*, *34*(10), 1544-1550.

Heetveld, M. J., Raaymakers, E. L. F. B., Luitse, J. S. K., Nijhof, M., & Gouma, D. J. (2007). Femoral neck fractures: Can physiologic status determine treatment choice? *Clinical Orthopaedics & Related Research*, *461*, 203-212.

McGillis Hall, L., Doran, D., & Pink, L. (2008). Outcomes of interventions to improve hospital nursing work environments. *Journal of Nursing Administration*, *38*(1), 40-46.

Peeters, M. G., Verhagen, A., de Bie, R. A., & Oostendorp, B. R. (2001). The efficacy of conservative treatment in patients with whiplash injury: A systematic review of clinical trials. *Spine*, *26*(4), E64-E73.

Presti, C., Puech-Leao, P., & Albers, M. (1999). Superficial femoral eversion endarterectomy combined with a vein segment as a composite artery-vein bypass graft for infrainguinal arterial reconstruction. *Journal of Vascular Surgery*, *29*(3), 413-421.

Whitney, C. M. (2004). Maintaining the square: How older adults with Parkinson's Disease sustain quality in their lives. *Journal of Gerontological Nursing*, *30*(1), 28-35.

Main Study

True experimental

Als, H., Lawhon, G., Duffy, F. H., McAnulty, G. B., Gibes-Grossman, R., & Blickman, J. G. (1994). Individualized developmental care for the very low-birth-weight preterm infant: Medical and neurofunctional effects. *JAMA*, *272*(11), 853-858.

Arts, M. P., Brand, R., van den Akker, E. M., Koes, B. W., Bartels, R. H., & Peul, W. C. (2009). Tubular diskectomy vs conventional microdiskectomy for sciatica: A randomized controlled trial. *JAMA*, *302*(2), 149-158.

Dembinski, R., Hochhausen, N., Terbeck, S., Uhlig, S., Dassow, C., Schneider, M., Schachtrupp, A., et al. (2007). Pumpless extracorporeal lung assist for protective mechanical ventilation in experimental lung injury. *Critical Care Medicine*, *35*(10), 2359-2366.

Dore, S., Buchan, D., Coulas, S., Hamber, L., Stewart, M., Cowan, D., & Jamieson, L. (1998). Alcohol versus natural drying for newborn cord care. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, *27*(6), 621-627.

van Gils, E. J. M., Veenhoven, R. H., Hak, E., Rodenburg, G. D., Bogaert, D., IJzerman, E. P., Bruin, J., et al. (2009). Effect of reduced-dose schedules with 7-valent pneumococcal conjugate vaccine on nasopharyngeal pneumococcal carriage in children: A randomized controlled trial. *JAMA*, *302*(2), 159-167.

Iwama, H., Ohmizo, H., Furuta, S., Ohmori, S., Watanabe, K., Kaneko, T., & Tsutsumi, K. (2002). Inspired superoxide anions attenuate blood lactate concentrations in postoperative patients. *Critical Care Medicine*, *30*(6), 1246-1249.

Lobo, S. M., Salgado, P. F., Castillo, V. G., Borim, A. A., Polachini, C. A., Palchetti, J. C., Brienzi, S. L., et al. (2000). Effects of maximizing oxygen delivery on morbidity and mortality in high-risk surgical patients. *Critical Care Medicine*, *28*(10), 3396-3404.

Mercat, A., Richard, J. M., Vielle, B., Jaber, S., Osman, D., Diehl, J., Lefrant, J., et al. (2008). Positive end-expiratory pressure setting in adults with acute lung injury and acute respiratory distress syndrome: A randomized controlled trial. *JAMA*, *299*(6), 646-655.

Rubertsson, S., Grenvik, A., Zemgulis, V., & Wiklund, L. (1995). Systemic perfusion pressure and blood flow before and after administration of epinephrine during experimental cardiopulmonary resuscitation. *Critical Care Medicine*, *23*(12), 1984-1996.

The TADS Team. (2007). The treatment for adolescents with depression study (TADS): Long-term effectiveness and safety outcomes. *Archives of General Psychiatry*, *64*(10), 1132-1143.

Quasi-experimental

Allen, H. M., Borden, S., Pikelny, D. B., Paralkar, S., Slavin, T., & Bunn, W. B. (2003). An intervention to promote appropriate management of allergies in a heavy manufacturing workforce: Evaluating health and productivity outcomes. *Journal of Occupational and Environmental Medicine*, *45*(9), 956-972.

Bergman-Evans, B. (2004). Beyond the basics: Effects of the eden alternative model on quality of life issues. *Journal of Gerontological Nursing*, *30*(6), 27-34.

Coughlin, T. A., & Long, S. K. (2000). Effects of Medicaid managed care on adults. *Medical Care*, *38*(4), 433-446.

Jablonski, R., Reed, D., & Maas, M. (2005). Care intervention for older adults with Alzheimer's disease and related dementias: effect of family involvement on cognitive and functional outcomes in nursing homes. *Journal of Gerontological Nursing*, *31*(6), 38-48.

Kaunonen, M. P., Tarkka, M. P., Laippala, P., & Paunonen-Ilmonen, M. P. (2000). The impact of supportive telephone call intervention on grief after the death of a family member. *Cancer Nursing*, *23*(6), 483-491.

Liu, L., Li, C., Tang, S. T., Huang, C., & Chiou, A. (2006). Role of continuing supportive cares in increasing social support and reducing perceived uncertainty among women with newly diagnosed breast cancer in Taiwan. *Cancer Nursing*, *29*(4), 273-282.

Mentes, J., & Culp, K. (2003). Reducing hydration-linked events in nursing home residents. *Clinical Nursing Research*, *12*(3), 210-225.

Mignone, J., & Guidotti, T. L. (1999). Support groups for injured workers: Process and outcomes. *Journal of Occupational and Environmental Medicine*, *41*(12), 1059-1064.

Motheral, B., & Fairman, K. A. (2001). Effect of a three-tier prescription copay on pharmaceutical and other medical utilization. *Medical Care*, *39*(12), 1293-1304.

Polanczyk, G., Zeni, C., Genro, J. P., Guimaraes, A. P., Roman, T., Hutz, M. H., & Rohde, L. A. (2007). Association of the adrenergic α2A receptor gene with methylphenidate improvement of inattentive symptoms in children and adolescents with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *64*(2), 218-224.

Single system

Behari, S., Nayak, S. R., Bhargava, V., Banerji, D., Chhabra, D. K., & Jain, V. K. (2003). Craniocervical tuberculosis: Protocol of surgical management. *Neurosurgery*, *52*(1), 72-81.

Chatellier, G., Day, M., Bobrie, G., & Menard, J. (1995). Feasibility study of N-of-1 trials with blood pressure self-monitoring in hypertension. *Hypertension*, *25*(2), 294-301.

Durham, S. R., McComb, J. G., & Levy, M. L. (2003). Correction of large (>25 cm2) cranial defects with "reinforced" hydroxyapatite cement: Technique and complications. *Neurosurgery*, *52*(4), 842-845.

Gosain, A. K., Santoro, T. D., Havlik, R. J., Cohen, S. R., & Holmes, R. E. (2002). Midface distraction following Le Fort III and Monobloc osteotomies: Problems and solutions. *Plastic & Reconstructive Surgery*, *109*(6), 1797-1808.

Jais, P., Haissaguerre, M., Shah, D. C., Chouairi, S., Gencel, L., Hocini, M., & Clementy, J. (1997). A focal source of atrial fibrillation treated by discrete radiofrequency ablation. *Circulation*, *95*(3), 572-576.

King, W. A., Wackym, P. A., Sen, C., Meyer, G. A., Shiau, J., & Deutsch, H. (2001). Adjunctive use of endoscopy during posterior fossa surgery to treat cranial neuropathies. *Neurosurgery*, *49*(1), 108-116.

Lundell, J. C., Silverman, D. G., Brull, S. J., O'Connor, T. Z., Kitahata, L. M., Collins, J. G., & LaMotte, R. (1996). Reduction of postburn hyperalgesia after local injection of ketorolac in healthy volunteers. *Anesthesiology*, *84*(3), 502-509.

Rao, N., & Regalla, D. M. (2006). Uncertain efficacy of daptomycin for prosthetic joint infections: A prospective case series. *Clinical Orthopaedics & Related Research*, *451*, 34-37. doi:10.1097/01.blo.0000224021.73163.61

Schluter, W., Judson, F., Baro'n, A., McGill, W., Marine, W., & Douglas, J. (1996). Usefulness of Human Immunodeficiency Virus post-test counseling by telephone for low-risk clients of an urban sexually transmitted diseases clinic. *Sexually Transmitted Diseases*, *23*(3), 190-197.

Smith, A., Lew, R., Shrimpton, C., Evans, R., & Abbenante, G. (2000). A novel stable inhibitor of endopeptidases EC 3.4.24.15 and 3.4.24.16 potentiates bradykinin-induced hypotension. *Hypertension*, *35*(2), 626-630.

Descriptive, exploratory or observational

Alexander, J. M., McIntire, D. M., & Leveno, K. J. (1999). Chorioamnionitis and the prognosis for term infants. *Obstetrics & Gynecology*, *94*(2), 274-278.

Bart, C., & Tabone, J. (1999). Mission statement content and hospital performance in the Canadian not-for-profit health care sector. *Health Care Management Review*, *24*(3), 18-29.

Bhattacharyya, N., & Fried, M. P. (2003). The accuracy of computed tomography in the diagnosis of chronic rhinosinusitis. *Laryngoscope*, *113*(1), 125-129.

Cournot, M., Marquie, J. C., Ansiau, D., Martinaud, C., Fonds, H., Ferrieres, J., & Ruidavets, J. B. (2006). Relation between body mass index and cognitive function in healthy middle-aged men and women. *Neurology*, *67*(7), 1208-1214.

Daly, R. M., Caine, D., Bass, S. L., Pieter, W., & Broekhoff, J. (2005). Growth of highly versus moderately trained competitive female artistic gymnasts. *Medicine & Science in Sports & Exercise*, *37*(6), 1053-1060.

Despriet, D. D., Klaver, C. C., Witteman, J. C., Bergen, A. A., Kardys, I., de Maat, M. P., Boekhoorn, S. S., et al. (2006). Complement factor H polymorphism, complement activators, and risk of age-related macular degeneration. *JAMA*, *296*(3), 301-309.

Jiang, R., Manson, J. E., Stampfer, M. J., Liu, S., Willett, W. C., & Hu, F. B. (2002). Nut and peanut butter consumption and risk of Type 2 diabetes in women. *JAMA*, *288*(20), 2554-2560.

Menges, T., Konig, I. R., Hossain, H., Little, S., Tchatalbachev, S., Thierer, F., Hackstein, H., et al. (2008). Sepsis syndrome and death in trauma patients are associated with variation in the gene encoding tumor necrosis factor. *Critical Care Medicine*, *36*(5), 1456-1462, Suppl. e1-e6.

Reid, S. A., Speedy, D. B., Thompson, J. M., Noakes, T. D., Mulligan, G., Page, T., Campbell, R. G., et al. (2004). Study of hematological and biochemical parameters in runners completing a standard marathon. *Clinical Journal of Sport Medicine*, *14*(6), 344-353.

Whalen, C. C., Nsubuga, P., Okwera, A., Johnson, J. L., Hom, D. L., Michael, N. L., Mugerwa, R. D., et al. (2000). Impact of pulmonary tuberculosis on survival of HIV-infected adults: a prospective epidemiologic study in Uganda. *AIDS*, *14*(9), 1219-1228.

Qualitative

Appelin, G., & Bertero, C. (2004). Patients' experiences of palliative care in the home: A phenomenological study of a Swedish sample. *Cancer Nursing*, *27*(1), 65-70.

Averitt, S. S. (2003). "Homelessness is not a choice!!" The plight of homeless women with preschool children living in temporary shelters. *Journal of Family Nursing*, *9*(1), 79-100.

Beck, C. (1996). Postpartum depressed mothers' experiences interacting with their children. *Nursing Research*, *45*(2), 98-104.

Benisovich, S., & King, A. C. (2003). Meaning and knowledge of health among older adult immigrants from Russia: A phenomenological study. *Health Education Research*, *18*(2), 135-144.

Eifried, S. (2003). Bearing witness to suffering: The lived experience of nursing students. *Journal of Nursing Education*, *42*(2), 59-67.

Goldbort, J. G. (2009). Women's lived experience of their unexpected birthing process. *MCN: The American Journal of Maternal Child Nursing*, *34*(1), 57-62.

Hinck, S. M. (2007). The meaning of time in oldest-old age. *Holistic Nursing Practice*, *21*(1), 35-41.

Marcinkowski, K., Wong, V., & Dignam, D. (2005). Getting back to the future: A grounded theory study of the patient perspective of total knee joint arthroplasty. *Orthopaedic Nursing*, *24*(3), 202-209.

Meighan, M., Davis, M., Thomas, S., & Droppleman, P. (1999). Living with postpartum depression: The father's experience. *MCN: The American Journal of Maternal Child Nursing*, *24*(4), 202-208.

Taleghani, F., Yekta, Z. P., Nasrabadi, A. N., & Kappeli, S. (2008). Adjustment process in Iranian women with breast cancer. *Cancer Nursing*, *31*(3), 32-41.

Systematic review

Bagshaw, S. M., Berthiaume, L. R., Delaney, A., & Bellomo, R. (2008). Continuous versus intermittent renal replacement therapy for critically ill patients with acute kidney injury: A meta-analysis. *Critical Care Medicine*, *36*(2), 610-617.

Benatar, M., & Kaminski, H. J. (2007). Evidence report: The medical treatment of ocular myasthenia (an evidence-based review): Report of the Quality Standards

Subcommittee of the American Academy of Neurology. *Neurology*, *68*(24), 2144-2149.

Hagen, K. B., Hilde, G., Jamtvedt, G., & Winnem, M. F. (2002). The Cochrane Review of advice to stay active as a single treatment for low back pain and sciatica. *Spine*, *27*(16), 1736-1741.

Jones, A. E., Brown, M. D., Trzeciak, S., Shapiro, N. I., Garrett, J. S., Heffner, A. C., & Kline, J. A. (2008). The effect of a quantitative resuscitation strategy on mortality in patients with sepsis: A meta-analysis. *Critical Care Medicine*, *36*(10), 2734-2739.

Kontorinis, N., Agarwal, K., & Dieterich, D. (2005). Treatment of hepatitis C virus in HIV patients: A review. *AIDS*, *19*(3), S166-S173.

Mellegers, M. A., Furlan, A. D., & Mailis, A. (2001). Gabapentin for neuropathic pain: Systematic review of controlled and uncontrolled literature. *Clinical Journal of Pain*, *17*(4), 284-295.

Salt, J., Cummings, G. G., & Profetto-McGrath, J. (2008). Increasing retention of new graduate nurses: A systematic review of interventions by healthcare organizations. *Journal of Nursing Administration*, *38*(6), 287-296.

Saposnik, G., & Del Brutto, O. H. (2003). Stroke in South America: A systematic review of incidence, prevalence, and stroke subtypes. *Stroke*, *34*(9), 2103-2107.

Singh, S., & Kumar, A. (2007). Wernicke encephalopathy after obesity surgery: A systematic review. *Neurology*, *68*(11), 807-811.

Sinuff, T., Adhikari, N. K. J., Cook, D. J., Schunemann, H. J., Griffith, L. E., Rocker, G., & Walter, S. D. (2006). Mortality predictions in the intensive care unit: Comparing physicians with scoring systems. *Critical Care Medicine*, *34*(3), 878-885.

Chapter 6 – Reliability study

Exploration of the reliability of scores obtained from the proposed critical appraisal tool (CAT) is the aim of Objective 5. Three different approaches to test theory can be used to assess reliability: classical test theory (CTT), generalizability theory (G theory), and item response theory (IRT) [1]. The reasons why CTT and G theory were used and why IRT was not used are outlined here are. A short introduction to G theory is also included.

Classical test theory became popular in the mid-1960s and is expressed by the formula [2]:

Observed score = True score + error

CTT was used in this study because it is easily recognised and the results can be compared with other CATs that normally calculate reliability using CTT statistics, such as Cronbach's alpha or the intraclass correlation coefficient (ICC). However, these measures of reliability do not break down the reasons for error into component parts. Therefore, CTT cannot be used to identify exactly where errors have occurred in the proposed CAT and cannot help to reduce those errors.

The following components may affect the observed score in the proposed CAT: the paper, the rater, the research design used in the paper, and the categories in the

proposed CAT. The use of G theory makes it possible to calculate how each of these components may affect the papers' score. In CTT terms, this can be expressed as [3]:

$$X = \mu + e_{paper} + e_{rater} + e_{research\ design} + e_{categories} + e_{residual}$$

Where

X = Observed score, μ = Universe (true) score, e = error,

$e_{residual}$ = error not specifically included

However, instead of looking at error, one aspect of G theory called a G study calculates the total observed score variance. This is the sum of all the possible component variances that may affect the total observed score variance. The G study also calculates how each of these components influences every other component to give a far more comprehensive and complex view of the total observed score variance. For example, the total observed score variance may be affected by an interaction between:

- Paper crossed with Rater: $(p \times r)$ or pr
- Paper crossed with Research design: $(p \times d)$ or pd
- Paper crossed with Rater crossed with Research design: $(p \times r \times d)$ or prd.

In this study, the total observed score variance can be expressed as [4 (pp. 4-8)]:

$$\sigma^2_{X_{prdc}} = \sigma^2_p + \sigma^2_r + \sigma^2_d + \sigma^2_c + \sigma^2_{pr} + \sigma^2_{pd} + \sigma^2_{pc} + \sigma^2_{rd} + \sigma^2_{rc} + \sigma^2_{dc} + \sigma^2_{prd} + \sigma^2_{prc} + \sigma^2_{rdc} + \sigma^2_{prdc}$$

Where

$\sigma^2_{X_{prdc}}$ = Total observed score variance, p = Paper, r = Rater,

d = Research design, c = Category,

σ^2_{pr} = Interaction between Paper crossed with Rater variance, and so forth.

This additional information makes it possible to use G theory to calculate exactly where variances in scores have occurred and to give some clues about how these variances may be reduced. Calculation of variances is achieved through ANOVA estimators or the expected mean square (EMS) of the variances, which are estimates of actual variances [4 (pp. 4-8)]. Estimates of variances can be interpreted in the same way as actual variances but are easier to use.

Item response theory was not used in this study because the proposed CAT could not meet the second of its main assumptions. These assumptions are [1 (pp. 301-302)]:

1. A scale is unidimensional.
2. The probability of answering one item in a certain way is unrelated to answering another item in the same way for papers with similar traits.

It was argued in the previous chapter that each category within the proposed CAT is unidimensional but that overall the proposed CAT is multidimensional. Therefore, the proposed CAT could meet the first assumption if each category were analysed separately. However, an extensive IRT analysis of the proposed CAT would be required to satisfy the second assumption (each item in a category is unrelated to any other item in the same category). Assuming that potential CAT items represent a two-parameter polytomous IRT model, a simple estimate of the potential sample size was 500 raters [1 (pp. 322-323)].

On the other hand, G theory could indicate where problems were occurring in the proposed CAT, which is very beneficial in the early stages of tool development and evaluation, with a sample size of five raters and 24 papers (for details of sample size, see section 6.3, p. 150). Therefore, it was decided that IRT analysis was not appropriate at this stage. However, an IRT analysis could be undertaken in the future.

The remainder of this chapter consists of an article accepted for publication on 9 August 2011 and available online 11 November 2011 (Appendix C.5):

> Crowe, M., & Sheppard, L. (2011). Reliability analysis of a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology*, (Online). doi:10.1016/j.jclinepi.2011.08.006

Changes may be made to the article when published. In the event that copyright permission may be required for this article, it can be found in Appendix A.3.

Reliability analysis of a proposed critical appraisal tool demonstrated value for diverse research designs

6.1 ABSTRACT

Objective – To examine the reliability of scores obtained from a proposed critical appraisal tool (CAT).

Study design and setting – Based on four raters and a random sample of 24 health-related research papers, the scores obtained from the proposed CAT were examined using intraclass correlation coefficients (ICC), generalizability theory, and participants' feedback.

Results – The ICC for all research papers was 0.83 (consistency) and 0.74 (absolute agreement) for four participants. The highest ICC (consistency) for individual research designs was for qualitative research (0.91) and the lowest was for descriptive, exploratory or observational research (0.64). The G study showed a moderate research design effect (32%) for scores averaged across all papers. The research design effect was mainly in the *Sampling*, *Results*, and *Discussion* categories (44%, 36%, and 34% respectively). When scores for each research design were analysed, there was a majority paper effect for each (53–70%), with small to moderate rater or paper × rater interaction effects (0–27%).

Conclusions – Reasons for the research design effect were participant unfamiliarity with some of the research designs and that papers were not matched to participants' expertise. Even so, the proposed CAT shows great promise as a tool that can be used across a wide range of health research designs.

6.2 BACKGROUND

Critical appraisal, a core technique in evidence-based practice (EBP) and systematic reviews, is a standardised way of assessing research so that decisions can be made based on the best evidence available [5, 6 (pp. 1-5)]. A large number of CATs have been developed to achieve an efficient approach to critical appraisal [5, 7]. Unfortunately, many of these CATs have fundamental flaws that prevent them from being truly useful for appraising research. These problems include: tools that are limited in the research designs that can be assessed; tools that lack comprehensiveness in their appraisal approach; and tools that use inappropriate scoring methods, which can hide poor research [5, 7, 8, 9, 10]. The greatest concern is that CATs are used to assess the research validity and reliability, but many CATs have been designed with little or no evidence for score validity or reliability [5, 11, 12]. A review of 45 papers (Chapter 4), which reported on how CATs were designed, found that 38 (84%) had little or no evidence of score validity, and 34 (76%) had no evidence of score reliability [5].

A new structure for a CAT was proposed based on the review and the evidence available for designing CATs in Chapter 4 [5]. The proposed CAT attempted to overcome the shortfalls in previous CATs by having a structure that could be used across all research types, comprehensively assessing research, and having an appropriate scoring system [5, 13]. The structure was based initially on research validity but that method was abandoned because:

1. Assessment of the research was often limited internal research validity, ignoring external and conclusion validity.
2. Issues such as clear objectives or reasons certain decisions were made did not readily fit research validity but were still considered important within critical appraisal.

3. Using research validity criteria to assess different types of research was difficult and time consuming [5].

Therefore, a structure for the proposed CAT was developed based on seven reporting guidelines, research methods theory, and a qualitative analysis of existing CATs [5]. The proposed CAT consisted of eight categories such that each category contained items which were most similar but the categories themselves were dissimilar. The categories were: *Preamble*, *Introduction*, *Design*, *Sampling*, *Data collection*, *Ethical matters*, *Results*, and *Discussion*. To appraise papers there were numerous item descriptors to be examined within each category, as seen in Table 6.1.

The next step was to determine whether the proposed CAT could validly appraise different research designs [13]. The validation process closely followed the guidelines outlined in the *Standards for educational and psychological testing*, which require a combination of theory, empirical evidence, and a context for validity testing [14 (pp. 9-17)]. The validity study had two major aims: (1) to develop a scoring system for the proposed CAT; and (2) to determine whether each of the categories were necessary to appraise research [13]. The scoring system did not require each item or item descriptor to be scored individually. Instead, items were marked as being present, absent but should be present, or not applicable based on the research design used in the paper appraised. However, this was not a simple check list that could lead to inflexibility and inaccuracy. The appraiser made a decision about what score a category should receive based on the marked item descriptors plus their overall assessment of that category. Scoring each category was on a scale from zero (no evidence) to five (highest evidence), where only whole numbers (integers) were used. Furthermore, the evidence must be stated in the paper and could not be assumed. This was in keeping with other CATs, reporting guidelines, and procedures for conducting systematic reviews [7, 15, 16]. A user

guide was developed for the proposed CAT to assist with scoring and as a necessary component of validity [13] (see section 6.9.1, p. 170).

Table 6.1 Proposed critical appraisal tool (CAT)

Category Item	**Item descriptor** [☑ Present; ☒ Absent; ■ Not applicable]	**Score** [0–5]
Preamble		
Text	1. Sufficient detail others could reproduce ☐ 2. Clear/concise writing ☐, table(s) ☐, diagram(s) ☐, figure(s) ☐	**Preamble score**
Title	1. Includes study aims ☐ and design ☐	
Abstract	1. Key information ☐ 2. Balanced ☐ and informative ☐	
Introduction		
Background	1. Summary of current knowledge ☐ 2. Specific problem(s) addressed ☐ and reason(s) for addressing ☐	**Introduction score**
Objective	1. Primary objective(s), hypothesis(es), or aim(s) ☐ 2. Secondary question(s) ☐	
Design		
Research design	1. Research design(s) chosen ☐ and why ☐ 2. Suitability of research design(s) ☐	**Design score**
Intervention, Treatment, Exposure	1. Intervention(s)/treatment(s)/exposure(s) chosen ☐ and why ☐ 2. Precise details of the intervention(s)/treatment(s)/exposure(s) ☐ for each group ☐ 3. Intervention(s)/treatment(s)/exposure(s) valid ☐ and reliable ☐	
Outcome, Output, Predictor, Measure	1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen ☐ and why ☐ 2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s) ☐ 3. Outcome(s)/output(s)/predictor(s)/measure(s) valid ☐ and reliable ☐	
Bias, etc	1. Potential bias ☐, confounding variables ☐, effect modifiers ☐, interactions ☐ 2. Sequence generation ☐, group allocation ☐, group balance ☐, and by whom ☐ 3. Equivalent treatment of participants/cases/groups ☐	
Sampling		
Sampling method	1. Sampling method(s) chosen ☐ and why ☐ 2. Suitability of sampling method ☐	**Sampling score**
Sample size	1. Sample size ☐, how chosen ☐, and why ☐ 2. Suitability of sample size ☐	
Sampling protocol	1. Target/actual/sample population(s): description ☐ and suitability ☐ 2. Participants/cases/groups: inclusion ☐ and exclusion ☐ criteria 3. Recruitment of participants/cases/groups ☐	
Data collection		
Collection method	1. Collection method(s) chosen ☐ and why ☐ 2. Suitability of collection method(s) ☐	**Data collection score**
Collection protocol	1. Include date(s) ☐, location(s) ☐, setting(s) ☐, personnel ☐, materials ☐, processes ☐ 2. Method(s) to ensure/enhance quality of measurement/instrumentation ☐ 3. Manage non-participation ☐, withdrawal ☐, incomplete/lost data ☐	
Ethical matters		
Participant ethics	1. Informed consent ☐, equity ☐ 2. Privacy ☐, confidentiality/anonymity ☐	**Ethical matters score**
Researcher ethics	1. Ethical approval ☐, funding ☐, conflict(s) of interest ☐ 2. Subjectivities ☐, relationship(s) with participants/cases ☐	
Results		
Analysis, Integration, Interpretation method	1. A.I.I. method(s) for primary outcome(s)/output(s)/predictor(s) chosen ☐ and why ☐ 2. Additional A.I.I. methods (e.g. subgroup analysis) chosen ☐ and why ☐ 3. Suitability of analysis/integration/interpretation method(s) ☐	**Result score**
Essential analysis	1. Flow of participants/cases/groups through each stage of research ☐ 2. Demographic and other characteristics of participants/cases/groups ☐ 3. Analyse raw data ☐, response rate ☐, non-participation/withdrawal/incomplete/lost data ☐	
Outcome, Output, Predictor analysis	1. Summary of results ☐ and precision ☐ for each outcome/output/predictor/measure 2. Consideration of benefits/harms ☐, unexpected results ☐, problems/failures ☐ 3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes) ☐	
Discussion		
Interpretation	1. Interpretation of results in the context of current evidence ☐ and objectives ☐ 2. Draw inferences consistent with the strength of the data ☐ 3. Consideration of alternative explanations for observed results ☐ 4. Account for bias ☐, confounding/effect modifiers/interactions/imprecision ☐	**Discussion score**
Generalisation	1. Consideration of overall practical usefulness of the study ☐ 2. Description of generalisability (external validity) of the study ☐	
Concluding remarks	1. Highlight study's particular strengths ☐ 2. Suggest steps that may improve future results (e.g. limitations) ☐ 3. Suggest further studies ☐	

The validity process also tested the proposed CAT against five other CATs that had validity and reliability data available [12, 17, 18, 19, 20]. This showed that all the categories used in the proposed CAT, except *Preamble*, could be considered suitable for critical appraisal. There was insufficient evidence to make a decision on the *Preamble* category because it could only be compared to one of the five alternative CATs. Therefore, it was decided to include the *Preamble* category in the proposed CAT until more evidence could be gathered [13].

The third step, after design and validity, was to examine reliability of scores obtained from the proposed CAT – the overall objective of this study. Reliability, in this instance, refers to how closely a number of different raters agree on the score that should be given to a particular piece of research (inter-rater reliability). The intraclass correlation coefficient (ICC) is a common method for measuring reliability [12, 20, 21]. However, the ICC is based on classical test theory, and breaks the score into true score and error. This does not allow for further analysis of exactly where the error in the score occurred. Generalizability theory (G theory), in a process known as a G study, breaks scores down into the universal (or true) score plus where errors occur due to the tool, raters, environmental conditions, or any other factor that may have potentially influenced the score [1, 3, 22]. This ability to find where errors occur was seen as vital to understanding of the proposed CAT. However, it should be noted that ICCs and G coefficients (from G studies) cannot be directly compared, except in specific circumstances [1].

Given the previous use of ICCs and the flexibility of generalizability theory, the aims of this reliability study were to determine:

1. Whether the scores obtained by the proposed CAT were reliable as determined by the ICC.
2. Which factors (or facets) contributed most to the mean variances in scores (G study).

3. How many raters may be required per paper to obtain optimal reliability of the scores (D study).
4. Participants' feedback on the proposed CAT and user guide.

6.3 METHODS

6.3.1 Design

The study design was exploratory because the purpose was to discover whether the scores obtained by the proposed CAT were reliable. Each participant was given a random selection of papers to appraise. The papers selected were based on six research designs:

1. True experimental.
2. Quasi-experimental.
3. Single system.
4. Descriptive, exploratory or observational (DEO).
5. Qualitative.
6. Systematic review.

The reasons these broad classes of research design were chosen and how the individual papers were selected have been fully outlined in the previous chapter [13]. Briefly, the six research designs were based on broad groupings of research designs from the literature [23, 24, 25] and Chapter 3. A pre-determined search strategy was used to find papers based on the research design types and limit results to substantial research papers. The papers were randomly selected from the full text journals subscribed to by James Cook University (JCU) through OvidSP (Ovid, New York) in September 2009, using the random sequence generator available from RANDOM.ORG [26]. Each paper was read by the author to confirm the research

design. This ensured the selected papers belonged to the required research design. A full list of the papers used in this study is available in section 6.9.2, p. 173.

Ethical approval for this study was obtained from James Cook University Human Research Ethics Committee (No. H3415). Informed consent was obtained from each participant before they voluntarily took part in the study. An example of the information sheet, informed consent form, and questions for participants are contained in Appendix D. Participants could withdraw at any stage without explanation or prejudice. There were no conflicts of interest or funding sources to declare.

A sample size calculation was based on the work of Walter, Eliasziw, & Donner [27]. The formulae for the sample size were:

$$k = 1 + \frac{2(z + (1-\beta))^2 n}{(\ln C_0)^2 (n-1)}$$

Where

$$C_0 = \frac{1 + \left(n\left(\frac{r_{min}}{1-r_{min}}\right)\right)}{1 + \left(n\left(\frac{r}{1-r}\right)\right)}$$

k, number of raters
z, z-score
$1 - \beta$, power
n, number of papers
ln, natural log
r, required reliability coefficient
r_{min}, minimum acceptable reliability coefficient

A Microsoft Excel worksheet function of the formulae and a decision table were developed (section 6.9.3, p. 176). This calculated that a sample size of five participants appraising four papers each were required to obtain an ICC (r) of 0.90 (α = 95%, 1–β = 0.85, r_{min} = 0.40). A total of 24 papers was required per participant because six research designs were being tested. Participants were recruited from a convenience sample of academic staff from the James Cook University Schools of Public Health, Tropical Medicine and Rehabilitation Science, and Medicine and Dentistry. Staff were emailed regarding the study and a total of six participants voluntarily agreed to enrol.

It was decided that missing data from a participant would be scored based on the median score of the same category, from the same research design, for that participant, rounded to the nearest integer. This decision was made before any data were collected. If the missing value for a participant was from a true experimental design paper, in the *Sampling* category, for example, then the median of the *Sampling* category from the remaining true experimental design papers for that participant would constitute the missing value. This strategy was used because it had the least effect on scores and was the most conservative value of Mean, Median and Mode, given the statistical analysis used.

It has been stated in the literature that reliability scores for non-clinical tools should be at least 0.70 and for clinical tools at least 0.90, although the method used to calculate these scores is generally not included [1]. However, for this study there were three reasons why the reliability scores might be lower than expected or might show some inconsistency. First, the papers were a random selection, which meant there was no coherence between them. This, in turn, might make appraisal more difficult to accomplish. Second, the papers were taken from the broad field of health research, whereas most appraisal of research is confined to one or a limited number of related fields. Third, some or all of the papers might have been outside a participant's expertise, making it more difficult for them to accurately rate papers. Even given these issues, the design, sample and data collection was conducted in a manner where a high reliability coefficient was expected.

6.3.2 Data collection

All participants were supplied with a guide to using the proposed CAT and appraisal forms. The form (Table 6.1) and user guide were the same as those used in Chapter 5, which showed that the form could be considered a valid method of obtaining scores when appraising research [13]. Each paper and form had an

identification label so that the participants and subsequent analysis could be attributed to individual papers. Furthermore, the research design for each paper was printed on the paper and critical appraisal form. This was done to eliminate the possibility that participants might mistake the research design used in a paper. Identifying the research design meant this variable was controlled without affecting the overall purpose of the study, and participants could concentrate on critically appraising the papers rather than assessing the research design used.

Each participant was given a folder containing the papers, forms, and a user guide. Participants were instructed to read the user guide before appraising any papers because it had information regarding how to use the proposed CAT. Participants were given a six-week period between March and April 2010 to read and appraise the papers, and return the forms. Participants were emailed every two weeks to remind them of the completion date and to check whether there were any problems. Participants were also informed to contact the author if they had any problems with the tool or appraising the papers. None of the participants requested assistance during this time.

After appraising the papers, participants were questioned, by means of a semi-structured questionnaire, about their research experience, and their perception of the proposed CAT and the guide to using the tool. The purpose was to determine whether there was any difference in how participants appraised a paper based on their research experience, and to gain feedback on the proposed CAT and user guide so that both could be improved for future use.

6.4 RESULTS

Five of the six participants who volunteered returned appraisal data. Two of the participants self-rated themselves as being very experienced researchers (Raters II

and V). The remaining three participants self-rated themselves as moderately experienced researchers (Raters I, III, and IV).

Overall, there were five incidents of accidental missing data, which were scored based on the method outlined in section 6.3.1. However, in three out of five cases, participants had purposely marked the *Ethical matters* category as not applicable for all systematic review papers even though the user guide clearly stated that all categories should be scored for all research designs. When questioned about this, all three participants stated that they thought that *Ethical matters* for systematic reviews was irrelevant because approval from an ethics committee was not required to complete this type of research. When asked whether they thought sources of funding or conflicts of interest were ethical issues that should be stated in a systematic review paper, the three participants agreed that this was true and that they should have included an *Ethical matters* score, but had limited their thinking to participant ethics rather than including researcher ethics.

Due to this unforseen circumstance, the missing data strategy for *Ethical matters* was altered:

1. Where the *Ethical matters* category was calculated for systematic reviews, the ICC only used two participant scores.
2. G_String_III [28], the software used for G theory calculations, automatically replaced missing data with the grand mean for the category being calculated.
3. Where the total score % for systematic reviews were calculated, the median value from the two participants that had scored the *Ethical matters* category for systematic reviews replaced the missing *Ethical matters* scores. This prevented the *Total scores* for the participants that had not scored the *Ethical matters* categories for systematic reviews being much lower than expected and, thereby, negatively biasing the results, while at the same time this method did not positively bias the results.

It was assumed that in calculations for the ICCs, G study and D study, the paper and rater effects were random (the papers and raters were not the only possible papers or raters) and the category effects were fixed (there were no additional categories) [4, 29]. Two types of coefficients can be calculated for an ICC and a G study. These are coefficients for consistency (C) and absolute agreement (A) in ICCs, and relative error ($E\rho^2$) and absolute error (Φ) coefficients in G studies. In terms of consistency and relative error, the coefficients are calculated based on whether the raters rank the entity being measured in the same order regardless of the real score given. In terms of absolute agreement and absolute error, the coefficients are calculated based on whether the raters give the same real scores to the entity being measured. As a result, the consistency/relative error coefficient is normally higher than the absolute agreement/absolute error coefficient [1, 4, 30].

In general, when there are fewer raters, the ICC and G coefficients are lower. This occurred in this study when raters I, II, IV, and V were individually removed from analysis so that only four raters remained in each case. However, when Rater III was removed from analysis, the ICC and G coefficients increased, particularly in true experimental (23%), DEO (43%), and qualitative research designs (28%), and in the *Sample* (8%) and *Ethical matters* (15%) categories. In other words, Rater III scored papers much differently than other raters. Conversations with Rater III made it clear that they had not read the user guide and, as a result, had not scored papers in accordance with the nature of the CAT. It was, therefore, decided to exclude Rater III scores from data analysis.

6.4.1 Intraclass correlation coefficient (ICC)

Each ICC was calculated using SPSS Statistics version 18.0.2 (SPSS, Chicago IL) using the RELIABILITY command. Since the assumption was that the paper effects were random and the category effects were fixed, the MODEL subcommand used the

MIXED value, the TYPE subcommand was calculated for consistency and absolute agreement, and other subcommands used were the defaults [31 (pp. 1704-1712)].

The total score % for all research designs had an ICC for consistency of 0.83 and absolute agreement of 0.74 (Table 6.2). The total score % for each research design had ICCs for consistency of (highest to lowest): qualitative 0.91; systematic review 0.89; single system 0.85; true experimental 0.75; quasi-experimental 0.72; and DEO 0.64. The total score % ICCs for absolute agreement were (highest to lowest): true experimental 0.73; qualitative 0.67; systematic review 0.67; DEO 0.65; quasi-experimental 0.60; and single system 0.57.

The ICCs for consistency for each category were (highest to lowest): *Ethical matters* 0.84; *Results* 0.75; *Design* 0.73; *Introduction* 0.70; *Discussion* 0.68; *Sample* 0.66; *Preamble* 0.58; and *Data collection* 0.54. The ICCs for absolute agreement for the categories were (highest to lowest): *Ethical matters* 0.78; *Design* 0.65; *Sample* 0.62; *Discussion* 0.62; *Results* 0.60; *Introduction* 0.53; *Data collection* 0.52; and *Preamble* 0.50.

Table 6.2 Summary of ICCs (k = 4, excludes Rater III)

Category	Research designs													
	TE (n=4)		**QE** (n=4)		**SS** (n=4)		**DEO** (n=4)		**QL** (n=4)		**SR** (n=4)		**All** (N=24)	
	C	*A*	*C*	*A*	*C*	*A*	*C*	*A*	*C*	*A*	*C*	*A*	*C*	*A*
Preamble	0.80	0.80	0.49	0.28	0	0	0.26	0.24	0.68	0.65	0.89	0.78	0.58	0.50
Introduction	0	0	0	0	0.91	0.57	0.56	0.38	0.78	0.71	0.33	0.08	0.70	0.53
Design	0.81	0.72	0.81	0.76	0	0	0	0	0.81	0.57	0.92	0.75	0.73	0.65
Sample	0.69	0.73	0	0	0	0	0.67	0.63	0.84	0.46	0.59	0.21	0.66	0.62
Data collection	0	0	0.80	0.70	0.65	0.48	0	0	0.95	0.85	0.56	0.57	0.54	0.52
Ethical matters	0.72	0.73	0.76	0.62	0.93	0.87	0.81	0.79	0.79	0.62	*0.24	*0.20	0.84	0.78
Results	0	0	0.55	0.40	0.30	0.13	0	0	0.14	0.06	0.75	0.62	0.75	0.60
Discussion	0.53	0.37	0.81	0.58	0.59	0.39	0	0	0.84	0.75	0	0	0.68	0.62
Total score %	0.75	0.73	0.72	0.60	0.85	0.57	0.64	0.65	0.91	0.67	0.89	0.67	0.83	0.74

TE, true experimental; QE, quasi-experimental; SS, single system; DEO, descriptive, exploratory, and observational; QL, qualitative; SR, systematic review; C, consistency; A, absolute agreement; *k*, number of raters; n, papers per research design; N, total papers.

* *k* = 2 (missing data)

6.4.2 G and D study

The G and D study results were calculated using a combination of SPSS Statistics version 18.0.2 (SPSS, Chicago IL) and G_String_III [28]. In SPSS, the command used was VARCOMP, and the METHOD subcommand used was Minique(1) [29]. In the G study, the object of measure was the paper (p), or paper nested within a research design ($p{:}d$). The majority of mean variance should be accounted for in the object of measure. Main effects are research design (d), rater (r), and category (c) which may have an influence on the object of measure. Interaction effects are interactions between the object of measure and other items, or interactions just between other items, which may have an influence on the object of measure (for example, rater crossed with paper (pr) or research design crossed with category nested within research design ($pc{:}d$)) [4].

The percentage mean variance components for all papers were analysed to obtain a sense of where variances were occurring (Table 6.3, Part 1). This showed that 38% of variance was due to the paper effect (p) and 32% was from research design effect (d).

The percentage mean variance components for average research design scores were analysed to explore how research design affected variances (Table 6.3, Part 2). This showed that the majority of variance was for the object of measure (p) (53–70%). The interaction effect of paper crossed with rater (pr) for DEO was 27% of variance. The rater effect (r) was 27% of variance for qualitative and 22% for systematic review. Interaction effects for paper crossed with category (pc), rater crossed with category (rc), and paper crossed with rater crossed with category (prc) were minimal in each research design.

Table 6.3, Part 3 shows how average category scores were affected by variance. $Data\ collection$ and $Ethical\ matters$ categories had the majority of variance from paper nested within research design ($p{:}d$) at 52% and 72% respectively. $Sampling$, $Results$,

Table 6.3 Percentage mean variance components (k = 4, excludes Rater III)

Part 1 Average all papers

Effect	%
p:d	38
d	32
r	8
dr	5
dc	0
pr:d	7
pc:d	4
rc	1
drc	0
prc:d	3

Part 2 Average each research design

Effect	Research design					
	TE	*QE*	*SS*	*DEO*	*QL*	*SR*
p	60	70	59	53	60	65
r	0	0	15	0	27	22
pr	13	10	8	27	0	3
pc	7	5	9	9	6	1
rc	10	7	6	2	3	5
prc	10	9	4	8	4	5

Part 3 Average each category

Effect	Pre-amble	Intro-duction	Design	Samp-ling	Data collect.	Ethical matter	Results	Disc-ussion	Total score
p:d	34	28	43	21	52	72	17	29	44
d	17	25	23	44	1	4	46	34	31
r	12	24	10	3	3	7	18	8	9
dr	9	0	9	12	5	2	2	8	5
pr:d	28	23	16	21	40	15	17	21	10

Part 4 Individual research design and category

Category	Effect	Research design					
		TE	*QE*	*SS*	*DEO*	*QL*	*SR*
Pre-amble	*p*	80	28	0	24	65	78
	r	0	44	63	6	5	12
	pr	20	28	37	70	30	10
Intro-duction	*p*	0	0	57	38	71	8
	r	42	5	37	32	9	76
	pr	58	95	6	30	20	16
Design	*p*	72	76	0	0	57	75
	r	12	6	67	38	29	19
	pr	17	17	33	62	13	6
Samp-ling	*p*	69	0	0	63	46	21
	r	0	0	38	5	46	64
	pr	31	100	62	32	9	15
Data collect.	*p*	0	70	48	0	85	56
	r	0	12	26	0	11	0
	pr	100	18	26	100	5	44
Ethical matters	*p*	72	62	87	78	62	*38
	r	0	19	7	3	22	*13
	pr	28	19	6	18	16	*49
Results	*p*	0	40	13	0	6	62
	r	71	27	56	9	60	18
	pr	29	33	31	91	34	21
Disc-ussion	*p*	37	58	39	0	75	0
	r	30	29	33	0	10	59
	pr	33	13	27	100	15	41

p, paper;
d, research design;
r, rater (random);
c, category (fixed);
p or p:d, object of measure;
TE, true experimental;
QE, quasi-experimental;
SS, single system;
DEO, descriptive, exploratory, and observational;
QL, qualitative;
SR, systematic review;
k, number of raters.

* k = 2.

All percentages rounded to the nearest integer.

and *Discussion* categories, and total score % had high variance contributed by the research design effect (d) (44%, 46%, 34%, and 31% respectively). The *Introduction* category had a combination of 49% variance attributable to research design (d = 25%) and rater (r = 24%) effects, while variance for *Preamble* (28%) and *Data collection* (40%) categories was due to an interaction effect in paper crossed with rater nested within research design ($pr:d$).

Examination of individual categories within each research design (Table 6.3, Part 4) showed that 15 of the 48 possible combinations had a 0–10% paper effect (p). Six of these 15 had a 90–100% paper crossed with rater (pr) interaction effect. Qualitative research showed the best results, with six categories having majority paper effects (57–85%). Next were true experimental, quasi-experimental, and systematic review with four categories having majority paper effects (56–80%), followed by DEO (two categories) and single system (one category). The *Ethical matters* category had the best results across research designs, with five out of six designs showing majority paper effects (62–87%). The next best results were for *Design*, with four research designs having majority paper effects. This was followed by *Preamble* and *Data collection* with three each.

Finally, a D study was undertaken to determine the total score % coefficients with different numbers of raters per paper and with all other variables kept equal (Table 6.4). The number of raters calculated were 1–3, 5, and 10. The greatest change in G coefficients was between one and two raters, with the change between two and three raters and beyond being progressively less, similar to the law of diminishing returns where every extra rater returns a diminishing increase in reliability coefficients.

Table 6.4 D study (excludes Rater III)

No. of Raters	Research designs													
	TE (n=4)		**QE** (n=4)		**SS** (n=4)		**DEO** (n=4)		**QL** (n=4)		**SR** (n=4)		**All** (N=24)	
	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ
1	0.43	0.40	0.40	0.27	0.59	0.25	0.30	0.30	0.73	0.34	0.67	0.34	0.52	0.25
2	0.60	0.58	0.57	0.43	0.74	0.40	0.47	0.47	0.84	0.51	0.81	0.50	0.68	0.35
3	0.69	0.67	0.66	0.53	0.81	0.50	0.57	0.57	0.89	0.61	0.86	0.60	0.76	0.41
4	0.75	0.73	0.72	0.60	0.85	0.57	0.64	0.64	0.91	0.67	0.89	0.67	0.81	0.44
5	0.79	0.77	0.77	0.65	0.88	0.62	0.69	0.69	0.93	0.72	0.91	0.72	0.84	0.47
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	0.88	0.87	0.87	0.79	0.94	0.77	0.81	0.81	0.96	0.84	0.95	0.84	0.92	0.52

TE, true experimental; QE, quasi-experimental; SS, single system; DEO, descriptive, exploratory, and observational; QL, qualitative; SR, systematic review; G coefficient: $E\rho^2$, relative error; Φ, absolute error; n, papers per research design; N, total papers.

6.4.3 Participant reactions

Participants thought the strengths of the proposed CAT were that it covered all areas of research methods; it separated research methods into individual categories so that the appraiser got an impression of which parts of a paper were good or bad, as well as an overall impression of the paper; and that areas such as *Ethical matters* and *Sampling* were included, which were not covered in other CATs. Weaknesses of the proposed CAT were that sometimes participants caught themselves giving a low rating to a category because a number of items were missing. However, those items should have been marked as not applicable. Other comments made on weaknesses of the proposed CAT were that it was easier to use at the ends of the research continuum (for example, true experimental and qualitative research) rather than the middle; some items were confusing, such as outlying data and subgroup analysis; and a lot of items were not applicable, especially in qualitative research.

Strengths of the user guide included that it clarified how to use the tool in each research design and category; it was useful to refer back to when a participant was unsure how to score something; and it had the right amount of information on research. Weaknesses of the user guide were that it was both too short and slightly

too long; it needed more information on research designs and methods; it was difficult to decide if something was not applicable or absent; and there were not enough examples to help guide appraisal of the papers.

Finally, participants thought that other uses for the proposed tool could be a template for writing a research paper; a tool to peer review articles; teaching how to critically appraise research; and to appreciate the complexity of research. Other comments by participants were that they were more comfortable appraising research methods with which they were familiar and found it difficult to appraise papers that were outside their field of expertise.

6.5 Discussion

The most unexpected result was that three out of five of the participants did not appraise the *Ethical matters* category for systematic reviews. These participants stated that they had seen *Ethical matters*, in systematic reviews only, in purely participant terms rather than in participant/researcher terms, as stated in the proposed CAT and user guide. Why conducting a systematic review should be any different from other types of research was a question that could not be answered in this study. However, it opens the issue to further study. Also, it could not be determined whether there was a tendency for these three participants to score other research designs more from a participant viewpoint than in participant/researcher terms. This too requires further study.

The decision to remove the scores supplied by Rater III could be considered doubtful. However, Rater III had not consulted the user guide to rate papers and the user guide forms an integral part of score validity in any testing regimen [13, 14]. Therefore, it was decided that removal of Rater III's data was the most appropriate course of action. Keeping Rater III's scores simply because they were collected

would negate the validity of the scores and, thereby, void the reliability results for the proposed CAT.

When the total score % given to each research design was examined, there were ICCs for consistency of between 0.72 and 0.91 for all research designs except DEO (0.64). Therefore, although participants were cognisant of the difficulty in appraising papers outside their experience of research methods, they still rated papers reasonably consistently (above the 0.70 level for non-clinical tools indicated earlier) [1]. This was also evident from the G study, which showed that the majority of mean variance was due to the paper effect (p) across each of the research designs. Only in DEO research was there a noteworthy interaction effect, paper crossed with rater (pr), which also showed as the lowest ICC for consistency. Furthermore, ICCs for absolute agreement were high in each research design (0.57–0.73), which showed that the actual score raters gave to each paper were reasonably similar.

A core tenet for the proposed CAT was that the total score % should not be the sole indicator of how a research paper was appraised and each category score should stand as an indicator of the standard of a paper. However, although the ICCs for consistency were reasonably high for each category (0.54–0.84), they were still lower than those for research designs. The reason for this became apparent from the G study, which showed that the *Introduction*, *Design*, *Sampling*, *Results*, and *Discussion* categories had a substantial research design effect (d, 23–46%) – the scores in these categories were affected by the research design of the paper being appraised. Among the three remaining categories, *Preamble* and *Data collection* had a substantial interaction effect ($pr{:}d$, 28% and 40% respectively) – a combination of paper crossed with rater and nested within research design influenced score variance. The exception was the *Ethical matters* category, which had no large main or interaction effects.

A possible explanation for the variability in the results could be that the categories were not appropriate for each research design, which led to fluctuations in scoring. However, where the percent mean variance component for the category facet was extracted (*pc*, *rc*, *prc* in Table 6.3, Part 2), it was very low (0–10%). Therefore, the most likely causes of these variations were two-fold. First, participants stated that they had greater experience with some research designs over others, and that they found it more difficult to appraise papers which used research designs they were less familiar with. Second, participants' expertise was not matched to the papers in the study because the papers were randomly selected. A result of this was that papers which were outside a participant's expertise were more difficult to rate and more likely to be rated more inconsistently than papers where the participant was more familiar with the subject matter.

As an example of these two issues influencing scores, all participants stated that they were most uncomfortable with the single system papers because they were unfamiliar with the research design and lacked the knowledge for the topics covered in those papers. The ICC for single system designs reflected these issues because the difference between the consistency (0.85) and absolute (0.57) coefficients was the highest (0.28) for any research design, meaning that although participants ranked the single system papers similarly, they did not agree on the real score the papers should receive. Also, the percentage mean variance components for single system designs across each category were unimpressive, with only two majority paper effects for *Ethical matters* (87%) (which was most consistent across each research design) and *Introduction* (57%), and three 0% paper effects for *Preamble*, *Design*, and *Sampling*, which reflects the participants' lack of familiarity with the research design (Table 6.3, Part 4). Similar statements of unfamiliarity with DEO research designs and the subject matter in those papers was evident in the ICC for consistency of 0.64, which was the lowest of all research designs. In the G study for

DEO research, there were only two majority paper effects (*Sampling* 63% and *Ethical matters* 78%) and four o% paper effects (*Design*, *Data collection*, *Results*, and *Discussion*).

Finally, the D study showed that a minimum of three raters should be used to achieve consistently high relative error ($E\rho^2$ approx. 0.70) or absolute error (Φ approx. 0.50) coefficients. This differs from conventional thinking on systematic reviews, which states that a minimum of two raters are required and a third, or subsequent rater, is only necessary when the other raters cannot come to a consensus. These results indicate that further investigation should be undertaken to empirically determine the optimum number of raters required to appraise research papers.

6.6 Conclusion

When interpreting the scores obtained by this research, three things need to be kept in mind: (1) a random selection of 24 papers was used across six research designs; (2) participants' expertise did not necessarily match the subject matter in the papers, even though the papers and participants were from health-related disciplines; and (3) participants' knowledge of research designs was self-described as being limited to those they had experience using.

Even given these limits, the proposed CAT shows great promise in being a viable tool that can be used across a wide range of research designs and appraisal situations. There were little or no category effects across the research designs, meaning that the categories are appropriate for different types of research design. Much of the variability in scores may be due to the diverse subject matter of papers, and participants' unfamiliarity with some research designs. Improvements to the user guide to overcome this variability can be made by including examples for each

research design in each category. However, the problems with the proposed CAT should not be overstated or taken out of context because they are less likely to feature in situations where raters are familiar with the subject matter, and the research designs used to gather data for that subject matter.

6.7 IN SUMMARY

- The reliability of scores from the proposed CAT were tested using the intraclass correlation coefficient (ICC) and generalizability theory (G theory).
- The proposed CAT obtained consistency ICCs from 0.91 to 0.64, depending on the research design.
- G theory showed that raters had difficulty appraising papers where they were unfamiliar with the subject matter or lacked experience with the research design.
- The next chapter compares the use of a structured CAT with no CAT when appraising research papers (Objective 6).

6.8 REFERENCES

1. Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford: Oxford University Press.

2. Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi:10.1207/S15327752JPA8001_18

3. Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527-551). San Diego, CA: Academic Press.

4. Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

5. Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

6. Glasziou, P., Irwig, L., Bain, C., & Colditz, G. (2001). *Systematic reviews in health care: A practical guide*. Cambridge, MA: Cambridge University Press.

7. Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitc, C., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27). doi:10.3310/hta7270

8. Armijo Olivo, S., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156-175. doi:10.2522/ptj.20070147

9. Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1060. doi:10.1001/jama.282.11.1054

10. Moyer, A., & Finney, J. W. (2005). Rating methodological quality: Toward improved assessment and investigation. *Accountability in Research, 12*(4), 299-313. doi:10.1080/08989620500440287

11. Bialocerkowski, A. E., Grimmer, K. A., Milanese, S. F., & Kumar, S. (2004). Application of current research evidence to clinical physiotherapy practice. *Journal of Allied Health, 33*(4), 230-237.

12. Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

13. Crowe, M., & Sheppard, L. (2011). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies,* (Online). doi:10.1016/j.ijnurstu.2011.06.004

14. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

15. Khan, K. S., ter Riet, G., Glanville, J., Sowden, A. J., & Kleijen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews (CRD Report 4). York, England: University of York.

16. Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA, 285*(15), 1992-1995. doi:10.1001/jama.285.15.1992

17. Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *JAMA, 272*(2), 101-104. doi:10.1001/jama.1994.03520020027007

18. Reis, S., Hermoni, D., Van-Raalte, R., Dahan, R., & Borkan, J. M. (2007). Aggregation of qualitative studies - From theory to practice: Patient priorities and family medicine/general practice evaluations. *Patient Education and Counseling, 65*(2), 214-222. doi:10.1016/j.pec.2006.07.011

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

20. Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18*(4), 385-401. doi:10.1080/09602010802009201

21. Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12. doi:10.1016/0197-2456(95)00134-4

22. Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. doi:10.1177/0013164404266386

23. Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

24. Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Pearson.

25. Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

26. Haadr, M. (2009). Random.org: Random sequence generator. Retrieved 29 January 2011, from http://www.random.org/sequences/

27. Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*(1), 101-110. doi:10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e

28. Bloch, R. (2010). G_String_III (Version 5.4.6). Hamilton, ON: Programme for Educational Research and Development. Retrieved from http://fhsperd.mcmaster.ca/g_string/

29. Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 43-58). Thousand Oaks, CA: Sage.

30. Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

31. SPSS Inc. (2009). *PASW Statistics 18 command syntax reference*. Chicago, IL: IBM SPSS Inc.

6.9 ADDITIONAL MATERIAL

6.9.1 User guide for the proposed CAT (reliability study)

The difference between the user guide for the evaluation of validity (section 5.8.3) and the user guide for the reliability study is in the introductory material. There is no difference in the descriptions of Categories and Items (see section 5.8.3.3).

Introduction

The critical appraisal tool assumes an awful lot. It assumes that the individual using the tool is familiar with research designs, sampling techniques, ethics, data collection methods, and statistical and non-statistical data analysis techniques.

It may be helpful to have a general research methods text available to refer to when appraising papers.

The papers being appraised are unlikely to have the information sought in the sequence outlined in the critical appraisal form. Therefore, it is suggested to read each paper quickly from start to finish, getting an overall sense of what is being discussed. Then re-read the paper and fill in the scores.

Research design and Paper ID

Each paper and each critical appraisal form has two pieces of information at the top:

1. Research design – To make appraising each paper a little easier, the research design used is written on the paper and on the form. This means that you do not need to decide which research design was used and can concentrate on appraising the paper based on the research design indicated. Two of the research designs have alternative names that you may be more familiar with: single system designs include n-of-1, time-series, single-subject, and within subject designs; descriptive, explanatory, observational designs are also know as quantitative non-experimental designs.
2. Paper ID – This is to identify each paper and cross-reference it with the critical appraisal form so that scores from each appraiser can be compared. The Paper ID is made up of the first or main author's surname and the year the paper was published. Please ensure that the Paper ID on the paper you are reading corresponds with the Paper ID on the critical appraisal form where you enter the scores.

Scoring

The appraisal form is divided into eight categories and 22 items. An item has multiple parts which describe the item and make it easier to appraise and score a category. Each category receives its own score on a 6 point scale from 0–5. A score of 0 is the lowest score a category can achieve, while a score of 5 is the highest.

In the appraisal form, there are tick boxes (☐) beside descriptions of items. The tick box is useful to indicate if the item descriptor is:

- Present (☑) – For an item descriptor to be marked as present, there should be evidence of it being present rather than an assumption of presence.
- Absent (☒) – For an item descriptor to be marked as absent, it is implied that it should be present in the first place.
- Not applicable (■) – For an item descriptor to be marked as not applicable, the item descriptor must not be relevant given the characteristics of the paper being appraised and is, therefore, not considered when assigning a score to a category.

Whether an item descriptor is present, absent, or not applicable is further explored in the section *Categories and items*.

While it may be tempting to add up all the present marks (☑) and all the absent marks (☒) in each category and to use the proportion of one to the other to calculate the score for the category, this is strongly discouraged. It is strongly discouraged because not all item descriptors in a category are of equal importance. For example, in the *Introduction* category there are two items (*Background* and *Objective*) and a total of five tick boxes. If a paper being appraised has all boxes marked as present (☑) except for *Primary objective(s), hypothesis(es), or aim(s)*, should the paper be scored 4/5 for that category? It could be argued that a research paper without a primary objective, hypothesis, or aim is fundamentally flawed and, as a result, should be scored 0/5 even though the other four tick boxes were marked as present.

Therefore, the tick marks for present, absent, or not applicable are to be used as a guide to scoring a category rather than as a simple check list. It is up to the appraiser to take into consideration all aspects of each category and, based on both the tick marks and judgement, assign a score to the category.

Similarly, the research design used in each paper should be appraised on its own merits and not relative to some preconceived notion of a hierarchy of research designs. What is most important is that the paper used an appropriate research

design based on the research question it was addressing, rather than what research design in itself was used.

Finally, it is not the purpose of this tool to present a single score upon which an overall assessment of a paper can be made. Just like not all item descriptors are of equal importance, neither are all categories the same. Categories and as an extension all scores are dissimilar, not equivalent, and cannot be added:

4. Each category is designed to be separate from every other category, while items within each category are as similar as possible. As a result, scores from each category are dissimilar.
5. The scores are ordinal or rank-order scales and because categories are dissimilar, a specific category scoring X is not necessarily the same as another category scoring X. That is, scores are not equivalent.
6. As a result of scores being dissimilar and not equivalent, scores cannot be added. For example, if you collected information on a person, such as how they rate a book, a movie, and a night club on a 5-star rating system, it would not make much sense to add these data together. However, the data can still be used to build a picture of the individual. In the same way, it does not make sense to add together the scores for the *Introduction* and *Discussion* categories or any other combination of categories. However, the data can be used to build up a picture of the paper being appraised.

Categories and items

There were no changes made to Categories or Items between evaluation of validity and this study. The explanation of Categories and Items can be seen in section 5.8.3.3.

6.9.2 List of papers used for testing reliability

True experimental

Arts, M. P., Brand, R., van den Akker, E. M., Koes, B. W., Bartels, R. H., & Peul, W. C. (2009). Tubular diskectomy vs conventional microdiskectomy for sciatica: A randomized controlled trial. *JAMA*, *302*(2), 149-158. doi:10.1001/jama.2009.972

van Gils, E. J. M., Veenhoven, R. H., Hak, E., Rodenburg, G. D., Bogaert, D., IJzerman, E. P., Bruin, J., et al. (2009). Effect of reduced-dose schedules with 7-valent pneumococcal conjugate vaccine on nasopharyngeal pneumococcal carriage in children: A randomized controlled trial. *JAMA*, *302*(2), 159-167. doi:10.1001/jama.2009.975

Lobo, S. M., Salgado, P. F., Castillo, V. G., Borim, A. A., Polachini, C. A., Palchetti, J. C., Brienzi, S. L., et al. (2000). Effects of maximizing oxygen delivery on morbidity and mortality in high-risk surgical patients. *Critical Care Medicine*, *28*(10), 3396-3404.

The TADS Team. (2007). The treatment for adolescents with depression study (TADS): Long-term effectiveness and safety outcomes. *Archives of General Psychiatry*, *64*(10), 1132-1143.

Quasi-experimental

Bergman-Evans, B. (2004). Beyond the basics: Effects of the eden alternative model on quality of life issues. *Journal of Gerontological Nursing*, *30*(6), 27-34.

Kaunonen, M. P., Tarkka, M. P., Laippala, P., & Paunonen-Ilmonen, M. P. (2000). The impact of supportive telephone call intervention on grief after the death of a family member. *Cancer Nursing*, *23*(6), 483-491.

Mignone, J., & Guidotti, T. L. (1999). Support groups for injured workers: Process and outcomes. *Journal of Occupational and Environmental Medicine*, *41*(12), 1059-1064.

Polanczyk, G., Zeni, C., Genro, J. P., Guimaraes, A. P., Roman, T., Hutz, M. H., & Rohde, L. A. (2007). Association of the adrenergic α2A receptor gene with methylphenidate improvement of inattentive symptoms in children and adolescents with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *64*(2), 218-224.

Single system

Behari, S., Nayak, S. R., Bhargava, V., Banerji, D., Chhabra, D. K., & Jain, V. K. (2003). Craniocervical tuberculosis: Protocol of surgical management. *Neurosurgery*, *52*(1), 72-81.

Gosain, A. K., Santoro, T. D., Havlik, R. J., Cohen, S. R., & Holmes, R. E. (2002). Midface distraction following Le Fort III and Monobloc osteotomies: Problems and solutions. *Plastic & Reconstructive Surgery*, *109*(6), 1797-1808.

Jais, P., Haissaguerre, M., Shah, D. C., Chouairi, S., Gencel, L., Hocini, M., & Clementy, J. (1997). A focal source of atrial fibrillation treated by discrete radiofrequency ablation. *Circulation*, *95*(3), 572-576.

Smith, A., Lew, R., Shrimpton, C., Evans, R., & Abbenante, G. (2000). A novel stable inhibitor of endopeptidases EC 3.4.24.15 and 3.4.24.16 potentiates bradykinin-induced hypotension. *Hypertension*, *35*(2), 626-630.

Descriptive, exploratory, observational

Alexander, J. M., McIntire, D. M., & Leveno, K. J. (1999). Chorioamnionitis and the prognosis for term infants. *Obstetrics & Gynecology*, *94*(2), 274-278.

Bhattacharyya, N., & Fried, M. P. (2003). The accuracy of computed tomography in the diagnosis of chronic rhinosinusitis. *Laryngoscope*, *113*(1), 125-129.

Cournot, M., Marquie, J. C., Ansiau, D., Martinaud, C., Fonds, H., Ferrieres, J., & Ruidavets, J. B. (2006). Relation between body mass index and cognitive function in healthy middle-aged men and women. *Neurology*, *67*(7), 1208-1214. doi:10.1212/01.wnl.0000238082.13860.50

Whalen, C. C., Nsubuga, P., Okwera, A., Johnson, J. L., Hom, D. L., Michael, N. L., Mugerwa, R. D., et al. (2000). Impact of pulmonary tuberculosis on survival of HIV-infected adults: a prospective epidemiologic study in Uganda. *AIDS*, *14*(9), 1219-1228.

Qualitative

Beck, C. (1996). Postpartum depressed mothers' experiences interacting with their children. *Nursing Research*, *45*(2), 98-104.

Goldbort, J. G. (2009). Women's lived experience of their unexpected birthing process. *MCN: The American Journal of Maternal Child Nursing*, *34*(1), 57-62. doi:10.1097/01.NMC.0000343867.95108.b3

Hinck, S. M. (2007). The meaning of time in oldest-old age. *Holistic Nursing Practice*, *21*(1), 35-41.

Meighan, M., Davis, M., Thomas, S., & Droppleman, P. (1999). Living with postpartum depression: The father's experience. *MCN: The American Journal of Maternal Child Nursing*, *24*(4), 202-208.

Systematic review

Bagshaw, S. M., Berthiaume, L. R., Delaney, A., & Bellomo, R. (2008). Continuous versus intermittent renal replacement therapy for critically ill patients with acute kidney injury: A meta-analysis. *Critical Care Medicine*, *36*(2), 610-617. doi:10.1097/01.CCM.0B013E3181611F552

Mellegers, M. A., Furlan, A. D., & Mailis, A. (2001). Gabapentin for neuropathic pain: Systematic review of controlled and uncontrolled literature. *Clinical Journal of Pain*, *17*(4), 284-295.

Saposnik, G., & Del Brutto, O. H. (2003). Stroke in South America: A systematic review of incidence, prevalence, and stroke subtypes. *Stroke*, *34*(9), 2103-2107.

Singh, S., & Kumar, A. (2007). Wernicke encephalopathy after obesity surgery: A systematic review. *Neurology*, *68*(11), 807-811. doi:10.1212/01.wnl.0000256812.29648.86

6.9.3. Worksheet function and decision table

Microsoft Excel worksheet function

```
'Walter, Eliasziw, Donner (1998) Sample size and optimal designs
for reliability studies. Statistics in Medicine, 17(1), 101-110.
Sample size based on H (theta)
Function rSampleSizeH(Observations As Integer, EstR As Single, MinR
     As Single, zScore As Single, Power As Single)

   Dim MinTheta, EstTheta, Critical, Top, Bottom As Double

   'Make sure inputs are in the correct range
   If Observations < 2 Then
      rSampleSizeH = "Observations must be integer, >=2"
   ElseIf EstR >= 1 Or EstR <= 0 Then
      rSampleSizeH = "Est reliability must be >0 and <1"
   ElseIf MinR >= 1 Or MinR < 0 Then
      rSampleSizeH = "Min reliability must be >=0 and <1"
   ElseIf MinR >= EstR Then
      '1-tailed test, EstR must be > MinR
      rSampleSizeH = ""
   Else
      'Calculation
      MinTheta = MinR / (1 - MinR)
      EstTheta = EstR / (1 - EstR)
      Critical = (1 + (Observations * MinTheta)) / (1 + (Observations
           * EstTheta))
      Top = 2 * ((zScore + Power) ^ 2) * Observations
      Bottom = ((Log(Critical)) ^ 2) * (Observations - 1)
      'Round up [WorksheetFunction.RoundUp()] final answer
           [1+(Top/Bottom] to the nearest integer [,0]
      rSampleSizeH = WorksheetFunction.RoundUp(1 + (Top / Bottom), 0)
   End If
End Function
```

Decision table

	Min *r*	0.40		***z*$_\alpha$**	1.65		**1–*β***	0.85			
n	**Estimate of *r***										
	0.95	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**	**0.60**	**0.55**	**0.50**	**0.45**
2	5	8	11	16	24	36	56	94	178	426	1,798
3	4	6	8	11	15	22	34	56	104	244	1,011
4	3	5	7	9	13	18	27	44	81	187	765
5	3	4	6	8	11	16	24	38	70	160	646
6	3	4	6	8	11	15	22	35	63	143	577
7	3	4	6	7	10	14	21	33	59	133	532
8	3	4	5	7	10	13	20	31	56	125	501
9	3	4	5	7	9	13	19	30	53	120	477
10	3	4	5	7	9	13	18	29	52	116	459

n, number of observations

Chapter 7 – Compare CAT with no CAT

The purpose of this chapter was originally to compare the proposed critical appraisal tool (CAT) with no CAT when appraising research papers because there was little evidence of whether a CAT affected how research papers are appraised (Objective 6). However, due to questions raised during the reliability study, two more points have been added to this part of the research: (1) does subject matter knowledge affect appraisal of research papers?; and (2) does research design knowledge affect appraisal of research papers?

The chapter consists of an article accepted for publication on 11 August 2011 and published December 2011 (Appendix C.6):
Crowe, M., Sheppard, L. & Campbell, A. (2011). A comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: A randomised trial. *International Journal of Evidence-Based Healthcare*, *9*(4), 444-449. doi:10.1111/j.1744-1609.2011.00237.x

Changes have been made to the submitted article to ensure consistency. In the event that copyright permission may be required for this article, it can be found in Appendix A.4.

A comparison of the effects of using the proposed critical appraisal tool versus informal appraisal in assessing health research: A randomised trial

7.1 ABSTRACT

Background – In systematic reviews, evidence-based practice, and journal clubs critical appraisal tools are used to rate research papers. Little evidence exists on whether a critical appraisal tool, subject matter knowledge, or research design knowledge affect the appraisal of research papers. The aim was compare the proposed critical appraisal tool (PCAT) with an informal appraisal (IA) of research papers, where raters indicate their level of subject matter knowledge and research design knowledge.

Methods – 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 the IA group and the PCAT group. Each participant independently appraised five research papers using informal appraisal or the proposed CAT.

Results – The intraclass correlation coefficient (ICC) for absolute agreement was 0.76 for the IA group and 0.88 for the PCAT group (a difference of 0.12). The G study showed that in the IA group 24% of variance in scores was attributable to either the rater or paper × rater (pr) interactions whereas this was 12% in the PCAT group. Analysis of covariance showed that there were significant results in the IA 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 $\eta^2 = 0.50$).

Discussion – The proposed CAT was more reliable than an informal appraisal of research papers. In the IA group there were significant effects for rater and subject matter knowledge, whereas the proposed CAT almost eliminated the rater effect and no subject matter knowledge effect was apparent. There was no research design knowledge effect in either group.

Conclusions – The proposed CAT provided much better score reliability and should help readers with different levels and types of knowledge to reach similar conclusions about a health research paper.

7.2 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 (CASP) tools [5], Assessment of Multiple Systematic Reviews (AMSTAR) [6], and Single-Case Experimental Design (SCED) 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 which use a diverse range of qualitative and quantitative research designs or which 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 which are incompatible. Second, the majority of CATs lack the depth to fully appraise research [8, 9] or have scoring systems which are insufficient to accurately reflect the content of research papers [10, 11, 12]. In 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 proposed CAT [1, 15, 16] was designed to overcome the problems outlined above. First, the proposed CAT was built based on a review of the design evidence for 45 critical appraisal tools 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, 20, 21, 22, 23, 24], and research methods theory [25, 26, 27]. This analysis led to the development of a tool which 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 to help readers through the appraisal process. 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 a score reliability study [16] were completed for the proposed CAT. These preliminary assessments showed that the scores obtained had a reasonable degree of validity and the proposed CAT could be considered a reliable means of appraising health research in a wide range of research designs. The proposed CAT and user guide, as used in this chapter, are available in section 7.9.2, p. 199).

However, while undertaking previous research into the proposed CAT, Chapters 5 and 6 [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 which is new to or contradicts with a reader's knowledge may be discarded, no matter how worthy.

Teaching and implementation of evidence based practice (EBP) 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 affected how readers appraise a sample of health research papers.
2. To examine whether subject matter knowledge or research design knowledge affected how readers appraise a sample of health research papers.

7.3 METHODS

The CAT used in the study was the proposed CAT. The proposed CAT was used because it was known to the author; score validity and reliability data were available; and the proposed CAT 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 proposed CAT or informal appraisal.

7.3.1 Design

Participants were match paired by the author 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 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 used to match participants rather than as a conclusive measure of researcher experience.

When all participants had been match paired, they were randomly assigned by the author to either the informal appraisal group (IA group) or the proposed CAT group (PCAT group), using the random sequence generator available from RANDOM.ORG [29]. The author was not blinded to the groups participants were allocated. Blinding was not considered necessary because participants individually scored papers without input from the author. Participants were informed that they could contact the author if they had any general questions regarding the study. However, questions concerning how to score a research paper, whether using the proposed CAT 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 those participants, until the study was completed.

7.3.2 Sampling

A sample size calculation based on the work of Walter, Eliasziw, & Donner was used [30]. The formulae for the sample size were:

$$k = 1 + \frac{2(z + (1-\beta))^2 n}{(\ln C_0)^2 (n-1)}$$

Where

$$C_0 = \frac{1 + \left(n\left(\frac{r_{min}}{1 - r_{min}}\right)\right)}{1 + \left(n\left(\frac{r}{1-r}\right)\right)}$$

k, number of raters
z, z-score
$1 - \beta$, power
n, number of papers
ln, natural log
r, required reliability coefficient
r_{min}, minimum acceptable reliability coefficient

A Microsoft Excel worksheet function of the formulae and a decision table were developed (section 7.9.1, p. 196). Based on these calculations, a sample size of six raters reading five papers each was required to achieve an intraclass correlation coefficient (ICC) of 0.90 (α = 95%, 1–β = 0.79, r_{min} = 0.55. 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 RANDOM.ORG [29]. The research papers were selected from a larger pool of papers that was used in Chapters 5 and 6 [15, 16]. In brief, the larger pool of research papers was randomly selected from OvidSP's (Ovid, New York) 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, M. P., Brand, R., van den Akker, E. M., Koes, B. W., Bartels, R. H., & Peul, W. C. (2009). Tubular diskectomy vs conventional microdiskectomy for sciatica: A randomized controlled trial. *JAMA*, *302*(2), 149-158.
2. Quasi-experimental: Polanczyk, G., Zeni, C., Genro, J. P., Guimaraes, A. P., Roman, T., Hutz, M. H., & Rohde, L. A. (2007). Association of the adrenergic α2A receptor gene with methylphenidate improvement of inattentive

symptoms in children and adolescents with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *64*(2), 218-224.

3. Single system: Jais, P., Haissaguerre, M., Shah, D. C., Chouairi, S., Gencel, L., Hocini, M., & Clementy, J. (1997). A focal source of atrial fibrillation treated by discrete radiofrequency ablation. *Circulation*, *95*(3), 572-576.
4. Qualitative: Beck, C. (1996). Postpartum depressed mothers' experiences interacting with their children. *Nursing Research*, *45*(2), 98-104.
5. Systematic review: Singh, S., & Kumar, A. (2007). Wernicke encephalopathy after obesity surgery: A systematic review. *Neurology*, *68*(11), 807-811.

7.3.3 Data collection

Potential participants were asked to take part in the study through a series of invitations emailed to academic and research staff, and post-graduate 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. 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. For the IA group (section 7.9.2, p. 196), 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 PCAT group (section 7.9.3, p. 199), the participants were asked to read each paper thoroughly and to fill out a proposed CAT form for each paper. The proposed CAT 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 a self assessed and reported scale from 0 (no knowledge) to 5 (extensive knowledge).

7.3.4 Data analysis

When the appraisal forms were returned, the total scores for the PCAT group were checked by adding the individual category scores. Total scores for the research papers in the IA and PCAT groups were then converted to percentage scores so that the rating of papers could be compared. Reliability of scores was calculated using the intraclass correlation coefficient (ICC) and generalizability 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.

7.3.5 Ethics

Ethical approval for this study was obtained from James Cook University Human Ethics Committee (No. H3415) and the study conformed to the Declaration of Helsinki [31]._ENREF_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. There were no conflicts of interest or funding sources to declare.

7.4 RESULTS

A total of 19 people responded to the invitation to participate in the study. Ten of these participants completed the study. Despite numerous efforts to attract further

participants to the study, no other participants were found. Based on responses to the questionnaire, four participants (2 pairs) had a low level of research experience, four participants (2 pairs) had a medium level of research experience, and two participants (1 pair) had a high level of research experience. The flow of participants through the study is indicated in Figure 7.1.

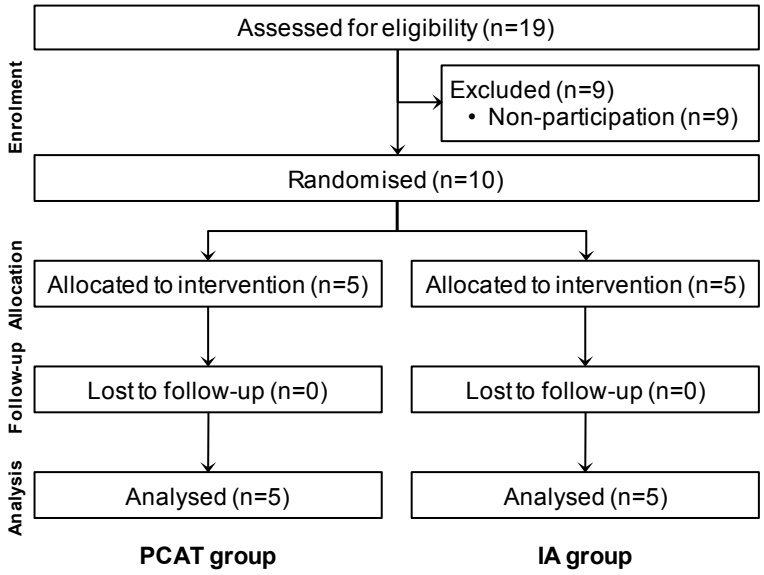

Figure 7.1 Flow of participants

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 PCAT 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.

Subject matter knowledge for both the IA group and the PCAT group were positively skewed (many 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 (using Mann-Whitney U) between the IA group and PCAT group for subject matter knowledge (U = 298.5,

z = -0.29, p = 0.78 1-tailed) or research design knowledge (U = 270.0, z = -0.85, p = 0.40 1-tailed).

Reliability, based on total score %, was calculated in SPSS version 18.02 (SPSS, Chicago, IL) using the intraclass correlation coefficient (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 7.1, Part 1). The PCAT group had an ICC for consistency of 0.89 and for absolute agreement of 0.88.

A generalizability theory G study (Table 7.1, Part 2), using G_String_III [32], demonstrated where error occurred in the total scores. The IA group had 76% of variance attributable to the paper (p), 10% attributable to the rater (r), and 14% attributable to paper × rater interaction (pr). The PCAT group had 88% of variance attributable to the paper (p), 1% attributable to the rater (r), and 11% attributable to paper × rater interaction (pr). Taking an $a\ priori$ minimum acceptable G coefficient of 0.75, a D study (Table 7.1, Part 3) 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 PCAT group, two raters would be required to achieve both the relative and absolute G coefficients.

Analysis of covariance – ANCOVA – (Table 7.2) 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 1-tailed, partial η^2 = 0.28) and rater (F(4,18) = 4.57, p < 0.05 1-tailed, partial η^2 = 0.50). There were no significant results for the PCAT group.

Table 7.1 Reliability (total score %, k=5, n=5)

Part 1 ICC (Intraclass correlation coefficient)

	IA group	PCAT group
Consistency	0.84	0.89
Absolute agreement	0.76	0.88

Part 2 G study

	IA group	PCAT group
Paper (*p*)	76	88
Rater (*r*)	10	1
Paper × Rater (*pr*)	14	11

Part 3 D study

k	IA group		PCAT group	
	Eρ²	Φ	Eρ²	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, Number of raters per group;
n, Number of papers per rater;
IA, informal appraisal;
PCAT, proposed critical appraisal tool;
Eρ², relative error G coefficient;
Φ, absolute error G coefficient.

Table 7.2 Analysis of covariance

Main effects	IA group				PCAT group			
	F	Sig	Part η^2	*f*	*F*	Sig	Part η^2	*f*
Knowledge subject matter (*df* 1,18)	7.03	0.02	0.28	0.63	0.33	0.57	0.02	0.14
Knowledge research design (*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

IA, informal appraisal; PCAT, proposed critical appraisal tool; *F*, *F* statistic;
Sig, significance, α 0.05 1-tailed; Part η^2, partial eta squared; *f*, effect size;
df, degrees of freedom.

7.5 DISCUSSION

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

With regards to reliability, it was expected that the scores from the PCAT 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 PCAT 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 for the IA group. A confidence interval was not calculated for the difference between ICCs because there is no agreed computation method [33, 34, 35].

Furthermore, the proposed CAT almost eliminated the rater effect (variance in total scores due to variability in how a rater scored a paper), with the PCAT 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 proposed CAT than using informal appraisal especially where absolute agreement was sought (2 versus 5 raters).

ANCOVA for the IA group showed that there was a significant subject matter knowledge effect ($f = 0.63$). 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. 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.

The G study, ANCOVA, and D study results show that using the proposed CAT 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 proposed CAT 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.

7.6 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 proposed CAT 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 proposed CAT 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 proposed CAT may in turn lead to improved understanding of findings and application of the evidence.

7.7 IN SUMMARY

- Little is known on whether using a CAT or not using a CAT, subject matter knowledge, and research design knowledge affect the appraisal of research papers.
- A match-paired randomised trial was conducted to explore this question.
- The results showed that the proposed CAT prevented research appraisal from being affected by the rater and subject matter knowledge.
- The next chapter brings together the thesis conclusions.

7.8 REFERENCES

1. Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

2. Khan, K. S., ter Riet, G., Glanville, J., Sowden, A. J., & Kleijen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews (CRD Report 4). York, England: University of York.

3. Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12. doi:10.1016/0197-2456(95)00134-4

4. de Vet, H. C. W., de Bie, R. A., van der Heijden, G. J. M. G., Verhagen, A. P., Sijpkes, P., & Knipschild, P. G. (1997). Systematic reviews on the basis of methodological criteria. *Physiotherapy, 83*(6), 284-289. doi:10.1016/S0031-9406(05)66175-5

5. NHS Public Health Resources Unit. (2010). CASP: Critical Appraisal Skills Programme. Retrieved 29 January 2011, from http://www.sph.nhs.uk/what-we-do/public-health-workforce/resources/critical-appraisals-skills-programme/

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

7. Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18*(4), 385-401. doi:10.1080/09602010802009201

8. Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitc, C., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27). doi:10.3310/hta7270

9. Moyer, A., & Finney, J. W. (2005). Rating methodological quality: Toward improved assessment and investigation. *Accountability in Research, 12*(4), 299-313. doi:10.1080/08989620500440287

10. Heller, R. F., Verma, A., Gemmell, I., Harrison, R., Hart, J., & Edwards, R. (2008). Critical appraisal for public health: A new checklist. *Public Health, 122*(1), 92-98. doi:10.1016/j.puhe.2007.04.012

11. Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1060. doi:10.1001/jama.282.11.1054

12. Kuper, A., Lingard, L., & Levinson, W. (2008). Critically appraising qualitative research. *BMJ, 337*(7671), 687-689. doi:10.1136/bmj.a1035

13. Burnett, J., Kumar, S., & Grimmer, K. (2005). Development of a generic critical appraisal tool by consensus: Presentation of first round Delphi survey results. *Internet Journal of Allied Health Sciences and Practice, 3*(1), 22. Retrieved from http://ijahsp.nova.edu/

14. Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

15. Crowe, M., & Sheppard, L. (2011). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies, 48*(12), 1505-1516. doi:10.1016/j.ijnurstu.2011.06.004

16. Crowe, M., Sheppard, L., & Campbell, A. (2011). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology,* (Online). doi:10.1016/j.jclinepi.2011.08.006

17. Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity, 36*(4), 391-409. doi:10.1023/A:1020909529486

18. Dye, J. F., Schatz, I. M., Rosenberg, B. A., & Coleman, S. T. (2000). Constant comparison method: A kaleidoscope of data. *The Qualitative Report, 4*(1/2). Retrieved from http://www.nova.edu/ssss/QR/

19. Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet, 354*(9193), 1896-1900. doi:10.1016/S0140-6736(99)04149-5

20. Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA, 285*(15), 1992-1995. doi:10.1001/jama.285.15.1992

21. Ogrinc, G., Mooney, S. E., Estrada, C., Foster, T., Goldmann, D., Hall, L. W., ... Watts, B. (2008). The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: Explanation and elaboration. *Quality and Safety in Health Care, 17*(Supplement 1), i13-i32. doi:10.1136/qshc.2008.029058

22. Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., ... Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA, 283*(15), 2008-2012. doi:10.1001/jama.283.15.2008

23. 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*(6), 349-357. doi:10.1093/intqhc/mzm042

24. von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine, 4*(10), e296. doi:10.1371/journal.pmed.0040296

25. Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

26. Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

27. Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

28. MacAuley, D., McCrum, E., & Brown, C. (1998). Randomised controlled trial of the READER method of critical appraisal in general practice. *BMJ, 316*(7138), 1134-1137.

29. Haadr, M. (2009). Random.org: Random sequence generator. Retrieved 29 January 2011, from http://www.random.org/sequences/

30. Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*(1), 101-110. doi:10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e

31. World Medical Association. (2008). *Declaration of Helsinki: Ethical principles for medical research involving human subjects*. Paper presented at the 59th World Medical Association General Assembly, Seoul, South Korea. Retrieved from http://www.wma.net/e/policy/b3.htm

32. Bloch, R. (2010). G_String_III (Version 5.4.6). Hamilton, ON: Programme for Educational Research and Development. Retrieved from http://fhsperd.mcmaster.ca/g_string/

33. Burch, B. D. (2011). Assessing the performance of normal-based and REML-based confidence intervals for the intraclass correlation coefficient. *Computational Statistics & Data Analysis, 55*(2), 1018-1028. doi:16/j.csda.2010.08.007

34. Newcombe, R. G. (2011). Propagating imprecision: Combining confidence intervals from independent sources. *Communications in Statistics, 40*(17), 3154-3180. doi:10.1080/03610921003764225

35. Ramasundarahettige, C. F. F., Donner, A., & Zou, G. Y. (2009). Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Statistics in Medicine, 28*(7), 1041-1053. doi:10.1002/sim.3523

7.9 ADDITIONAL MATERIAL

7.9.1 Worksheet function and decision table

Microsoft Excel worksheet function

```
'Walter, Eliasziw, Donner (1998) Sample size and optimal designs
for reliability studies. Statistics in Medicine, 17(1), 101-110.
Sample size based on H (theta)
Function rSampleSizeH(Observations As Integer, EstR As Single, MinR
     As Single, zScore As Single, Power As Single)

  Dim MinTheta, EstTheta, Critical, Top, Bottom As Double

  'Make sure inputs are in the correct range
  If Observations < 2 Then
    rSampleSizeH = "Observations must be integer, >=2"
  ElseIf EstR >= 1 Or EstR <= 0 Then
    rSampleSizeH = "Est reliability must be >0 and <1"
  ElseIf MinR >= 1 Or MinR < 0 Then
    rSampleSizeH = "Min reliability must be >=0 and <1"
  ElseIf MinR >= EstR Then
    '1-tailed test, EstR must be > MinR
    rSampleSizeH = ""
  Else
    'Calculation
    MinTheta = MinR / (1 - MinR)
    EstTheta = EstR / (1 - EstR)
    Critical = (1 + (Observations * MinTheta)) / (1 + (Observations
         * EstTheta))
    Top = 2 * ((zScore + Power) ^ 2) * Observations
    Bottom = ((Log(Critical)) ^ 2) * (Observations - 1)
    'Round up [WorksheetFunction.RoundUp()] final answer
         [1+(Top/Bottom] to the nearest integer [,0]
    rSampleSizeH = WorksheetFunction.RoundUp(1 + (Top / Bottom), 0)
  End If
End Function
```

Decision table

Min *r*	0.55		z_α	1.65		**1–*β***	0.79
n	**Estimate of *r***						
	0.95	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.65**
2	6	10	16	27	49	97	243
3	4	7	11	18	32	63	153
4	4	6	10	16	27	52	125
5	4	6	9	14	24	46	110
6	4	5	8	13	22	43	102
7	3	5	8	13	21	41	96
8	3	5	8	12	21	39	92
9	3	5	8	12	20	38	90
10	3	5	8	12	20	37	87

n, number of observations

7.9.2 Appraisal materials for IA group

Instructions

1. Each paper has two pieces of information:
   a. Paper ID – This is to identify each paper so that scores from each appraiser can be compared. The Paper ID is made up of the first or main author's surname and the year the paper was published. Please ensure that the Paper ID on the paper you have read corresponds with the Paper ID on the form where you write the paper's rank order.
   b. Research design – To make appraising each paper a little easier, the research design used is written on the paper and in each section below. This means that you do not need to decide which research design was used and can concentrate on appraising the paper based on the research design indicated. Two of the research designs have alternative names that you may be more familiar with: Single system designs include n-of-1, time-series, single-subject, and within subject designs; Descriptive, explanatory or observational designs are also known as quantitative non-experimental designs.
2. Read each paper thoroughly.
3. Having read a paper:
   a. Write a score, from 0 (the lowest score) and 10 (the highest score), in the "Score" box based how good you think the paper covered the topic discussed
   b. Each research design should be appraised on its own merits, not to a 'gold standard'
   c. Scores are whole numbers only (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
   d. If in doubt use your best judgement, there is no right or wrong answer.
4. Please indicate your current level of knowledge for the:
   a. Topic discussed in each paper.
   b. Research design used in each paper.

Appraisal form

Paper ID	**Research Design**	**Score** [out of 10]
Arts 2009	True experimental	

Topic Please circle the appropriate number to indicate your level of knowledge for the *topic* discussed in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Research design Please circle the appropriate number to indicate your level of knowledge for the *research design* used in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Paper ID	**Research Design**	**Score** [out of 10]
Polanczyk 2007	Quasi-experimental	

Topic Please circle the appropriate number to indicate your level of knowledge for the *topic* discussed in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Research design Please circle the appropriate number to indicate your level of knowledge for the *research design* used in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Paper ID	**Research Design**	**Score** [out of 10]
Jais 1997	Single system	

Topic Please circle the appropriate number to indicate your level of knowledge for the *topic* discussed in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Research design Please circle the appropriate number to indicate your level of knowledge for the *research design* used in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Paper ID	**Research Design**	**Score** [out of 10]
Beck 1996	Qualitative	

Topic Please circle the appropriate number to indicate your level of knowledge for the *topic* discussed in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Research design Please circle the appropriate number to indicate your level of knowledge for the *research design* used in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Paper ID	**Research Design**	**Score** [out of 10]
Singh 2007	Systematic review	

Topic Please circle the appropriate number to indicate your level of knowledge for the *topic* discussed in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

Research design Please circle the appropriate number to indicate your level of knowledge for the *research design* used in this paper:

No knowledge — Extensive knowledge

0 1 2 3 4 5

7.9.3 Appraisal materials for PCAT group

Instructions and user guide

Summary

1. Read each paper thoroughly.
2. Each research design should be appraised on its own merits, not to a 'gold standard'.
3. All categories must be scored – it does not matter what research design is used.
   a. Category scores are whole numbers only (0, 1, 2, 3, 4, 5)
   b. The lowest score possible for a category is 0
   c. The highest score possible for a category is 5.
4. Items may be marked ☑ present, ☒ absent, or ■ not applicable.
   a. Tick marks are not a checklist to be totalled – they are a guide to scoring a category.
5. If in doubt use your best judgement, there is no right or wrong answer.
6. Please indicate your current level of knowledge for the:
   a. Topic discussed in each paper
   b. Research design used in each paper.

Introduction

The critical appraisal tool assumes an awful lot. It assumes that the individual using the tool is familiar with research designs, sampling techniques, ethics, data collection methods, and statistical and non-statistical data analysis techniques.

It may be helpful to have a general research methods text available to refer to when appraising papers.

The papers being appraised are unlikely to have the information sought in the sequence outlined in the critical appraisal form. Therefore, it is suggested to read each paper quickly from start to finish, getting an overall sense of what is being discussed. Then re-read the paper and fill in the scores.

Paper ID and Research design

Each paper and each critical appraisal form has two pieces of information at the top:

3. Paper ID – This is to identify each paper and cross-reference it with the critical appraisal form so that scores from each appraiser can be compared. The Paper ID is made up of the first or main author's surname and the year the paper was published. Please ensure that the Paper ID on the paper you are reading

corresponds with the Paper ID on the critical appraisal form where you enter the scores.

4. Research design – To make appraising each paper a little easier, the research design used is written on the paper and on the form. This means that you do not need to decide which research design was used and can concentrate on appraising the paper based on the research design indicated. Two of the research designs have alternative names that you may be more familiar with: Single system designs include n-of-1, time-series, single-subject, and within subject designs; Descriptive, explanatory or observational designs are also known as quantitative non-experimental designs.

Scoring method

The appraisal form is divided into eight categories and 22 items. An item has multiple parts which describe the item and make it easier to appraise and score a category. Each category receives its own score on a 6 point scale from 0–5. A score of 0 is the lowest score a category can achieve, while a score of 5 is the highest.

Categories can only be scored as a whole number or integer, i.e. 0, 1, 2, 3, 4, or 5. Half marks are not allowed.

In the appraisal form, there are tick boxes (☐) beside item descriptors. The tick box is useful to indicate if the item descriptor is:

- Present (☑) – For an item descriptor to be marked as present, there should be evidence of it being present rather than an assumption of presence.
- Absent (☒) – For an item descriptor to be marked as absent, it is implied that it should be present in the first place.
- Not applicable (■) – For an item descriptor to be marked as not applicable, the item descriptor must not be relevant given the characteristics of the paper being appraised and is, therefore, not considered when assigning a score to a category.

Whether an item descriptor is present, absent, or not applicable is further explored in the section *Categories and items*.

All categories must be scored. Although items may be marked ‘not applicable’, all categories are applicable in all research designs.

While it may be tempting to add up all the present marks (☑) and all the absent marks (☒) in each category and to use the proportion of one to the other to calculate

the score for the category, this is strongly discouraged. It is strongly discouraged because not all item descriptors in any category are of equal importance. For example, in the *Introduction* category there are two items (*Background* and *Objective*) and a total of five tick boxes. If a paper being appraised has all boxes marked as present (☑) except for *Primary objective(s), hypothesis(es), or aim(s)*, should the paper be scored 4/5 for that category? It could be argued that a research paper without a primary objective, hypothesis, or aim is fundamentally flawed and, as a result, should be scored 0/5 even though the other four tick boxes were marked as present.

Therefore, the tick marks for present, absent, or not applicable are to be used as a guide to scoring a category rather than as a simple check list. It is up to the appraiser to take into consideration all aspects of each category and, based on both the tick marks and judgement, assign a score to the category.

Similarly, the research design used in each paper should be appraised on its own merits and not relative to some preconceived notion of a hierarchy of research designs. What is most important is that the paper used an appropriate research design based on the research question it was addressing, rather than what research design in itself was used.

Finally, it is not the purpose of this tool to present a single score upon which an overall assessment of a paper can be made. Just like not all item descriptors are of equal importance, neither are all categories the same. Categories and as an extension all scores are dissimilar, not equivalent, and cannot be added.

Level of knowledge

At the end of each form there is a space to indicate your current level of knowledge regarding the topic discussed and the research design used in each paper you have read. Please circle the appropriate number which corresponds to your current level of knowledge in each case.

Scoring categories and items

1. Preliminary

Title

1. Includes study aims and design
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.

Abstract

1. Contains key information
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.
2. Balanced and informative
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.

Text

Note This item can only be assessed when the article has been read in full.

1. Sufficient detail others could reproduce
    - This is an over-arching concept and should be present throughout the study.
2. Clear, concise writing/table(s)/diagram(s)/figure(s)
    - This is an over-arching concept and should be present throughout the study.

2. Introduction

Background

1. Summary of current knowledge
    - Current and applicable knowledge provides a context for the study.
2. Specific problem(s) addressed and reason(s) for addressing
    - Description of why the study was undertaken.
    - Links current knowledge and stated objective(s), hypothesis(es), or aim(s).

Objective

1. Primary objective(s), hypothesis(es), aim(s)
    - The study must have at least one stated objective, hypothesis, or aim.
2. Secondary question(s)
    - Secondary question(s) may sometimes arise based on the primary objective(s), hypothesis(es), or aim(s).
    - Since this is not always the case, a study without secondary questions should not be penalised.

3. Design

Research design

1. Research design(s) chosen and why
    - Description of the research design chosen and why it was chosen.
2. Suitability of research design(s)
    - The research design should be congruent with *Background*, *Objective*, *Intervention(s)/treatment(s)/exposure(s)*, and *Outcome(s)/output(s)/ predictor(s)*.

Intervention, Treatment, Exposure

1. Intervention(s)/treatment(s)/exposure(s) chosen and why
    - Where a study does not normally have an intervention/treatment/exposure, it should not be penalised when none is present.
    - Statement for every intervention/treatment/exposure chosen and why it was chosen.
    - Each intervention/treatment/exposure must be congruent with *Background*, *Objective*, and *Research design*.
2. Precise details of the intervention(s)/treatment(s)/exposure(s) for each group
    - Full details are presented for every intervention/treatment/exposure for every participant/case/group so that other studies could duplicate.
3. Intervention(s)/treatment(s)/exposure(s) valid and reliable
    - A statement of reliability/validation or why there is no validation/reliability for each intervention/treatment/exposure.

Outcome, Output, Predictor, Measure

1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen and why
    - All research has at least one expected outcome/output/predictor/measure.
    - Statement for each outcome/output/predictor/measure chosen and why it was chosen.

- Each outcome/output/predictor/measure must be congruent with *Background*, *Objective*, *Research design*, and *Intervention/treatment/exposure*.

2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s)
   - Full details are presented of every expected outcome/output/predictor/measure for every participant/case/group so that other studies could duplicate.
3. Outcome(s)/output(s)/predictor(s)/measure(s) valid and reliable
   - A statement of reliability/validation or why there is no validation/reliability for each outcome/output/predictor/measure.

Note In some cases the *Outcome(s)/output(s)/predictor(s)/measure(s)* may be similar to or the same as the *Objective(s), hypothesis(es), aim(s)*. However, in most cases to achieve the *Objective(s), hypothesis(es), aim(s)* a series of *Outcome(s) /output(s)/predictor(s)/measure(s)* are required.

Bias, etc.

1. Potential sources of bias, confounding variables, effect modifiers, interactions
   - Identification of potential sources of:
     - Bias – e.g. attrition, detection, experimental, information, interview, observation, performance, rater, recall, selection.
     - Confounding variables or factors – A variable which interferes between the intervention/treatment/exposure and the outcome/output/predictor/measure.
     - Effect modification – A variable which modifies the association between the intervention/treatment/exposure and the outcome/output/predictor/measure.
     - Interaction effects – When various combinations of intervention(s)/treatment(s)/exposure(s) cause different outcome(s)/output(s)/predictor(s)/measure(s).
   - Should be identified, as far as possible, within the *Research design* before data collection begins in order to minimise their effect.
   - See also *Sampling* and *Data collection*.
2. Sequence generation, group allocation, group balance, and by whom
   - In studies where participants/cases are allocated to groups, the methods used should be stated and procedures established before recruitment or data collection begins (e.g. blinding, method used to randomise, allocate to or balance groups).

3. Equivalent treatment of participants/cases/groups
    - Each participant/case/group must be treated equivalently apart from any intervention/treatment/exposure.
    - If participants/cases/groups are not treated equivalently a statement regarding why this was not possible, how this may affect results, and procedures in place for managing participants/cases/groups.
    - See also *Sampling protocol*, *Collection protocol*, and *Participant ethics*.

4. Sampling

Sampling method

1. Sampling method(s) chosen and why
    - Description of the sampling method chosen and why it was chosen.
    - Sampling methods are normally probability or non-probability based.
    - Examples include: Simple random, systematic, stratified, cluster, convenience, representative, purposive, snowball, and theoretical.
    - Also included here is the search strategy used for a systematic review (e.g. databases searched, search terms).
2. Suitability of sampling method
    - The sampling method should be decided and in place before recruitment or data collection begins.
    - The sampling method should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc*.

Sample size

1. Sample size, how chosen, and why
    - Description of the sample size, the method of sample size calculation, and why that method was chosen.
    - Sample size calculations are normally probability or non-probability based.
    - Examples of how calculations can be made include: Accuracy [e.g. confidence interval (α), population or sample variance (s^2, σ^2), effect size or index (ES, d), power (1-β)], analysis, population, redundancy, saturation, and budget.
2. Suitability of sample size
    - The sample size or estimate of sample size, with contingencies, should be described and calculated before recruitment/data collection begins.

- The sample size should be congruent with *Objective*, *Research design*, Intervention/*treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc.*

Note Sample size calculations are not required for systematic reviews, because it is not possible to know the number of papers that will meet the selection criteria, or for *some* single system designs.

Sampling protocol

1. Description and suitability of target/actual/sample population(s)
   - The target/actual/sample population(s) should be described.
   - The target/actual/sample population(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc.*
2. Inclusion and exclusion criteria for participants/cases/groups
   - Inclusion and exclusion criteria should be explicitly stated and established before recruitment/data collection begins.
   - The use of inclusion and exclusion criteria (especially exclusion criteria) should not be used in such a way as to bias the sample.
3. Recruitment of participants/cases/groups
   - Description of procedures for recruitment and contingencies put in place.
   - Recruitment should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Bias etc.*, and other aspects of *Sampling*.
   - See also *Participant ethics*, *Researcher ethics*, and *Collection protocol*.

Note For systematic reviews inclusion and exclusion criteria *only* need to be appraised, because they refer to the parameters used to select papers.

5. Data collection

Collection method

1. Collection method(s) chosen and why
   - Description of the method(s) used to collect data and why each was chosen.
   - In systematic reviews, this refers to how information was extracted from papers, because these are the data collected.
2. Suitability of collection method(s)
   - The data collection method(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, *Bias etc.*, and *Sampling*.

Collection protocol

1. Include date(s), location(s), setting(s), personnel, materials, processes
    - Description of and details regarding exactly how data were collected, especially any factor(s) which may affect *Outcome/output/predictor/measure* or *Bias etc.*
2. Method(s) to ensure/enhance quality of measurement/instrumentation
    - Description of any method(s) used to enhance or ensure the quality of data collected (e.g. pilot study, instrument calibration, standardised test(s), independent/multiple measurement, valid/reliable tools).
    - Also includes any method(s) which reduce or eliminate bias, confounding variables, effect modifiers, interactions which are not an integral part of the *Design* category (e.g. blinding of participants, intervention(s), outcome(s), analysis; protocols and procedures implemented).
    - In qualitative studies, this relates to concepts such as trustworthiness, authenticity, and credibility.
    - See also *Bias etc.*
3. Manage non-participation, withdrawal, incomplete/lost data
    - Description of any method(s) used to manage or prevent non-participation, withdrawal, or incomplete/lost data.
    - These include but are not limited to: Intention to treat analysis (ITT); last observation carried forward (LOCF); follow up (FU), e.g. equal length, adequate, or complete; and, completer analysis, e.g. on-treatment, on-protocol.

6. Ethical matters

Note Some studies may have been conducted before *Ethical matters* were a major point of concern. The research ethics standards of the time may need to be taken into consideration rather than the prevailing standards.

Participant ethics

1. Informed consent, equity
    - All participants must have provided their informed consent.
    - Equity includes, but is not limited to, cultural respect, just and equitable actions, no harm to participants, debriefing, and consideration for vulnerable individuals or groups.
2. Privacy, confidentiality/anonymity
    - The privacy and confidentiality and/or anonymity of participants must be catered for.

- If this is not possible, the informed and written consent of individuals affected must be obtained.

Researcher ethics

1. Ethical approval, funding, conflict(s) of interest
   - A statement of ethical approval from recognised Ethics Committee(s) or Board(s) suitable for the study being undertaken.
   - Any real, perceived, or potential conflict(s) of interest should be stated.
   - All sources of funding should be stated.
2. Subjectivities, relationship(s) with participants/cases
   - Description of how the researcher(s) could have potentially or did affect the outcomes of the study through their presence or behaviour.
   - Includes a description of procedures used to minimise this occurring.
   - See also *Bias etc.*

7. Results

Analysis, Integration, Interpretation method

1. A.I.I. (Analysis/Integration/Interpretation) method(s) for primary outcome(s)/output(s)/predictor(s) chosen and why
   - Description of statistical and non-statistical method(s) used to analyse/integrate/interpret Outcome(s)/output(s)/predictor(s)/measure(s) and why each was chosen.
2. Additional A.I.I. methods (e.g. subgroup analysis) chosen and why
   - Description of additional statistical and non-statistical method(s) used to analyse/integrate/interpret Outcome(s)/output(s)/predictor(s)/measure(s) and why each was chosen.
3. Suitability of analysis/integration/interpretation method(s)
   - The analysis/integration/interpretation method(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor*, *Bias etc.*, *Sampling*, and *Data collection*.

Essential analysis

1. Flow of participants/cases/groups through each stage of research
   - Description of how participants/cases/groups advanced through the study.
   - Explanation of course of intervention/treatment/exposure.
2. Demographic and other characteristics of participants/cases/groups
   - Description of baseline characteristics of participants/cases/groups so this can be integrated into the analysis.

3. Analyse raw data, response rate, non-participation, withdrawal, incomplete/ lost data
    - Unadjusted data should be analysed.
    - There may be differences between those that completed and those that did not complete the study.

Outcome, Output, Predictor analysis

1. Summary of results and precision for each outcome/output/predictor/measure
    - Results summarised with, where possible, an indicator of the precision and effect size of each result for each outcome/output/predictor/measure.
    - Where data are adjusted, make clear what was adjusted and why.
    - Where data are categorised, report of internal and external boundaries.
    - Use of quotations to illustrate themes/findings, privileging of subject meaning, adequate description of findings, evidence of reflexivity.
2. Consideration of benefits/harms, unexpected results, problems/failures
    - Description of all outcomes, not just ones being looked for.
    - Description of differences between planned and actual implementation, and the potential effect on results.
3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes)
    - Exploration of outliners because they may not be anomalous.

8. Discussion

Interpretation

1. Interpretation of results in the context of current evidence and objectives
    - Summarises key results in relation to *Background* and *Objective*.
    - Compare and contrast other research findings.
2. Draw inferences consistent with the strength of the data
    - Do not over or under represent data.
    - Draw inferences based on the entirety of available evidence.
    - See also *Sampling* and *Data collection*.
3. Consideration of alternative explanations for observed results
    - Exploration of reasons for differences between observed and expected.
    - Determines if other factors may lead to similar results.
4. Account for bias, confounding, interactions, effect modifiers, imprecision
    - Discussion on magnitude and direction of Bias etc. and how this may have affected the results.
    - See also *Essential analysis*.

Generalisation

1. Consideration of overall practical usefulness of the study
    - Discussion on practical vs. theoretical usefulness.
2. Description of generalisability (external validity) of the study
    - Dependent on *Design*, *Sampling*, and *Data collection*.

Concluding remarks

1. Highlight study's particular strengths
    - What did the study do well?
2. Suggest steps that may improve future results (e.g. limitations)
    - How could the study have been better?
3. Suggest further studies
    - Where should the next study begin?

Critical appraisal rating form

Category Item	**Item descriptor** [☑ Present; ☒ Absent; ■ Not applicable]	**Score** [0–5]
1. Preliminary		
Title	1. Includes study aims ☐ and design ☐	**Preliminary score**
Abstract	1. Contains key information ☐ 2. Balanced ☐ and informative ☐	
Text (asses this item last)	1. Sufficient detail others could reproduce ☐ 2. Clear/concise writing ☐, table(s) ☐, diagram(s) ☐, figure(s) ☐	
2. Introduction		
Background	1. Summary of current knowledge ☐ 2. Specific problem(s) addressed ☐ and reason(s) for addressing ☐	**Introduction score**
Objective	1. Primary objective(s), hypothesis(es), or aim(s) ☐ 2. Secondary question(s) ☐	
3. Design		
Research design	1. Research design(s) chosen ☐ and why ☐ 2. Suitability of research design(s) ☐	**Design score**
Intervention, Treatment, Exposure	1. Intervention(s)/treatment(s)/exposure(s) chosen ☐ and why ☐ 2. Precise details of the intervention(s)/treatment(s)/exposure(s) ☐ for each group ☐ 3. Intervention(s)/treatment(s)/exposure(s) valid ☐ and reliable ☐	
Outcome, Output, Predictor, Measure	1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen ☐ and why ☐ 2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s) ☐ 3. Outcome(s)/output(s)/predictor(s)/measure(s) valid ☐ and reliable ☐	
Bias, etc.	1. Potential bias ☐, confounding variables ☐, effect modifiers ☐, interactions ☐ 2. Sequence generation ☐, group allocation ☐, group balance ☐, and by whom ☐ 3. Equivalent treatment of participants/cases/groups ☐	
4. Sampling		
Sampling method	1. Sampling method(s) chosen ☐ and why ☐ 2. Suitability of sampling method ☐	**Sampling score**
Sample size	1. Sample size ☐, how chosen ☐, and why ☐ 2. Suitability of sample size ☐	
Sampling protocol	1. Target/actual/sample population(s): description ☐ and suitability ☐ 2. Participants/cases/groups: inclusion ☐ and exclusion ☐ criteria 3. Recruitment of participants/cases/groups ☐	
5. Data collection		
Collection method	1. Collection method(s) chosen ☐ and why ☐ 2. Suitability of collection method(s) ☐	**Data collection score**
Collection protocol	1. Include date(s) ☐, location(s) ☐, setting(s) ☐, personnel ☐, materials ☐, processes ☐ 2. Method(s) to ensure/enhance quality of measurement/instrumentation ☐ 3. Manage non-participation ☐, withdrawal ☐, incomplete/lost data ☐	
6. Ethical matters		
Participant ethics	1. Informed consent ☐, equity ☐ 2. Privacy ☐, confidentiality/anonymity ☐	**Ethical matters score**
Researcher ethics	1. Ethical approval ☐, funding ☐, conflict(s) of interest ☐ 2. Subjectivities ☐, relationship(s) with participants/cases ☐	
7. Results		
Analysis, Integration, Interpretation method	1. A.I.I. method(s) for primary outcome(s)/output(s)/predictor(s) chosen ☐ and why ☐ 2. Additional A.I.I. methods (e.g. subgroup analysis) chosen ☐ and why ☐ 3. Suitability of analysis/integration/interpretation method(s) ☐	**Results score**
Essential analysis	1. Flow of participants/cases/groups through each stage of research ☐ 2. Demographic and other characteristics of participants/cases/groups ☐ 3. Analyse raw data ☐, response rate ☐, non-participation/withdrawal/incomplete/lost data ☐	
Outcome, Output, Predictor analysis	1. Summary of results ☐ and precision ☐ for each outcome/output/predictor/measure 2. Consideration of benefits/harms ☐, unexpected results ☐, problems/failures ☐ 3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes) ☐	
8. Discussion		
Interpretation	1. Interpretation of results in the context of current evidence ☐ and objectives ☐ 2. Draw inferences consistent with the strength of the data ☐ 3. Consideration of alternative explanations for observed results ☐ 4. Account for bias ☐, confounding/effect modifiers/interactions/imprecision ☐	**Discussion score**
Generalisation	1. Consideration of overall practical usefulness of the study ☐ 2. Description of generalisability (external validity) of the study ☐	
Concluding remarks	1. Highlight study's particular strengths ☐ 2. Suggest steps that may improve future results (e.g. limitations) ☐ 3. Suggest further studies ☐	

Topic Please circle the appropriate number to indicate your level of knowledge for the *topic* discussed in this paper:

No knowledge					Extensive knowledge
0	1	2	3	4	5

Research design Please circle the appropriate number to indicate your level of knowledge for the *research design* used in this paper:

No knowledge					Extensive knowledge
0	1	2	3	4	5

Chapter 8 – Conclusion

In Chapter 1, six objectives for this study were outlined. The first two objectives set the scene for the study by showing that:

1. Qualitative and quantitative research were on a continuum of research designs, meaning that qualitative and quantitative research could be assessed in the same critical appraisal tool (CAT) (Chapter 2, [1]).
2. The use of mind maps to outline research methods helped to develop an understanding of the scope and variety in research methods (Chapter 3, [2]).

From there, the remaining four objectives concentrated on the design of the proposed CAT, and evaluation of the validity and reliability of scores. These four objectives were:

3. Critically review the literature on the design of existing CATs and use this information to create a proposed CAT.
4. Refine the initial draft of the proposed CAT, develop a scoring system, and evaluate the validity of scores obtained by the proposed CAT.
5. Examine the reliability of scores obtained by the proposed CAT.
6. Compare structured critical appraisal, using the proposed CAT, with informal appraisal (no CAT) when appraising research papers.

These objectives were specifically devised to reduce or eliminate four criticisms of CATs and thereby achieve the overall aim of the study. The four criticisms of CATs were:

1. Limited research design appraisal.
2. Lack of depth to properly assess research papers.
3. Inappropriate scoring systems.
4. No validity or reliability data.

Outlined in the next two sections are the remaining four objectives and whether the criticisms of CATs have been covered sufficiently. Then the limitations of the study are explored and what future research could add to the work already undertaken. Finally, the conclusions of the study are discussed along with whether the aim of the study was achieved, the aim being:

> To design and evaluate a CAT that can be used across a broad range of qualitative and quantitative health research; has the depth to fully assess research papers; has an appropriate scoring system; and has validity and reliability data available to evaluate the scores obtained by the CAT.

8.1 DESIGN

> *Objective 3* – Critically review the literature on the design of existing CATs and use this information to create the proposed CAT.

> *Criticism 1* – Limited research design appraisal.
> *Criticism 2* – Lack of depth to properly assess research papers.

The first criticism of CATs has been addressed. The critical review (Chapter 4, [3]) specifically ensured that all details in the included papers were incorporated into the proposed CAT. Furthermore, knowledge of reporting guidelines and research

methods (Chapter 3) was used to help create the categories into which the items were divided, and what item descriptors belonged to which item.

This does not mean, of course, that the proposed CAT is comprehensive enough to properly assess health research papers. One criticism of the proposed CAT is that it was developed from CATs which, by the author's own admission, suffered from the same criticisms levelled at other critical appraisal tools. However, the proposed CAT at least is based on evidence for CAT design rather than on subjective or biased assessments of what should be included in such a tool.

There was also a discussion in the evaluation of validity (Chapter 5, [4]) about the possibility that due to the inclusive nature of the design, the proposed CAT could suffer from construct-irrelevant variance rather than having construct underrepresentation. However, there were, and still are, not enough data available to demonstrate construct-irrelevant variance. This is an issue to be addressed in future research, perhaps using item response theory (IRT), as mentioned in Chapter 6.

Overall, the proposed CAT has addressed the criticisms that critical appraisal tools:

1. Can be limited in the research designs appraised.
2. Lack depth to properly assess research papers.

As such, Objective 3 of the study has been achieved.

8.2 EVALUATION

Objective 4 – Refine the initial draft of the potential CAT, develop a scoring system, and evaluate the validity of scores obtained by the potential CAT.

Objective 5 – Examine the reliability of scores obtained by the proposed CAT.

Objective 6 – Compare structured critical appraisal, using the proposed CAT, with informal appraisal (no CAT) when appraising research papers.

Criticism 3 – Inappropriate scoring systems.

Criticism 4 – No validity or reliability data.

Decisions on a scoring system are intertwined in the evaluation of validity (Chapter 5, [4, 5 (pp. 9-11), 6]). This is why the scoring system was considered under evaluation rather than the design of the proposed CAT. It was decided that because each of the categories in the proposed CAT consisted of a unidimensional construct, each category should be scored separately on a six-point scale from 0–5. It was hoped that each research paper and category could be cross checked to appraise papers. However, it became apparent that such a procedure was unmanageable with an exponential rise in the number of cross checks as the number of research papers increased. Fortunately, the total score appeared to be sufficient for score interpretation without impairing precision. The total score was calculated by adding all eight categories, without weightings, and then converting this to a percentage. This was allowable under validity theory where multiple scores from unidimensional constructs can be totalled to create a multidimensional construct score [5 (pp. 9-17)]. The caveat with the scoring system was that scores for each category and the total score should be published in any appraisal of the literature so that weak scores in a category are not hidden.

Beyond developing a scoring system, other aspects of construct validity were also examined. These included the evaluation of test content, response processes, internal structure, relations to other variables, and consequences of testing (Chapter 5, [4, 5, 6]). Much of the research for evaluating construct validity in Chapter 5 involved gathering and analysing data about the relation to other variables. These data showed that the proposed critical appraisal tool was at least

comparable to four of the five alternative tools: Physiotherapy Evidence Database (PEDro, true experimental) [7]; Cho and Bero scale (quasi-experimental and DEO) [8]; Reis et al scale (qualitative) [9]; and Assessment of Multiple Systematic Reviews scale (AMSTAR, systematic reviews) [10]. There was a problem with the fifth tool, Single-case Experimental Design scale (SCED, single system), but this was attributed to the scale's focus on brain impairment treatment research rather than a wider range of single system designs [11]. However, SCED was the only scale available for single system designs that met the selection criteria for scales with which the proposed CAT could be compared (the scale needed to have available validity and reliability data).

When data on the reliability of the proposed CAT scores were analysed, there was no evidence that the appraisal of single system designs was compromised when compared to other research designs (Chapter 6, [12]). In fact, all research designs (Table 6.2) showed high intraclass correlation coefficients (ICC) for the total score % based on each research design (ICC for consistency = 0.64–0.91; ICC for absolute agreement = 0.57–0.73) and all research designs taken together (ICC for consistency = 0.74; ICC for absolute agreement = 0.83). In the G study, each research design also showed a majority paper effect (53–70%), which is exactly what is required from a scale (Table 6.3).

It was pointed out at the time that differences in the scores could be due to different levels of knowledge regarding the subject matter or the research designs used in the papers. Another evaluation of reliability, that compared raters using the proposed CAT with raters using informal appraisal, was undertaken to test this (Chapter 7, [13]). It showed no knowledge of subject matter effect and no research design effect in the proposed CAT. That CAT versus no CAT study showed very high ICCs for the proposed CAT (ICC for consistency = 0.89; ICC for absolute agreement = 0.88) and a large majority paper effect in the G study (88%) based on total score %. A reason

for no subject matter knowledge or research design knowledge effect in the CAT versus no CAT study (Chapter 7) may be that the raters were made aware of their subject matter knowledge and research design knowledge because they had to rate this on the appraisal forms. In the reliability study (Chapter 6), raters did not have to do this. As a result, in the CAT versus no CAT study raters may have been more cognisant of their levels of knowledge, which could have affected the results. A second reason could be that the raters in the second reliability study had five papers to read whereas the raters in the first reliability study had 24 papers to read. Therefore, fatigue, motivation, or time may have been a factor in the results.

Therefore, the proposed CAT has addressed the final two criticisms, which were:

3. Inappropriate scoring systems.
4. No validity or reliability data.

Criticism 3 has been directly countered by the combination of component and summative scoring. This means that individual category and the total score % must be made available for each paper that is scored using the proposed CAT.

On a shallow level, Criticism 4 was addressed by the simple fact that score validity and reliability data are now available for the proposed CAT. On a deeper level, the scores from the proposed CAT showed a good level of construct validity and the reliability data showed that the proposed CAT can match, if not exceed, other tools.

It should be noted, however, that the validity and reliability studies undertaken are considered a starting point for the proposed CAT. Each time a scale, or in this case the proposed CAT, is used, validity and reliability checking should be undertaken by the person administering the scale [5]. This is because validity and reliability are ongoing processes rather than one-off determinants of a scale's veracity.

As such, Objectives 4, 5, and 6 were achieved.

8.3 LIMITATIONS

Minor limitations of the study have been discussed under *Design* and *Evaluation* (sections 8.1 and 8.2). However, the major limitations of this study are that no validity or reliability data exist outside:

1. Health research.
2. Academic staff and post-graduate students from the School of Public Health, Tropical Medicine and Rehabilitation Science; School of Nursing, Midwifery and Nutrition; and School of Medicine and Dentistry at James Cook University, Townsville, Australia.

The limitation that the proposed CAT only has validity and reliability data for health research was planned. The aim of the study was to develop a CAT that could be used in a health research context only. Whether the proposed CAT could be used outside health research cannot be answered without the collection of validity and reliability data within those contexts.

The more significant limitation is that the proposed CAT was only tested by academic staff and post-graduate students in three Schools at James Cook University. Strictly speaking, the validity and reliability data cannot be generalised outside this context, unless the participants, Schools, and University are somehow representative of a larger population of participants, Schools, and Universities. No data are available to indicate whether this is so.

On the other hand, if this study had collected data from a number of health faculties in a range of universities throughout Australia, the proposed CAT could not be generalised outside that context either. This is the nature of measurement scales, which is often overlooked. Validity and reliability data must be collected for each context in which a tool is used. Publishing these data assists other users of the tool

by showing them how the scale has performed in various contexts. In other words, validity and reliability data must be collected each time a scale is used, otherwise the veracity of the scores cannot be claimed [5, 6].

Therefore, although there are two major limitations, these were either planned or, no matter what, additional reliability and validity data would need to be collected one context at a time.

8.4 FUTURE RESEARCH

Further research should be undertaken into the proposed CAT. The main area of investigation is whether the number of item descriptors can be reduced. Users of the proposed CAT have indicated that the number of item descriptors can be intimidating but once they began using the tool it became more easily managed. It was stated earlier that item response theory (IRT) could be used as a means of reducing the number of item descriptors. However, a possibly better alternative is Q methodology (also known as the q-sort method), which examines patterns of assessment through factor analysis [14]. Q methodology suits the critical appraisal of research papers because it is partially a subjective process. Further, by reducing item descriptors, certain items or categories may also become obsolete. The need for the *Preliminaries* and *Introduction* categories have been questioned before in Chapter 5, but there are still not enough data available to decide one way or the other.

Unsurprisingly, collection of more validity and reliability data should continue. The more these types of data are gathered and published, the better an understanding of the proposed CAT can be achieved. Furthermore, if it can be shown that many different raters assess research papers in the same way, then a database of ratings could be made available to future researchers where they would not need to re-

assess research papers that already have data available. In other words, with additional data, it could be possible to use the proposed CAT to criterion-reference health research papers, like PEDro (Physiotherapy Evidence Database) [15], OTseeker (Occupational Therapy Systematic Evaluation of Evidence) [16], PsychBITE (Psychological Database for Brain Impairment Treatment Efficacy) [17], and speechBITE (Speech Pathology Best Interventions and Treatment Efficacy) [18], without being limited to specific professions or research designs.

The final area of research that would bring all the above together would be to create an electronic version of the proposed CAT. An electronic tool would display only those item descriptors suitable to each type of research design and automatically upload the results to a database. The database could be personal, institutional, or available to researchers worldwide.

8.5 Conclusion

The aim of the study was achieved, based on fulfilment of all objectives. A CAT was developed and evaluated, and the proposed CAT:

1. Can be used across a broad range of qualitative and quantitative health research.
2. Has the depth to fully assess research papers.
3. Has an appropriate scoring system.
4. Has score validity and reliability data available.

Therefore, the proposed CAT, published as Appendix F, has been named the Crowe Critical Appraisal Tool (CCAT). The CCAT differs from the proposed CAT because it consists of two pages. The first page is for inserting details of the research paper and was developed from the mind maps created in Chapter 3. The second page of the CCAT is almost identical to the other versions of the proposed CAT, except there is

space to write where information was found on the research paper. Finally, the CCAT user guide has a more general introduction than the other published user guides; otherwise information on the categories, items, and item descriptors is the same.

The procedures used to evaluate the CCAT set a higher standard for current and future CATs. Simply stating that a CAT was once evaluated for 'face' or 'content' or 'construct' validity is not enough. Neither is a bold, nonsensical assertion that a CAT is 'valid' or 'reliable' [4]. Explicit evaluation of validity and reliability is required, and should be sought before and after using a CAT.

Finally, the CCAT should be viewed in relation to two ongoing developments in research. First, systematic reviews are no longer seen as a combination of randomised controlled trials and meta-analysis. Instead, when systematic reviews are properly executed they include multiple sources using a range of research designs [19]. This requires a single CAT to assess the research because if multiple CATs are used the scores obtained cannot be compared. The reason is that the assumptions underlying the different CATs may be incompatible and the CATs may not be measuring the same constructs in the same way, unless a body of evidence can show otherwise. Second, greater emphasis is now placed on using multiple interventions in healthcare, leading to the use of multiple or mixed methods research. This is happening from health professionals in practice [20] through to the World Health Organisation (WHO) in policy [21]. Integrating these types of research into systematic reviews or simply assessing them requires a CAT that can be used across research designs. The CCAT can meet these developments in research and can help assess, understand, and communicate research knowledge.

8.6 REFERENCES

1. Crowe, M., & Sheppard, L. (2010). Qualitative and quantitative research designs are more similar than different. *Internet Journal of Allied Health Sciences and Practice, 8*(4). Retrieved from http://ijahsp.nova.edu/

2. Crowe, M., & Sheppard, L. (2011). Mind mapping research methods. *Quality and Quantity,* (Online First). doi:10.1007/s11135-011-9463-8

3. Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

4. Crowe, M., & Sheppard, L. (2011). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies, 48*(12), 1505-1516. doi:10.1016/j.ijnurstu.2011.06.004

5. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

6. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

7. Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

8. Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *JAMA, 272*(2), 101-104. doi:10.1001/jama.1994.03520020027007

9. Reis, S., Hermoni, D., Van-Raalte, R., Dahan, R., & Borkan, J. M. (2007). Aggregation of qualitative studies - From theory to practice: Patient priorities and family medicine/general practice evaluations. *Patient Education and Counseling, 65*(2), 214-222. doi:10.1016/j.pec.2006.07.011

10. Shea, B., Grimshaw, J., Wells, G., Boers, M., Andersson, N., Hamel, C., ... Bouter, L. (2007). Development of AMSTAR: A measurement tool to assess the

methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*(10). doi:10.1186/1471-2288-7-10

11. Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18*(4), 385-401. doi:10.1080/09602010802009201

12. Crowe, M., Sheppard, L., & Campbell, A. (2011). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology,* (Online). doi:10.1016/j.jclinepi.2011.08.006

13. Crowe, M., Sheppard, L., & Campbell, A. (2011). A comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: A randomised trial. *International Journal of Evidence-Based Healthcare, 9*(4), 444-449. doi:10.1111/j.1744-1609.2011.00237.x

14. Barker, J. H. (2008). Q-methodology: An alternative approach to research in nurse education. *Nurse Education Today, 28*(8), 917-925. doi:10.1016/j.nedt.2008.05.010

15. The George Institute for Global Health. (2011). PEDro: Physiotherapy Evidence Database. Retrieved 29 January 2011, from http://www.pedro.org.au/

16. School of Health and Rehabilitation Sciences University of Queensland. (2011). OTseeker: Occupational Therapy Systematic Evaluation of Evidence. Retrieved 29 January 2011, from http://www.otseeker.com/

17. Rehabilitation Studies Unit, R. R. C. S. (2010). PsycBITE: Psychological Database for Brain Impairment Treatment Efficacy. Retrieved 29 January 2011, from http://www.psycbite.com/

18. University of Sydney, & Speech Pathology Australia. (2010). speechBITE: Speech Pathology Database for Best Interventions and Treatment Efficacy. Retrieved 29 January 2011, from http://www.speechbite.com/

19. Khan, K. S., ter Riet, G., Glanville, J., Sowden, A. J., & Kleijen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance

for those carrying out or commissioning reviews (CRD Report 4). York, England: University of York.

20. O'Cathain, A., Murphy, E., & Nicholl, J. (2008). Multidisciplinary, interdisciplinary, or dysfunctional? Team working in mixed-methods research. *Qualitative Health Research, 18*(11), 1574-1585. doi:10.1177/1049732308325535

21. Allotey, P., Reidpath, D., & Pokhrel, S. (2010). Social sciences research in neglected tropical diseases 1: the ongoing neglect in the neglected tropical diseases. *Health Research Policy and Systems, 8*(32). doi:10.1186/1478-4505-8-32

Appendix A – Copyright permissions

A.1 Nova Southeastern University

Retrieved 3 March 2010, from http://ijahsp.nova.edu/guide.html

Internet Journal of Allied Health Sciences and Practice
Chapter 2 – Qualitative and quantitative research

1) Authorship

All persons designated as authors must meet the criteria for authorship detailed in the copyright release form and as listed in the Friedman article. All authors must sign, date and submit a copy of the copyright release form to the IJAHSP. It may also be faxed toll free or mailed to the journal office.

Note. A manuscript with 6 or more authors attach [sic] a detailed statement of the contribution of each author to the copyright release form.

Please read the Friedman article regarding categories of authorship to which the IJAHSP subscribes.

Please keep in mind, if you are a professor advising on a student paper in a teacher/student relationship, the IJAHSP does not consider this authorship. The professor should be acknowledged, but should not receive authorship credit.

Manuscripts submitted by authors who were employees of the United States federal government at the time the subject of their work was investigated and the piece was written are not subject to the Copyright Act; these authors must inform the Editor of their status as federal employees.

Authors transfer their copyright to the IJAHSP but will not lose the right to reprint material from the articles. Any reprint will be required to acknowledge and credit the Internet Journal of Allied Health Sciences and Practice. If a manuscript is not

accepted, or is withdrawn before publication, the transfer of copyright is null and void. [emphasis added]

A.2 SPRINGER

Retrieved 3 March 2010, from http://www.springer.com/open+access/authors+rights?SGWID=0-176704-12-683201-0

Quality and Quantity
Chapter 3 – Research methods

The copyright to this article is transferred to Springer (respective to owner if other than Springer and for U.S. government employees: to the extent transferable) effective if and when the article is accepted for publication. The author warrants that his/her contribution is original and that he/she has full power to make this grant. The author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors. The copyright transfer covers the exclusive right and license to reproduce, publish, distribute and archive the article in all forms and media of expression now known or developed in the future, including reprints, translations, photographic reproductions, microform, electronic form (offline, online) or any other reproductions of similar nature.

An author may self-archive an author-created version of his/her article on his/her own website and or in his/her institutional repository. He/she may also deposit this version on his/her funder's or funder's designated repository at the funder's request or as a result of a legal obligation, provided it is not made publicly available until 12 months after official publication. He/she may not use the publisher's PDF version, which is posted on www.springerlink.com, for the purpose of selfarchiving or deposit. Furthermore, the author may only post his/her version provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at www.springerlink.com".

Prior versions of the article published on non-commercial pre-print servers like arXiv.org can remain on these servers and/or can be updated with the author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgement needs to be given to the final publication and a link should be inserted to the published article on Springer's website,

accompanied by the text "The final publication is available at springerlink.com". *The author retains the right to use his/her article for his/her further scientific career by including the final published journal article in other publications such as dissertations and postdoctoral qualifications provided acknowledgement is given to the original source of publication. [emphasis added]*

The author is requested to use the appropriate DOI for the article. Articles disseminated via www.springerlink.com are indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia. After submission of the agreement signed by the corresponding author, changes of authorship or in the order of the authors listed will not be accepted by Springer.

A.3 ELSEVIER

Retrieved 3 March 2010, from http://www.elsevier.com/wps/find/authorsview.authors/copyright

Journal of Clinical Epidemiology
Chapter 4 – Review of critical appraisal tool design
Chapter 6 – Reliability study

International Journal of Nursing Studies
Chapter 5 – Evaluation of validity

As a journal author, you retain rights for a large number of author uses, including use by your employing institute or company. These rights are retained and permitted without the need to obtain specific permission from Elsevier. These include:

- the right to make copies (print or electric) of the journal article for their own personal use, including for their own classroom teaching use;
- the right to make copies and distribute copies (including via e-mail) of the journal article to research colleagues, for personal use by such colleagues (but not for Commercial Purposes*, as listed below);
- the right to post a pre-print version of the journal article on Internet web sites including electronic pre-print servers, and to retain indefinitely such version on such servers or sites (see also our information on electronic preprints for a more detailed discussion on these points);

- the right to post a revised personal version of the text of the final journal article (to reflect changes made in the peer review process) on the author's personal or institutional web site or server, incorporating the complete citation and with a link to the Digital Object Identifier (DOI) of the article;
- the right to present the journal article at a meeting or conference and to distribute copies of such paper or article to the delegates attending the meeting;
- for the author's employer, if the journal article is a 'work for hire', made within the scope of the author's employment, the right to use all or part of the information in (any version of) the journal article for other intra-company use (e.g. training), including by posting the article on secure, internal corporate intranets;
- patent and trademark rights and rights to any process or procedure described in the journal article;
- *the right to include the journal article, in full or in part, in a thesis or dissertation; [emphasis added]*
- the right to use the journal article or any part thereof in a printed compilation of works of the author, such as collected writings or lecture notes (subsequent to publication of the article in the journal); and
- the right to prepare other derivative works, to extend the journal article into book-length form, or to otherwise re-use portions or excerpts in other works, with full acknowledgement of its original publication in the journal.

* Commercial Purposes includes the use or posting of articles for commercial gain including the posting by companies or their employee-authored works for use by customers of such companies (e.g. pharmaceutical companies and physician-prescribers); commercial exploitation such as directly associating advertising with such postings; the charging of fees for document delivery or access; or the systematic distribution to others via e-mail lists or list servers (to parties other than known colleagues), whether for a fee or for free.

A.4 WILEY-BLACKWELL

Retrieved 3 March 2010, from http://authorservices.wiley.com/bauthor/faqs_copyright.asp#1.7

International Journal of Evidence-based Healthcare
Chapter 7 – Compare CAT with no-CAT

7. What rights do I retain?

The Contributor or, if applicable, the Contributor's Employer, retains all proprietary rights other than copyright, such as patent rights, in any process, procedure or article of manufacture described in the contribution.

Contributors may re-use unmodified abstracts for any non-commercial purpose. For online use of the abstract, Wiley-Blackwell encourages but does not require linking back to the final published contribution.

Contributors may use the articles in teaching duties and in other works such as theses. [emphasis added]

Contributors may re-use figures, tables, data sets, artwork, and selected text up to 250 words from their contributions, provided the following conditions are met:

- Full and accurate credit must be given to the contribution.
- Modifications to the figures, tables and data must be noted. Otherwise, no changes may be made.
- The reuse may not be made for direct commercial purposes, or for financial consideration to the Contributor.
- Re-use rights shall not be interpreted to permit dual publication in violation of journal ethical practices.

Additional re-use rights are set forth in the actual copyright Agreement.

Appendix B – Ethics approval

A copy of ethics approval (H3415) from the James Cook University, Human Research Ethics Committee is reproduced on the following page.

Administrative documentation
has been removed

Appendix C – Published articles

Copies of published papers are reproduced on the following pages.

C.1 Qualitative and quantitative research designs are more similar than different (Chapter 2)

This article was removed due to copyright restrictions

C.2 Mind mapping research methods (Chapter 3)

This article was removed due to copyright restrictions

C.3 A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed (Chapter 4)

This article was removed due to copyright restrictions

C.4 A general critical appraisal tool: An evaluation of construct validity (Chapter 5)

This article was removed due to copyright restrictions

C.5 Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs (Chapter 6)

This article was removed due to copyright restrictions

C.6 Comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: A randomised trial (Chapter 7)

Appendix D – Material for participants, reliability study

Materials for participants in the *Reliability study* (Chapter 6) are reproduced on the following pages.

D.1 INFORMED CONSENT FORM

Principal investigator	Michael Crowe, MIT, BSc (Mgmt), ADMT, PhD student
Supervisor	Lorraine Sheppard, PhD
Study title	The development of a critical appraisal tool for qualitative and quantitative research in health – A PhD research study
School	Public Health, Tropical Medicine and Rehabilitation Science

I understand the aim of this research study is to investigate the use of critical appraisal tools in health research and to develop a valid, reliable tool which can be used across a broad spectrum of qualitative and quantitative health research so that the content of research can be thoroughly compared when reviewing the literature.

I understand that my participation will involve independently sorting and critically appraising 24 research articles using a critical appraisal tool and contributing to feedback on the process. I agree that the researcher may use the results as described in the information sheet.

I acknowledge that: *[Please tick (✓) appropriate box]*	**Yes**	**No**
Any risks and possible effects of participating in the study have been explained to my satisfaction	☐	☐
Taking part in this study is voluntary and I am aware that I can stop taking part at any time without explanation or prejudice and to withdraw any unprocessed data I have provided	☐	☐
Any information I give will be kept strictly confidential and my name will not be used to identify me with this study	☐	☐
I consent to participate in this study	☐	☐
I consent to provide feedback on the critical appraisal tool	☐	☐

Name _____ **Phone** _____

Please print

Email _____

Please print

Signature _____ **Date** _____

D.2 INFORMATION SHEET

Principal investigator	Michael Crowe, MIT, BSc (Mgmt), ADMT, PhD student
Supervisor	Lorraine Sheppard, PhD
Study title	The development of a critical appraisal tool for qualitative and quantitative research in health – A PhD research study
School	Public Health, Tropical Medicine and Rehabilitation Science

You are invited to take part in a research study about the use of critical appraisal tools in health research and to help develop a valid, reliable tool which can be used across a broad spectrum of qualitative and quantitative health research. Based on this tool, the content of research can be thoroughly compared when reviewing the literature or undertaking a systematic review.

The study is being conducted by Michael Crowe and will contribute to a PhD at James Cook University. Michael's supervisor is Professor Lorraine Sheppard.

If you agree to be involved in the study, you will be asked to independently sort and critically appraise 24 research articles using a critical appraisal tool. The research articles and critical appraisal tool will be supplied to you by the researcher. You will also be asked some questions by the researcher to gather information regarding the tool, such as ease of use and overall impression.

There are no risks associated with the study and taking part in this study is completely voluntary. You can stop taking part in the study at any time without explanation or prejudice. You may also withdraw any unprocessed data from the study. Your responses and contact details are strictly confidential. The data from the study will be used in research publications but you will not be identified in any way in these publications.

If you know of others that might be interested in this study, please pass on this information sheet to them so they can contact me to volunteer for the study.

If you have any questions, please contact Michael Crowe or Lorraine Sheppard:

D.3 QUESTIONS FOR PARTICIPANTS

Name	
Are you:	
Staff □ Post-graduate □ Under-graduate □	
What is your highest level of third level education?	
Bachelor □ Grad cert/diploma □ Masters □ PhD □	
Broadly speaking, what is your main area or what are your main areas of research?	
What research design or designs do you most commonly use?	
Qualitative	Narrative | Phenomenology | Grounded theory | Ethnography | Narrative case study
Descriptive, Exploratory, Observational	Cross-sectional | Longitudinal | Retrospective | Prospective | Correlational | Predictive
	Cohort | Case-control | Survey | Developmental | Normative | Case study
True experimental	Pre-test/post-test control group | Solomon four-group | Post-test only control group | Randomised two-factor | Placebo controlled trial
Quasi-experimental	Post-test only | Non-equivalent control group | Counter balanced *(cross-over)* | Separate sample pre-test post-test [no Control] [Control] | Multiple time series
Single system	One-shot experimental *(case study)* | Simple time series | One group pre-test/post-test | Within subjects *(Equivalent time, Repeated measures, Multiple treatment)* | Multiple baseline | Interactive
Mixed Method	Sequential | Concurrent | Transformative
Synthesis	Systematic review | Critical review | Thematic synthesis | Meta-ethnography
For how many years have you been involved in research?	
0–2 □ 3–5 □ 6–8 □ 9+ □	
On how many research projects have you worked?	
0–2 □ 3–5 □ 6–8 □ 9+ □	
On how many research projects have you been the lead researcher:	
0 □ 1–2 □ 3–4 □ 5+ □	
How would you rate your experience as a researcher, on a scale of 1 to 5, with 1 being an novice to 5 being a expert:	
1 2 3 4 5	
Looking at the critical appraisal tool, in your opinion what were the strengths of the tool?	
What were the weaknesses of the tool?	
Looking at the ***guide*** for the critical appraisal tool, what were the strengths of the guide?	
What were the weaknesses of the guide?	
Besides being used as a critical appraisal tool, do you think the tool could have any other uses?	
Is there anything else you would like to add?	

Appendix E – Material for participants, compare CAT with no CAT

Materials for participants in the *Compare CAT with no CAT* study (Chapter 7) are reproduced on the following pages.

E.1 INFORMED CONSENT FORM

Principal investigator	Michael Crowe, MIT, BSc (Mgmt), ADMT, PhD student
Supervisor	Lorraine Sheppard, PhD
Study title	The development of a critical appraisal tool for qualitative and quantitative research in health – A PhD research study
School	Public Health, Tropical Medicine and Rehabilitation Science

I understand the aim of this research study is to investigate the use of critical appraisal tools in health research and to develop a tool which can be used across a broad spectrum of qualitative and quantitative health research so that the content of research can be thoroughly compared when reviewing the literature.

I understand that my participation will involve independently critically appraising 5 (five) research articles using one of two critical appraisal tools and completing a questionnaire to obtain feedback on the process. I agree that the researcher may use the results as described in the information sheet.

I acknowledge that: *[Please tick (✓) the appropriate box]*	**Yes**	**No**
Any risks and possible effects of participating in the study have been explained to my satisfaction	☐	☐
Taking part in this study is voluntary and I am aware that I can stop taking part at any time without explanation or prejudice and to withdraw any unprocessed data I have provided	☐	☐
Any information I give will be kept strictly confidential and my name will not be used to identify me with this study	☐	☐
I consent to participate in this study	☐	☐
I consent to provide information regarding my research experience	☐	☐
I consent to provide feedback on the appraisal method used	☐	☐

Name _____ **Phone** _____

Email _____

Signature _____ **Date** _____

E.2 INFORMATION SHEET

Principal investigator	Michael Crowe, MIT, BSc (Mgmt), ADMT, PhD student
Supervisor	Lorraine Sheppard, PhD
Study title	The development of a critical appraisal tool for qualitative and quantitative research in health – A PhD research study
School	Public Health, Tropical Medicine and Rehabilitation Science

You are invited to take part in a research study about the use of critical appraisal tools in health research and to help develop a tool which can be used across a broad spectrum of qualitative and quantitative health research. Based on this tool, the content of research can be thoroughly compared when reviewing the literature or undertaking a systematic review.

The study is being conducted by Michael Crowe and will contribute to a PhD at James Cook University. Michael's supervisor is Professor Lorraine Sheppard.

If you agree to be involved in the study, you will be asked to independently appraise five (5) research articles using one of two different critical appraisal methods. You will be matched with another participant based on research experience, established through a short questionnaire, and you will then be randomly assigned to one of the appraisal methods. The purpose of this is to is to see how researchers appraise papers under the two different methods being investigated.

The research articles, tool, and instruction on the appraisal method will be given to you by the principal investigator. After appraising the papers, you will be asked some questions by the researcher to gather information regarding the tool, such as ease of use and overall impression.

There are no risks associated with the study and taking part is completely voluntary. You can stop taking part in the study at any time without explanation or prejudice. You may also withdraw any unprocessed data from the study. Your responses and contact details are strictly confidential. The data from the study will be used in research publications but you will not be identified in any way in these publications.

If you are interested, please contact me before close of business on Friday 30 July 2010. I will be distributing the papers and appraisal tools by Friday 6 August 2010 and require the appraisal forms to be returned to me by Friday 27 August 2010.

If you know of others that might be interested in this study, you can pass this information sheet on to them so they can contact me to volunteer for the study.

If you have any questions, please contact Michael Crowe or Lorraine Sheppard:

E.3 PRE-APPRAISAL QUESTIONS

Principal investigator Michael Crowe, MIT, BSc (Mgmt), ADMT, PhD student
Supervisor Lorraine Sheppard, PhD
Study title The development of a critical appraisal tool for qualitative and quantitative research in health – A PhD research study
School Public Health, Tropical Medicine and Rehabilitation Science

The purpose of this short questionnaire is to establish your research experience. This is needed to match you with another participant and then you will be randomly assigned to one of the appraisal methods being investigated. The results will be analysed to see how novice, intermediate, and expert researchers appraise research papers. Collection of your name and contact details are for the sole purpose of administering this study, no personally identifiable information will be used to analyse the data or in publication of results.

Name		
Contact details		
Phone/Extension	Building no. Office no.	Email
Which of these describes you best? (tick one only)		
☐ Staff	☐ Post-graduate	☐ Under-graduate
What is your *current* highest attainment in third level education? (tick one only)		
☐ Bachelor	☐ Grad cert/diploma	☐ Masters ☐ PhD
Broadly speaking, what is your main area or what are your main areas of research?		
What research design or designs do you or have you most commonly use in your research? (tick all that apply)		
☐ Qualitative	Narrative | Phenomenology | Grounded theory | Ethnography | Narrative case study	
☐ Descriptive, Explora-tory, Observational	Cross-sectional | Longitudinal | Retrospective | Prospective | Correlational | Predictive	
	Cohort | Case-control | Survey | Developmental | Normative | Case study	
☐ True experimental	Pre-test/post-test control group | Solomon four-group | Post-test only control group | Randomised two-factor | Placebo controlled trial	
☐ Quasi-experimental	Post-test only | Non-equivalent control group | Counter balanced *(cross-over)* | Separate sample pre-test post-test [no Control] [Control] | Multiple time series	
☐ Single system	One-shot experimental *(case study)* | Simple time series | One group pre-test/post-test | Within subjects *(Equivalent time, Repeated measures, Multiple treatment)* | Multiple baseline | Interactive	
☐ Mixed Method	Sequential | Concurrent | Transformative	
☐ Synthesis	Systematic review | Critical review | Thematic synthesis | Meta-ethnography	
For how many years have you been involved in research? (tick one only)		
☐ 0–2	☐ 3–5	☐ 6–8 ☐ 9+
On how many research projects have you worked? (tick one only)		
☐ 0–2	☐ 3–5	☐ 6–8 ☐ 9+
On how many research projects have you been the *lead* researcher? (tick one only)		
☐ 0	☐ 1–2	☐ 3–4 ☐ 5+
How would you rate your experience as a researcher, on a scale of 1 to 5, with 1 being an novice to 5 being an expert: (tick one only)		
1 ☐	2 ☐	3 ☐ 4 ☐ 5 ☐

Please save this form. You can either print or email the form to me at these contact details:

E.4 POST-APPRAISAL QUESTIONS

Principal investigator Michael Crowe, MIT, BSc (Mgmt), ADMT, PhD student
Supervisor Lorraine Sheppard, PhD
Study title The development of a critical appraisal tool for qualitative and quantitative research in health – A PhD research study
School Public Health, Tropical Medicine and Rehabilitation Science

The purpose of this short questionnaire is to obtain feedback on the tool you used to appraise research papers. Collection of your name is for the sole purpose of administering this study, no personally identifiable information will be used to analyse the data or in publication of results.

Name
Looking at the critical appraisal tool, in your opinion what were the strengths of the tool?
What were the weaknesses of the tool?
Looking at the ***guide*** for the critical appraisal tool, what were the strengths of the guide?
What were the weaknesses of the guide?
Besides being used as a critical appraisal tool, do you think the tool could have any other uses?
Is there anything else you would like to add?

Appendix F – Crowe Critical Appraisal Tool and user guide

The Crowe Critical Appraisal Tool and user guide (referred to in Chapter 8) are reproduced on the following pages.

F.1 CROWE CRITICAL APPRAISAL TOOL (CCAT)

Crowe Critical Appraisal Tool (CCAT)

Ref Reviewer

This form must be used in conjunction with the CCAT user guide; otherwise validity and reliability may be severely compromised.

Citation

	Year

Research design (insert exact design if applicable)

☐ Not research	Article | Editorial | Report | Opinion | Guideline | Pamphlet | …	
☐ Historical	…	
☐ Qualitative	Narrative | Phenomenology | Ethnography | Grounded theory | Narrative case study | …	
☐ Descriptive, Exploratory, Observational	Cross-sectional | Longitudinal | Retrospective | Prospective | Correlational | Predictive | …	
	Cohort | Case-control | Survey | Developmental | Normative | Case study | …	
Experimental	☐ True experimental	Pre-test/post-test control group | Solomon four-group | Post-test only control group | Randomised two-factor | Placebo controlled trial | …
	☐ Quasi-experimental	Post-test only | Non-equivalent control group | Counter balanced *(cross-over)* | Multiple time series | Separate sample pre-test post-test [no Control] [Control] | …
	☐ Single system	One-shot experimental *(case study)* | Simple time series | One group pre-test/post-test | Interactive | Within subjects *(Equivalent time, repeated measures, multiple treatment)* | Multiple baseline | …
☐ Mixed Method	Action research | Sequential | Concurrent | Transformative | …	
☐ Synthesis	Systematic review | Critical review | Thematic synthesis | Meta-ethnography | …	
☐ Other	…	

Sampling

Total size		Group 1:		Group 2:		Group 3:		Group 4:		Control:	
Population & Sample									

Data collection (add if not listed)

Audit/Review	a) Primary | Secondary | … b) Authoritative | Partisan | Antagonist | … c) Literature | Systematic | …	Interview	a) Formal | Informal | … b) Structured | Semi-structured | Unstructured | … c) One-on-one | Group | Multiple | Self administered | …
Observation	a) Participant | Non-participant | … b) Structured | Semi-structured | Unstructured | … c) Covert | Candid | …	Testing	a) Standardised | Norm-ref | Criterion-ref | Ipsative | … b) Objective | Subjective | … c) One-on-one | Group | Self administered | …

Scores

Preliminaries		Design		Data Collection		Results		Total	
Introduction		Sampling		Ethical Matters		Discussion		Total %	

General notes

Appraise research on the merits of the research design used, not against other research designs.

Category Item	Item descriptors [☑ Present; ☒ Absent; ■ Not applicable]	Where?	Score [0–5]
1. Preliminaries			
Title	1. Includes study aims ☐ and design ☐		
Abstract (assess last)	1. Key information ☐ 2. Balanced ☐ and informative ☐		
Text (assess last)	1. Sufficient detail others could reproduce ☐ 2. Clear/concise writing ☐, table(s) ☐, diagram(s) ☐, figure(s) ☐		
		Preliminaries	
2. Introduction			
Background	1. Summary of current knowledge ☐ 2. Specific problem(s) addressed ☐ and reason(s) for addressing ☐		
Objective	1. Primary objective(s), hypothesis(es), or aim(s) ☐ 2. Secondary question(s) ☐		
	Is it worth continuing?	**Introduction**	
3. Design			
Research design	1. Research design(s) chosen ☐ and why ☐ 2. Suitability of research design(s) ☐		
Intervention, Treatment, Exposure	1. Intervention(s)/treatment(s)/exposure(s) chosen ☐ and why ☐ 2. Precise details of the intervention(s)/treatment(s)/exposure(s) ☐ for each group ☐ 3. Intervention(s)/treatment(s)/exposure(s) valid ☐ and reliable ☐		
Outcome, Output, Predictor, Measure	1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen ☐ and why ☐ 2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s) ☐ 3. Outcome(s)/output(s)/predictor(s)/measure(s) valid ☐ and reliable ☐		
Bias, etc	1. Potential bias ☐, confounding variables ☐, effect modifiers ☐, interactions ☐ 2. Sequence generation ☐, group allocation ☐, group balance ☐, and by whom ☐ 3. Equivalent treatment of participants/cases/groups ☐		
	Is it worth continuing?	**Design**	
4. Sampling			
Sampling method	1. Sampling method(s) chosen ☐ and why ☐ 2. Suitability of sampling method ☐		
Sample size	1. Sample size ☐, how chosen ☐, and why ☐ 2. Suitability of sample size ☐		
Sampling protocol	1. Target/actual/sample population(s): description ☐ and suitability ☐ 2. Participants/cases/groups: inclusion ☐ and exclusion ☐ criteria 3. Recruitment of participants/cases/groups ☐		
	Is it worth continuing?	**Sampling**	
5. Data collection			
Collection method	1. Collection method(s) chosen ☐ and why ☐ 2. Suitability of collection method(s) ☐		
Collection protocol	1. Include date(s) ☐, location(s) ☐, setting(s) ☐, personnel ☐, materials ☐, processes ☐ 2. Method(s) to ensure/enhance quality of measurement/instrumentation ☐ 3. Manage non-participation ☐, withdrawal ☐, incomplete/lost data ☐		
	Is it worth continuing?	**Data collection**	
6. Ethical matters			
Participant ethics	1. Informed consent ☐, equity ☐ 2. Privacy ☐, confidentiality/anonymity ☐		
Researcher ethics	1. Ethical approval ☐, funding ☐, conflict(s) of interest ☐ 2. Subjectivities ☐, relationship(s) with participants/cases ☐		
	Is it worth continuing?	**Ethical matters**	
7. Results			
Analysis, Integration, Interpretation method	1. A.I.I. method(s) for primary outcome(s)/output(s)/predictor(s) chosen ☐ and why ☐ 2. Additional A.I.I. methods (e.g. subgroup analysis) chosen ☐ and why ☐ 3. Suitability of analysis/integration/interpretation method(s) ☐		
Essential analysis	1. Flow of participants/cases/groups through each stage of research ☐ 2. Demographic and other characteristics of participants/cases/groups ☐ 3. Analyse raw data ☐, response rate ☐, non-participation/withdrawal/incomplete/lost data ☐		
Outcome, Output, Predictor analysis	1. Summary of results ☐ and precision ☐ for each outcome/output/predictor/measure 2. Consideration of benefits/harms ☐, unexpected results ☐, problems/failures ☐ 3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes) ☐		
		Results	
8. Discussion			
Interpretation	1. Interpretation of results in the context of current evidence ☐ and objectives ☐ 2. Draw inferences consistent with the strength of the data ☐ 3. Consideration of alternative explanations for observed results ☐ 4. Account for bias ☐, confounding/effect modifiers/interactions/imprecision ☐		
Generalisation	1. Consideration of overall practical usefulness of the study ☐ 2. Description of generalisability (external validity) of the study ☐		
Concluding remarks	1. Highlight study's particular strengths ☐ 2. Suggest steps that may improve future results (e.g. limitations) ☐ 3. Suggest further studies ☐		
		Discussion	
9. Total			
Total score	1. Add all scores for categories 1–8		
		Total	

F.2 Crowe Critical Appraisal Tool (CCAT) user guide

Summary of main points

- Read each paper thoroughly.
- Research designs should be appraised on their own merits, not to a 'gold standard'.
- All categories must be scored – it does not matter which research design is used.
    - Category scores are whole numbers only (i.e. 0, 1, 2, 3, 4, 5)
    - The lowest score is 0
    - The highest score is 5.
- Items may be marked ☑ present, ☒ absent, or ■ not applicable.
    - Tick marks are not a checklist to be totalled – they are a guide to scoring a category.
- If in doubt use your best judgement, there is no right or wrong answer.

Contents

Introduction

The CCAT is demanding. It assumes that you are familiar with research designs, sampling techniques, ethics, data collection methods, and statistical and non-statistical data analysis techniques. Therefore, it may be helpful to have a general research methods text book available when you appraise papers.

The information sought when appraising a paper is unlikely to be in the sequence outlined in the CCAT form. Therefore, it is suggested that you read each paper quickly from start to finish getting an overall sense of what is being discussed.

On the first reading of a paper, these sections on the first page of the CCAT form can be filled in before you begin scoring the paper:

- Paper ID – Used to keep track of papers appraised.
- Citation – Match a CCAT form with the paper appraised.
- Research design – Indicate the research design or designs used.
- Sampling – Write down the total sample size and the sample size for each group, where applicable. Briefly describe the sample and the population the sample was selected from. Note any questions which occur about the sample.
- Data collection – Indicate the data collection method or methods used.
- My notes – Add thoughts that occur to you during the appraisal process.

Next, re-read the paper and fill in the second page of the CCAT. Insert any notes or page numbers where you found relevant information as you read the paper. This will help to jog your memory if you need to go through the paper in the future or need to justify your appraisal.

Some categories have the prompt: Is it worth continuing? If there are serious flaws in a paper in any of these categories, you should determine if it is worth continuing to appraise the paper or whether appraisal should be abandoned and the paper rejected.

Finally, transfer the scores from the second page to the first page of the CCAT form. By doing this, the majority of the information required for the appraisal is on the first page.

Overview of scoring a paper

The CCAT form is divided into eight categories and 22 items. An item has multiple parts which describe the item and make it easier to appraise and score a category. Each category receives its own score on a 6 point scale from 0–5. A score of 0 is the lowest score a category can achieve, while a score of 5 is the highest.

Categories can only be scored as a whole number or integer, i.e. 0, 1, 2, 3, 4, or 5. Half marks are not allowed.

There are tick boxes (☐) beside item descriptors. The tick box is useful to indicate if the descriptor for the item is:

- Present (☑) – For an item descriptor to be marked as present, there should be evidence of it being present rather than an assumption of presence.
- Absent (☒) – For an item descriptor to be marked as absent, it is implied that it should be present in the first place.
- Not applicable (■) – For an item descriptor to be marked as not applicable, the item descriptor must not be relevant given the characteristics of the paper being appraised and is, therefore, not considered when assigning a score to a category.

Whether an item descriptor is present, absent, or not applicable is further explored in the section *Guidelines for scoring categories and items*.

All categories must be scored because all categories are applicable in all research designs. Only items may be marked 'not applicable'.

While it may be tempting to add up all the present marks (☑) and all the absent marks (☒) in each category and to use the proportion of one to the other to calculate the score for the category, this is strongly discouraged. It is strongly discouraged because not all item descriptors in any category are of equal importance. For example, in the *Introduction* category there are two items (*Background* and *Objective*) and a total of five tick boxes. If a paper being appraised has all boxes marked as present (☑) except for *Primary objective(s), hypothesis(es), or aim(s)*, should the paper be scored 4/5 for that category? It could be argued that a research paper without a primary objective, hypothesis, or aim is fundamentally flawed and, as a result, should be scored 0/5 even though the other four tick boxes were marked as present.

Therefore, the tick marks for present, absent, or not applicable are to be used as a guide to scoring a category rather than as a simple check list. It is up to you as the appraiser to take into consideration all aspects of each category and based on both the tick marks and judgement assign a score to a category.

Similarly, the research design used in each paper should be appraised on its own merits and not relative to some preconceived notion of a hierarchy of research designs. What is most important is that the paper used an appropriate research design based on the research question being addressed, rather than what research design in itself was used.

The total score given to a paper can be expressed as a percentage by dividing the total score by forty (40) and writing the result on the first page of the CCAT form. The total percent should be written to the nearest full percent. There is no need for decimal places because they do not add anything to the accuracy of the total percent score obtained.

Finally, the total or percent score a paper obtains is not the sole criterion on which an overall assessment of a paper is based. The total or total percent score is a useful summary but may not be applicable in all cases. When reporting on an appraisal using the CCAT, the total or total percent score should be stated along with the score obtained in every category. This prevents papers that score high overall but very poorly in one or more categories being hidden amongst papers which scored high throughout all categories. Based on the reasons for the appraisal, some papers which have a low score in certain category but which have high a high total score may be ranked lower than those with a lower total score but a high score in that particular category. These processes are up to you, as the appraiser, to detail before you begin appraising papers.

Guidelines for scoring categories and items

1. Preliminaries

Title

1. Includes study aims and design
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.

Abstract

1. Contains key information
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.
2. Balanced and informative
    - Traditionally only required for reporting research.
    - It has been assumed that this does not affect the overall quality of the research but there is little evidence one way or the other.

Text

Note – This item can only be assessed when the article has been read in full.

1. Sufficient detail others could reproduce
    - This is an over-arching concept and should be present throughout the study.
2. Clear, concise writing/table(s)/diagram(s)/figure(s)
    - This is an over-arching concept and should be present throughout the study.

2. Introduction

Background

1. Summary of current knowledge
    - Current and applicable knowledge provides a context for the study.
2. Specific problem(s) addressed and reason(s) for addressing
    - Description of why the study was undertaken.
    - Links current knowledge and stated objective(s), hypothesis(es), or aim(s).

Objective

1. Primary objective(s), hypothesis(es), aim(s)
    - The study must have at least one stated objective, hypothesis, or aim.
2. Secondary question(s)
    - Secondary question(s) may sometimes arise based on the primary objective(s), hypothesis(es), or aim(s).
    - Since this is not always the case, a study without secondary questions should not be penalised.

3. Design

Research design

1. Research design(s) chosen and why
    - Description of the research design chosen and why it was chosen.
2. Suitability of research design(s)
    - The research design should be congruent with *Background*, *Objective*, *Intervention(s)/treatment(s)/exposure(s)*, and *Outcome(s)/ output(s)/predictor(s)*.

Intervention, Treatment, Exposure

1. Intervention(s)/treatment(s)/exposure(s) chosen and why
    - Where a study does not normally have an intervention/treatment/exposure, it should not be penalised when none is present.
    - Statement for every intervention/treatment/exposure chosen and why it was chosen.
    - Each intervention/treatment/exposure must be congruent with *Background*, *Objective*, and *Research design*.

2. Precise details of the intervention(s)/treatment(s)/exposure(s) for each group
    - Full details are presented for every intervention/treatment/exposure for every participant/case/group so that other studies could duplicate.
3. Intervention(s)/treatment(s)/exposure(s) valid and reliable
    - A statement of reliability/validation or why there is no validation/ reliability for each intervention/treatment/exposure.

Outcome, Output, Predictor, Measure

1. Outcome(s)/output(s)/predictor(s)/measure(s) chosen and why
    - All research has at least one expected outcome/output/predictor/ measure.
    - Statement for each outcome/output/predictor/measure chosen and why it was chosen.
    - Each outcome/output/predictor/measure must be congruent with *Background*, *Objective*, *Research design*, and *Intervention/ treatment/exposure*.
2. Clearly define outcome(s)/output(s)/predictor(s)/measure(s)
    - Full details are presented of every expected outcome/output/ predictor/measure for every participant/case/group so that other studies could duplicate.
3. Outcome(s)/output(s)/predictor(s)/measure(s) valid and reliable
    - A statement of reliability/validation or why there is no validation/ reliability for each outcome/output/predictor/measure.

Note – In some cases the Outcome(s)/output(s)/predictor(s)/measure(s) may be similar to or the same as the Objective(s), hypothesis(es), aim(s). However, in most cases to achieve the Objective(s), hypothesis(es), aim(s) a series of Outcome(s)/output(s)/predictor(s)/measure(s) are required.

Bias, etc.

1. Potential sources of bias, confounding variables, effect modifiers, interactions
    - Identification of potential sources of:
        - Bias – e.g. attrition, detection, experimental, information, interview, observation, performance, rater, recall, selection.
        - Confounding variables or factors – A variable which interferes between the intervention/treatment/exposure and the outcome/ output/predictor/measure.
        - Effect modification – A variable which modifies the association between the intervention/treatment/exposure and the outcome/ output/predictor/measure.

        - Interaction effects – When various combinations of intervention(s)/treatment(s)/exposure(s) cause different outcome(s)/output(s)/predictor(s)/measure(s).
    - Should be identified, as far as possible, within the *Research design* before data collection begins in order to minimise their effect.
    - See also *Sampling* and *Data collection*.

2. Sequence generation, group allocation, group balance, and by whom
    - In studies where participants/cases are allocated to groups, the methods used should be stated and procedures established before recruitment or data collection begins (e.g. blinding, method used to randomise, allocate to or balance groups).

3. Equivalent treatment of participants/cases/groups
    - Each participant/case/group must be treated equivalently apart from any intervention/treatment/exposure.
    - If participants/cases/groups are not treated equivalently a statement regarding why this was not possible, how this may affect results, and procedures in place for managing participants/cases/groups.
    - See also *Sampling protocol*, *Collection protocol*, and *Participant ethics*.

4. Sampling

Sampling method

1. Sampling method(s) chosen and why
    - Description of the sampling method chosen and why it was chosen.
    - Sampling methods are normally probability or non-probability based.
    - Examples include: Simple random, systematic, stratified, cluster, convenience, representative, purposive, snowball, and theoretical.
    - Also included here is the search strategy used for a systematic review (e.g. databases searched, search terms).

2. Suitability of sampling method
    - The sampling method should be decided and in place before recruitment or data collection begins.
    - The sampling method should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/ predictor/measure*, and *Bias etc*.

Sample size

1. Sample size, how chosen, and why
    - Description of the sample size, the method of sample size calculation, and why that method was chosen.

- Sample size calculations are normally probability or non-probability based.
- Examples of how calculations can be made include: Accuracy [e.g. confidence interval (α), population or sample variance (s^2, σ^2), effect size or index (ES, d), power (1-β)], analysis, population, redundancy, saturation, and budget.

2. Suitability of sample size
    - The sample size or estimate of sample size, with contingencies, should be described and calculated before recruitment/data collection begins.
    - The sample size should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/ measure*, and *Bias etc.*

Note – Sample size calculations are not required for systematic reviews, because it is not possible to know the number of papers that will meet the selection criteria, or for some single system designs.

Sampling protocol

1. Description and suitability of target/actual/sample population(s)
    - The target/actual/sample population(s) should be described.
    - The target/actual/sample population(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor/measure*, and *Bias etc.*

2. Inclusion and exclusion criteria for participants/cases/groups
    - Inclusion and exclusion criteria should be explicitly stated and established before recruitment/data collection begins.
    - The use of inclusion and exclusion criteria (especially exclusion criteria) should not be used in such a way as to bias the sample.

3. Recruitment of participants/cases/groups
    - Description of procedures for recruitment and contingencies put in place.
    - Recruitment should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Bias etc.*, and other aspects of *Sampling*.
    - See also *Participant ethics*, *Researcher ethics*, and *Collection protocol*.

Note – For systematic reviews inclusion and exclusion criteria only need to be appraised, because they refer to the parameters used to select papers.

5. Data collection

Collection method

1. Collection method(s) chosen and why
    - Description of the method(s) used to collect data and why each was chosen.
    - In systematic reviews, this refers to how information was extracted from papers, because these are the data collected.
2. Suitability of collection method(s)
    - The data collection method(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/ output/predictor/measure*, *Bias etc.*, and *Sampling*.

Collection protocol

1. Include date(s), location(s), setting(s), personnel, materials, processes
    - Description of and details regarding exactly how data were collected, especially any factor(s) which may affect *Outcome/output/predictor/ measure* or *Bias etc.*
2. Method(s) to ensure/enhance quality of measurement/instrumentation
    - Description of any method(s) used to enhance or ensure the quality of data collected (e.g. pilot study, instrument calibration, standardised test(s), independent/multiple measurement, valid/reliable tools).
    - Also includes any method(s) which reduce or eliminate bias, confounding variables, effect modifiers, interactions which are not an integral part of the *Design* category (e.g. blinding of participants, intervention(s), outcome(s), analysis; protocols and procedures implemented).
    - In qualitative studies, this relates to concepts such as trustworthiness, authenticity, and credibility.
    - See also *Bias etc.*
3. Manage non-participation, withdrawal, incomplete/lost data
    - Description of any method(s) used to manage or prevent non-participation, withdrawal, or incomplete/lost data.
    - These include but are not limited to: Intention to treat analysis (ITT); last observation carried forward (LOCF); follow up (FU), e.g. equal length, adequate, or complete; and, completer analysis, e.g. on-treatment, on-protocol.

6. Ethical matters

Note – Some studies may have been conducted before Ethical matters were a major point of concern. The research ethics standards of the time may need to be taken into consideration rather than the current standards.

Note – All research requires Ethical matters consideration even if formal ethics committee or ethics board approval is not required. This includes systematic reviews.

Participant ethics

1. Informed consent, equity
    - All participants must have provided their informed consent.
    - Equity includes, but is not limited to, cultural respect, just and equitable actions, no harm to participants, debriefing, and consideration for vulnerable individuals or groups.
2. Privacy, confidentiality/anonymity
    - The privacy and confidentiality and/or anonymity of participants must be catered for.
    - If this is not possible, the informed and written consent of individuals affected must be obtained.

Researcher ethics

1. Ethical approval, funding, conflict(s) of interest
    - A statement of ethical approval from recognised Ethics Committee(s) or Board(s) suitable for the study being undertaken.
    - Any real, perceived, or potential conflict(s) of interest should be stated.
    - All sources of funding should be stated.
2. Subjectivities, relationship(s) with participants/cases
    - Description of how the researcher(s) could have potentially or did affect the outcomes of the study through their presence or behaviour.
    - Includes a description of procedures used to minimise this occurring.
    - See also *Bias etc.*

7. Results

Analysis, Integration, Interpretation method

1. A.I.I. (Analysis/Integration/Interpretation) method(s) for primary outcome(s)/output(s)/predictor(s) chosen and why
    - Description of statistical and non-statistical method(s) used to analyse/integrate/interpret Outcome(s)/output(s)/predictor(s)/measure(s) and why each was chosen.
2. Additional A.I.I. methods (e.g. subgroup analysis) chosen and why
    - Description of additional statistical and non-statistical method(s) used to analyse/integrate/interpret Outcome(s)/output(s)/predictor(s)/ measure(s) and why each was chosen.

3. Suitability of analysis/integration/interpretation method(s)
    - The analysis/integration/interpretation method(s) should be congruent with *Objective*, *Research design*, *Intervention/treatment/exposure*, *Outcome/output/predictor*, *Bias etc.*, *Sampling*, and *Data collection*.

Essential analysis

1. Flow of participants/cases/groups through each stage of research
    - Description of how participants/cases/groups advanced through the study.
    - Explanation of course of intervention/treatment/exposure.
2. Demographic and other characteristics of participants/cases/groups
    - Description of baseline characteristics of participants/cases/groups so this can be integrated into the analysis.
3. Analyse raw data, response rate, non-participation, withdrawal, incomplete/ lost data
    - Unadjusted data should be analysed.
    - There may be differences between those that completed and those that did not complete the study.

Outcome, Output, Predictor analysis

1. Summary of results and precision for each outcome/output/predictor/measure
    - Results summarised with, where possible, an indicator of the precision and effect size of each result for each outcome/output/predictor/ measure.
    - Where data are adjusted, make clear what was adjusted and why.
    - Where data are categorised, report of internal and external boundaries.
    - Use of quotations to illustrate themes/findings, privileging of subject meaning, adequate description of findings, evidence of reflexivity.
2. Consideration of benefits/harms, unexpected results, problems/failures
    - Description of all outcomes, not just ones being looked for.
    - Description of differences between planned and actual implementation, and the potential effect on results.
3. Description of outlying data (e.g. diverse cases, adverse effects, minor themes)
    - Exploration of outliners because they may not be anomalous.

8. Discussion

Interpretation

1. Interpretation of results in the context of current evidence and objectives
    - Summarises key results in relation to *Background* and *Objective*.
    - Compare and contrast other research findings.
2. Draw inferences consistent with the strength of the data
    - Do not over or under represent data.
    - Draw inferences based on the entirety of available evidence.
    - See also *Sampling* and *Data collection*.
3. Consideration of alternative explanations for observed results
    - Exploration of reasons for differences between observed and expected.
    - Determines if other factors may lead to similar results.
4. Account for bias, confounding, interactions, effect modifiers, imprecision
    - Discussion on magnitude and direction of *Bias etc.* and how this may have affected the results.
    - See also *Essential analysis*.

Generalisation

1. Consideration of overall practical usefulness of the study
    - Discussion on practical vs. theoretical usefulness.
2. Description of generalisability (external validity) of the study
    - Dependent on *Design*, *Sampling*, and *Data collection*.

Concluding remarks

1. Highlight study's particular strengths
    - What did the study do well?
2. Suggest steps that may improve future results (e.g. limitations)
    - How could the study have been better?
3. Suggest further studies
    - Where should the next study begin?

9. Total

Total score

1. Add all scores for categories 1–8
    - Total the scores for all categories.
    - To calculate the total percent, divide the total score by 40.

Complete reference list

WORKS CITED

Ad Hoc Working Group for Critical Appraisal of the Medical Literature. (1987). A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine, 106*(4), 598-604. doi:10.1059/0003-4819-106-4-598

Alasuutari, P., Bickman, L., & Brannen, J. (2008). Introduction: Social research in changing social conditions. In P. Alasuutari, L. Bickman & J. Brannen (Eds.), *The SAGE handbook of social research methods* (pp. 1-8). London: Sage.

Allotey, P., Reidpath, D., & Pokhrel, S. (2010). Social sciences research in neglected tropical diseases 1: the ongoing neglect in the neglected tropical diseases. *Health Research Policy and Systems, 8*(32). doi:10.1186/1478-4505-8-32

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2), 1-38.

Armijo Olivo, S., Macedo, L. G., Gadotti, I. C., Fuentes, J., Stanton, T., & Magee, D. J. (2008). Scales to assess the quality of randomized controlled trials: A systematic review. *Physical Therapy, 88*(2), 156-175. doi:10.2522/ptj.20070147

Avis, M. (1994). Reading research critically. I. An introduction to appraisal: Designs and objectives. *Journal of Clinical Nursing, 3*(4), 227-234. doi:10.1111/j.1365-2702.1994.tb00393.x

Barker, J. H. (2008). Q-methodology: An alternative approach to research in nurse education. *Nurse Education Today, 28*(8), 917-925. doi:10.1016/j.nedt.2008.05.010

Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: A critical review. *BMC Medical Research Methodology, 9*(59). doi:10.1186/1471-2288-9-59

Bialocerkowski, A. E., Grimmer, K. A., Milanese, S. F., & Kumar, S. (2004). Application of current research evidence to clinical physiotherapy practice. *Journal of Allied Health, 33*(4), 230-237.

Bloch, R. (2010). G_String_III (Version 5.4.6). Hamilton, ON: Programme for Educational Research and Development. Retrieved from http://fhsperd.mcmaster.ca/g_string/

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity, 36*(4), 391-409. doi:10.1023/A:1020909529486

Boutron, I., Moher, D., Tugwell, P., Giraudeau, B., Poiraudeau, S., Nizard, R., & Ravaud, P. (2005). A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. *Journal of Clinical Epidemiology, 58*(12), 1233-1240. doi:10.1016/j.jclinepi.2005.05.004

Braithwaite, R. B. (1968). *Scientific explanation: A study of the function of theory, probability, and law in science*. Cambridge, MA: Cambridge University Press.

Brannen, J. (2004). Working qualitatively and quantitatively. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Eds.), *Qualitative research practice* (pp. 312-326). London: Sage.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Burch, B. D. (2011). Assessing the performance of normal-based and REML-based confidence intervals for the intraclass correlation coefficient. *Computational Statistics & Data Analysis*, *55*(2), 1018-1028. doi:16/j.csda.2010.08.007

Burnett, J., Kumar, S., & Grimmer, K. (2005). Development of a generic critical appraisal tool by consensus: Presentation of first round Delphi survey results. *Internet Journal of Allied Health Sciences and Practice, 3*(1), 22. Retrieved from http://ijahsp.nova.edu/

Buzan, T., & Abbott, S. (2005). *The ultimate book of mind maps: Unlock your creativity, boost your memory, change your life*. London: Thorsons.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Centre for Reviews and Dissemination. (2009). *Systematic reviews: CRD's guidance for undertaking reviews in health care*. York: University of York.

Cesario, S., Morin, K., & Santa-Donato, A. (2002). Evaluating the level of evidence of qualitative research. *Journal of Obstetric, Gynecologic, and Neonatal Nursing, 31*(6), 708-714. doi:10.1177/0884217502239216

Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials, 2*(1), 31-49.

Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher, 21*(6), 13-17. doi:10.3102/0013189x021006013

Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *JAMA, 272*(2), 101-104. doi:10.1001/jama.1994.03520020027007

Cooper, H. (1986). *The integrative research review: A systematic approach* (2nd ed.). Beverly Hills, CA: Sage.

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Côté, L., & Turgeon, J. (2005). Appraising qualitative research articles in medicine and medical education. *Medical Teacher, 27*(1), 71-75.

Creswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Crombie, I. K. (1996). *The pocket guide to critical appraisal: A handbook for health care professionals*. London: BMJ.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. doi:10.1177/0013164404266386

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. St Leonards, NSW: Allen & Unwin.

Crowe, M., & Sheppard, L. (2010). Qualitative and quantitative research designs are more similar than different. *Internet Journal of Allied Health Sciences and Practice, 8*(4). Retrieved from http://ijahsp.nova.edu/

Crowe, M., & Sheppard, L. (2011). Mind mapping research methods. *Quality and Quantity*, (Online First). doi:10.1007/s11135-011-9463-8

Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigour: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89. doi:10.1016/j.jclinepi.2010.02.008

Crowe, M., & Sheppard, L. (2011). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies*, *14*(12). 1505-1516. doi:10.1016/j.ijnurstu.2011.06.004

Crowe, M., Sheppard, L., & Campbell, A. (2011). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology*, (Online). doi:10.1016/j.jclinepi.2011.08.006

Crowe, M., Sheppard, L., & Campbell, A. (2011). A comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: A randomised trial. *International Journal of Evidence-Based Healthcare*, *9*(4), 444-449. doi:10.1111/j.1744-1609.2011.00237.x

Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M., & Hughes, E. (2007). A hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology, 60*(1), 43-49. doi:10.1016/j.jclinepi.2006.03.014

D'Auria, J. P. (2007). Using an evidence-based approach to critical appraisal. *Journal of Pediatric Health Care, 21*(5), 343-346. doi:10.1016/j.pedhc.2007.06.002

de Vet, H. C. W., de Bie, R. A., van der Heijden, G. J. M. G., Verhagen, A. P., Sijpkes, P., & Knipschild, P. G. (1997). Systematic reviews on the basis of methodological criteria. *Physiotherapy, 83*(6), 284-289. doi:10.1016/S0031-9406(05)66175-5

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitc, C., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27). doi:10.3310/hta7270

DePaulo, P. (2000). Sample size for qualitative research. *Quirk's Marketing Research Review,* (Article ID: 20001202). Retrieved from http://www.quirks.com/articles/a2000/20001202.aspx

Devers, K. J. (1999). How will we know "good" qualitative research when we see it? Beginning the dialogue in health services research. *Health Services Research, 34*(5 Part II), 1153-1188.

Dixon-Woods, M., Bonas, S., Booth, A., Jones, D. R., Miller, T., Sutton, A. J., . . . Young, B. (2006). How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research, 6*(1), 27-44. doi:10.1177/1468794106058867

Dixon-Woods, M., Booth, A., & Sutton, A. J. (2007). Synthesizing qualitative research: A review of published reports. *Qualitative Research, 7*(3), 375-422. doi:10.1177/1468794107078517

Dixon-Woods, M., Shaw, R. L., Agarwal, S., & Smith, J. A. (2004). The problem of appraising qualitative research. *Quality and Safety in Health Care, 13*(3), 223-225. doi:10.1136/qshc.2003.008714

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377-384.

Duffy, M. E. (1985). A research appraisal checklist for evaluating nursing research reports. *Nursing & Health Care, 6*(December), 539-547.

DuRant, R. H. (1994). Checklist for the evaluation of research articles. *Journal of Adolescent Health, 15*(1), 4-8. doi:10.1016/1054-139X(94)90381-6

Dye, J. F., Schatz, I. M., Rosenberg, B. A., & Coleman, S. T. (2000). Constant comparison method: A kaleidoscope of data. *The Qualitative Report, 4*(1/2). Retrieved from http://www.nova.edu/ssss/QR/

Earl-Slater, A. (2001). Critical appraisal and hierarchies of the evidence. *British Journal of Clinical Governance, 6*(1), 59-63. doi:10.1108/14664100110385154

Evans, M., & Pollock, A. V. (1985). A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *British Journal of Surgery, 72*(4), 256-260. doi:10.1002/bjs.1800720403

Farrand, P., Hussain, F., & Hennessy, E. (2002). The efficacy of the 'mind map' study technique. *Medical Education, 36*(5), 426-431. doi:10.1046/j.1365-2923.2002.01205.x

Genaidy, A. M., Lemasters, G. K., Lockey, J., Succop, P., Deddens, J., Sobeih, T., & Dunning, K. (2007). An epidemiological appraisal instrument - A tool for evaluation of epidemiological studies. *Ergonomics, 50*(6), 920-960. doi:10.1080/00140130701237667

Glasziou, P., Irwig, L., Bain, C., & Colditz, G. (2001). *Systematic reviews in health care: A practical guide*. Cambridge, MA: Cambridge University Press.

Glenny, A.-M. (2005). No "gold standard" critical appraisal tool for allied health research. *Evidence-Based Dentistry, 6*(4), 100-101. doi:10.1038/sj.ebd.6400351

Glynn, L. (2006). A critical appraisal tool for library and information research. *Library Hi Tech, 24*(3), 387-399. doi:10.1108/07378830610692154

Guyatt, G. H., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J., & Cook, R. J. (1995). Users' guides to the medical literature: IX. A method for grading health care recommendations. *JAMA, 274*(22), 1800-1804. doi:10.1001/jama.1995.03530220066035

Haadr, M. (2009). Random.org: Random sequence generator Retrieved 29 January, 2011, from http://www.random.org/sequences/

Hammersley, M. (1992). Deconstructing the qualitative-quantitative divide. In J. Brannen (Ed.), *Mixing methods: Qualitative and quantitative research* (pp. 39-55). Aldershot: Avebury.

Hawker, S., Payne, S., Kerr, C., Hardey, M., & Powell, J. (2002). Appraising the evidence: Reviewing disparate data systematically. *Qualitative Health Research, 12*(9), 1284-1299. doi:10.1177/1049732302238251

Heacock, H., Koehoorn, M., & Tan, J. (1997). Applying epidemiological principles to ergonomics: A checklist for incorporating sound design and interpretation of studies. *Applied Ergonomics, 28*(3), 165-172. doi:10.1016/S0003-6870(96)00066-X

Heller, R. F., Verma, A., Gemmell, I., Harrison, R., Hart, J., & Edwards, R. (2008). Critical appraisal for public health: A new checklist. *Public Health, 122*(1), 92-98. doi:10.1016/j.puhe.2007.04.012

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions (Version 5.0.1)*. London: The Cochrane Collaboration.

Howe, K. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10-16.

Howe, K., & Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher, 19*(4), 2-9.

Hudson, L. A., & Ozanne, J. L. (1988). Alternative ways of seeking knowledge in consumer research. *Journal of Consumer Research, 14*(4), 508-521.

Hunt, D. L., & McKibbon, K. A. (1997). Locating and appraising systematic reviews. *Annals of Internal Medicine, 126*(7), 532-538.

Hunt, S. D. (1991). Positivism and paradigm dominance in consumer research: Toward critical pluralism and rapprochement. *Journal of Consumer Research, 18*(1), 32-44.

Jadad, A. R., Moher, D., & Klassen, T. P. (1998). Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? *Archives of Pediatrics and Adolescent Medicine, 152*(8), 812-817. doi:10.1001/archpedi.152.8.812

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12. doi:10.1016/0197-2456(95)00134-4

Jüni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ, 323*(7303), 42-46. doi:10.1136/bmj.323.7303.42

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA, 282*(11), 1054-1060. doi:10.1001/jama.282.11.1054

Kastelic, J. P. (2006). Critical evaluation of scientific articles and other sources of information: An introduction to evidence-based veterinary medicine. *Theriogenology, 66*(3), 534-542. doi:10.1016/j.theriogenology.2006.04.017

Katrak, P., Bialocerkowski, A., Massy-Westropp, N., Kumar, V. S. S., & Grimmer, K. (2004). A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology, 4*(1). doi:10.1186/1471-2288-4-22

Khan, K. S., ter Riet, G., Glanville, J., Sowden, A. J., & Kleijen, J. (2001). Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews (CRD Report 4). York, England: University of York.

Kuper, A., Lingard, L., & Levinson, W. (2008). Critically appraising qualitative research. *BMJ, 337*(7671), 687-689. doi:10.1136/bmj.a1035

Law, M., Stewart, D., Pollock, N., Letts, L., Bosch, J., Westmorland, M., & Philpot, A. (2008). Occupational Therapy Evidence-Based Practice Research Group Retrieved 29 January, 2011, from http://www.srs-mcmaster.ca/Default.aspx?tabid=630

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology, 62*(10), e1-e34. doi:10.1016/j.jclinepi.2009.06.006

Lichtenstein, M. J., Mulrow, C. D., & Elwood, P. C. (1987). Guidelines for reading case-control studies. *Journal of Chronic Diseases, 40*(9), 893-903. doi:10.1016/0021-9681(87)90190-1

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(Monograph Supplement 9), 635-694.

Loney, P. L., Chambers, L. W., Bennett, K. J., Roberts, J. G., & Stratford, P. W. (1998). Critical appraisal of the health research literature: Prevalence or incidence of a health problem. *Chronic Diseases in Canada, 19*(4), 170-176.

Long, A. F., & Godfrey, M. (2004). An evaluation tool to assess the quality of qualitative research studies. *International Journal of Social Research Methodology, 7*(2), 181-196. doi:10.1080/1364557032000045302

Luttrell, W. (2005). Crossing anxious borders: Teaching across the quantitative-qualitative 'divide'. *International Journal of Research & Method in Education, 28*(2), 183-195. doi:10.1080/01406720500256251

MacAuley, D. (1994). READER: An acronym to aid critical reading by general practitioners. *British Journal of General Practice, 44*(379), 83–85.

MacAuley, D., McCrum, E., & Brown, C. (1998). Randomised controlled trial of the READER method of critical appraisal in general practice. *BMJ, 316*(7138), 1134-1137.

Maher, C. G., Sheerington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713-721.

Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527-551). San Diego, CA: Academic Press.

Meijman, F. J., & de Melker, R. A. (1995). The extent of inter- and intra-reviewer agreement on the classification and assessment of designs of single-practice research. *Family Practice, 12*(1), 93-97. doi:10.1093/fampra/12.1.93

Melnyk, B. M., & Fineout-Overholt, E. (2005). Rapid critical appraisal of randomized controlled trials (RCTs): An essential skill for evidence-based practice (EBP). *Pediatric Nursing, 31*(1), 50-52.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Miller, A. (Fall 2008). Realism. In E. N. Zalta (Ed.), Stanford encyclopedia of philosophy. Stanford, CA: Stanford University. Retrieved from http://plato.stanford.edu/archives/fall2008/entries/realism/.

Miller, D. C., & Salkind, N. J. (2002). *Handbook of research design and social measurement* (6th ed.). Thousand Oaks, CA: Sage.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet, 354*(9193), 1896-1900. doi:10.1016/S0140-6736(99)04149-5

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16*(1), 62-73. doi:10.1016/0197-2456(94)00031-W

Moher, D., Jones, A., & Lepage, L. (2001). Use of the CONSORT statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA, 285*(15), 1992-1995. doi:10.1001/jama.285.15.1992

Moncrieff, J., Churchill, R., Drummond, D. C., & McGuire, H. (2001). Development of quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods in Psychiatric Research, 10*(3), 126-133. doi:10.1002/mpr.108

Moyer, A., & Finney, J. W. (2005). Rating methodological quality: Toward improved assessment and investigation. *Accountability in Research, 12*(4), 299-313. doi:10.1080/08989620500440287

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

National Health and Medical Research Council, Australian Research Council, & Universities Australia. (2007). *Australian code for the responsible conduct of research*. Canberra, ACT: NHMRC.

Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approaches*. Boston, MA: Pearson.

Newcombe, R. G. (2011). Propagating imprecision: Combining confidence intervals from independent sources. *Communications in Statistics*, *40*(17), 3154-3180. doi:10.1080/03610921003764225

NHS Public Health Resources Unit. (2010). CASP: Critical Appraisal Skills Programme Retrieved 29 January, 2011, from http://www.sph.nhs.uk/what-we-do/public-health-workforce/resources/critical-appraisals-skills-programme/

Nielsen, M. E., & Reilly, P. L. (1985). A guide to understanding and evaluating research articles. *Gifted Child Quarterly, 29*(2), 90-92. doi:10.1177/001698628502900210

O'Cathain, A., Murphy, E., & Nicholl, J. (2008). Multidisciplinary, interdisciplinary, or dysfunctional? Team working in mixed-methods research. *Qualitative Health Research, 18*(11), 1574-1585. doi:10.1177/1049732308325535

Ogrinc, G., Mooney, S. E., Estrada, C., Foster, T., Goldmann, D., Hall, L. W., . . . Watts, B. (2008). The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: Explanation and elaboration. *Quality and Safety in Health Care, 17*(Supplement 1), i13-i32. doi:10.1136/qshc.2008.029058

Onwuegbuzie, A. J., & Leech, N. L. (2005). Taking the "q" out of research: Teaching research methodology courses without the divide between quantitative and qualitative paradigms. *Quality & Quantity, 39*(3), 267-295. doi:10.1007/s11135-004-1670-0

Onwuegbuzie, A. J., & Leech, N. L. (2007). A call for qualitative power analyses. *Quality and Quantity, 41*(1), 105-121. doi:10.1007/s11135-005-1098-1

Oxman, A. D., & Guyatt, G. H. (1988). Guidelines for reading literature reviews. *Canadian Medical Association Journal, 138*(8), 697-703.

Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage.

Petticrew, M. (2001). Systematic reviews from astronomy to zoology: Myths and misconceptions. *BMJ, 322*(7278), 98-101. doi:10.1136/bmj.322.7278.98

Pluye, P., Gagnon, M.-P., Griffiths, F., & Johnson-Lafleur, J. (2009). A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies, 46*(4), 529-546. doi:10.1016/j.ijnurstu.2009.01.009

Polgar, S., & Thomas, S. A. (2007). *Introduction to research in the health sciences* (5th ed.). Edinburgh: Churchill Livingstone.

Portney, L. G., & Watkins, M. P. (2008). *Foundations of clinical research: Applications to practice* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Ramasundarahettige, C. F. F., Donner, A., & Zou, G. Y. (2009). Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Statistics in Medicine*, *28*(7), 1041-1053. doi:10.1002/sim.3523

Rangel, S. J., Kelsey, J., Colby, C. E., Anderson, J., & Moss, R. L. (2003). Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *Journal of Pediatric Surgery, 38*(3), 390-396. doi:10.1053/jpsu.2003.50114

Rasmussen, L., O'Conner, M., Shinkle, S., & Thomas, M. K. (2000). The basic research review checklist. *Journal of Continuing Education in Nursing, 31*(1), 13-17.

Rehabilitation Studies Unit, R. R. C. S. (2010). PsycBITE: Psychological Database for Brain Impairment Treatment Efficacy Retrieved 29 January, 2011, from http://www.psycbite.com/

Reis, S., Hermoni, D., Van-Raalte, R., Dahan, R., & Borkan, J. M. (2007). Aggregation of qualitative studies - From theory to practice: Patient priorities and family medicine/general practice evaluations. *Patient Education and Counseling, 65*(2), 214-222. doi:10.1016/j.pec.2006.07.011

Reisch, J. S., Tyson, J. E., & Mize, S. G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics, 84*(5), 815-827.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ, 312*(7023), 71-72.

Sanderson, S., Tatt, I. D., & Higgins, J. P. T. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36*(3), 666-676. doi:10.1093/ije/dym018

School of Health and Rehabilitation Sciences University of Queensland. (2011). OTseeker: Occupational Therapy Systematic Evaluation of Evidence Retrieved 29 January, 2011, from http://www.otseeker.com/

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

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Silverman, D., & Marvasti, A. B. (2008). *Doing qualitative research: A comprehensive guide*. Los Angeles, CA: Sage.

Simon, S. D. (2001). Is the randomized clinical trial the gold standard of research? *Journal of Andrology, 22*(6), 938-943.

Sindhu, F., Carpenter, L., & Seers, K. (1997). Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. *Journal of Advanced Nursing, 25*(6), 1262-1268. doi:10.1046/j.1365-2648.1997.19970251262.x

Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ, 327*(7429), 1459-1461. doi:10.1136/bmj.327.7429.1459

Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational inquirers. *Educational Researcher, 15*(1), 4-12.

SPSS Inc. (2009). *PASW Statistics 18 command syntax reference*. Chicago, IL: IBM SPSS Inc.

Steinmetz, G. (1998). Critical realism and historical sociology. A review article. *Comparattive Studies in Society and History, 40*(1), 170-186. doi:10.1017/s0010417598980069

Stige, B., Malterud, K., & Midtgarden, T. (2009). Toward an agenda for evaluation of qualitative research. *Qualitative Health Research, 19*(10), 1504-1516. doi:10.1177/1049732309348501

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25. doi:10.1146/annurev.clinpsy.032408.153639

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi:10.1207/S15327752JPA8001_18

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford: Oxford University Press.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., . . . Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA, 283*(15), 2008-2012. doi:10.1001/jama.283.15.2008

Sutherland, S. E. (2004). An introduction to systematic reviews. *Journal of Evidence Based Dental Practice, 4*(1), 47-51. doi:10.1016/j.jebdp.2004.02.021

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18*(4), 385-401. doi:10.1080/09602010802009201

The Equator Network. (n.d.). EQUATOR: Enhancing the QUAlity and Transparency Of health Research Retrieved 29 January, 2011, from http://www.equator-network.org/

The George Institute for Global Health. (2011). PEDro: Physiotherapy Evidence Database Retrieved 29 January, 2011, from http://www.pedro.org.au/

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 43-58). Thousand Oaks, CA: Sage.

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*(6), 349-357. doi:10.1093/intqhc/mzm042

Treloar, C., Champness, S., Simpson, P. L., & Higginbotham, N. (2000). Critical appraisal checklist for qualitative research studies. *Indian Journal of Pediatrics, 67*(5), 347-351.

Trochim, W. M. (2006). The research methods knowledge base Retrieved 29 January, 2011, from http://www.socialresearchmethods.net/kb/

Uebersax, J. (2010). Statistical methods for rater and diagnostic agreement Retrieved 7 September, 2010, from http://www.john-uebersax.com/

University of Sydney, & Speech Pathology Australia. (2010). speechBITE: Speech Pathology Database for Best Interventions and Treatment Efficacy Retrieved 29 January, 2011, from http://www.speechbite.com/

Urschel, J. D. (2005). How to analyze an article. *World Journal of Surgery, 29*(5), 557-560. doi:10.1007/s00268-005-7912-z

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130-149. doi:10.1037/1082-989X.13.2.130

Verhagen, A. P., de Vet, H. C. W., de Bie, R. A., Kessels, A. G. H., Boers, M., Bouter, L. M., & Knipschild, P. G. (1998). The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of Clinical Epidemiology, 51*(12), 1235-1241. doi:10.1016/S0895-4356(98)00131-0

Vickers, A. (1995). Critical appraisal: How to read a clinical research paper. *Complementary Therapies in Medicine, 3*(3), 158-166. doi:10.1016/S0965-2299(95)80057-3

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine, 4*(10), e296. doi:10.1371/journal.pmed.0040296

Walsh, D., & Downe, S. (2006). Appraising the quality of qualitative research. *Midwifery, 22*(2), 108-119. doi:10.1016/j.midw.2005.05.004

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*(1), 101-110. doi:10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.0.co;2-e

Williams, C., Williams, S., & Appleton, K. (1997). Mind maps: An aid to effective formulation. *Behavioural and Cognitive Psychotherapy, 25*(3), 261-267. doi:10.1017/s1352465800018555

Wilson, A., & Henry, D. A. (1992). Meta-analysis Part 2: Assessing the quality of published meta-analyses. *Medical Journal of Australia, 156*(3), 173-174,177-178,180,184-187.

World Medical Association. (2008). *Declaration of Helsinki: Ethical principles for medical research involving human subjects*. Paper presented at the 59th World Medical Association General Assembly, Seoul, South Korea.

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

PAPERS APPRAISED

Alexander, J. M., McIntire, D. M., & Leveno, K. J. (1999). Chorioamnionitis and the prognosis for term infants. *Obstetrics and Gynecology, 94*(2), 274-278.

Allen, H. M., Borden, S., Pikelny, D. B., Paralkar, S., Slavin, T., & Bunn, W. B. (2003). An intervention to promote appropriate management of allergies in a heavy manufacturing workforce: Evaluating health and productivity outcomes. *Journal of Occupational and Environmental Medicine, 45*(9), 956-972.

Als, H., Lawhon, G., Duffy, F. H., McAnulty, G. B., Gibes-Grossman, R., & Blickman, J. G. (1994). Individualized developmental care for the very low-birth-weight preterm infant: Medical and neurofunctional effects. *JAMA, 272*(11), 853-858.

Appelin, G., & Bertero, C. (2004). Patients' experiences of palliative care in the home: A phenomenological study of a Swedish sample. *Cancer Nursing, 27*(1), 65-70.

Arthur, H., Smith, K., Kodis, J., & McKelvie, R. (2002). A controlled trial of hospital versus home-based exercise in cardiac patients. *Medicine and Science in Sports and Exercise, 34*(10), 1544-1550.

Arts, M. P., Brand, R., van den Akker, E. M., Koes, B. W., Bartels, R. H., & Peul, W. C. (2009). Tubular diskectomy vs conventional microdiskectomy for sciatica: A randomized controlled trial. *JAMA, 302*(2), 149-158.

Averitt, S. S. (2003). "Homelessness is not a choice!!" The plight of homeless women with preschool children living in temporary shelters. *Journal of Family Nursing, 9*(1), 79-100.

Bagshaw, S. M., Berthiaume, L. R., Delaney, A., & Bellomo, R. (2008). Continuous versus intermittent renal replacement therapy for critically ill patients with acute kidney injury: A meta-analysis. *Critical Care Medicine, 36*(2), 610-617.

Bart, C., & Tabone, J. (1999). Mission statement content and hospital performance in the Canadian not-for-profit health care sector. *Health Care Management Review, 24*(3), 18-29.

Beck, C. (1996). Postpartum depressed mothers' experiences interacting with their children. *Nursing Research, 45*(2), 98-104.

Behari, S., Nayak, S. R., Bhargava, V., Banerji, D., Chhabra, D. K., & Jain, V. K. (2003). Craniocervical tuberculosis: Protocol of surgical management. *Neurosurgery, 52*(1), 72-81.

Benatar, M., & Kaminski, H. J. (2007). Evidence report: The medical treatment of ocular myasthenia (an evidence-based review): Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology, 68*(24), 2144-2149.

Benisovich, S., & King, A. C. (2003). Meaning and knowledge of health among older adult immigrants from Russia: A phenomenological study. *Health Education Research, 18*(2), 135-144.

Bergman-Evans, B. (2004). Beyond the basics: Effects of the Eden Alternative model on quality of life issues. *Journal of Gerontological Nursing, 30*(6), 27-34.

Bhattacharyya, N., & Fried, M. P. (2003). The accuracy of computed tomography in the diagnosis of chronic rhinosinusitis. *Laryngoscope, 113*(1), 125-129.

Chatellier, G., Day, M., Bobrie, G., & Menard, J. (1995). Feasibility study of N-of-1 trials with blood pressure self-monitoring in hypertension. *Hypertension, 25*(2), 294-301.

Coughlin, T. A., & Long, S. K. (2000). Effects of Medicaid managed care on adults. *Medical Care, 38*(4), 433-446.

Cournot, M., Marquie, J. C., Ansiau, D., Martinaud, C., Fonds, H., Ferrieres, J., & Ruidavets, J. B. (2006). Relation between body mass index and cognitive function in healthy middle-aged men and women. *Neurology, 67*(7), 1208-1214.

Daly, R. M., Caine, D., Bass, S. L., Pieter, W., & Broekhoff, J. (2005). Growth of highly versus moderately trained competitive female artistic gymnasts. *Medicine and Science in Sports and Exercise, 37*(6), 1053-1060.

Dembinski, R., Hochhausen, N., Terbeck, S., Uhlig, S., Dassow, C., Schneider, M., . . . Kuhlen, R. (2007). Pumpless extracorporeal lung assist for protective mechanical ventilation in experimental lung injury. *Critical Care Medicine, 35*(10), 2359-2366.

Despriet, D. D., Klaver, C. C., Witteman, J. C., Bergen, A. A., Kardys, I., de Maat, M. P., . . . de Jong, P. T. (2006). Complement factor H polymorphism, complement activators, and risk of age-related macular degeneration. *JAMA, 296*(3), 301-309.

Dore, S., Buchan, D., Coulas, S., Hamber, L., Stewart, M., Cowan, D., & Jamieson, L. (1998). Alcohol versus natural drying for newborn cord care. *Journal of Obstetric, Gynecologic, and Neonatal Nursing, 27*(6), 621-627.

Durham, S. R., McComb, J. G., & Levy, M. L. (2003). Correction of large (>25 cm2) cranial defects with "reinforced" hydroxyapatite cement: Technique and complications. *Neurosurgery, 52*(4), 842-845.

Eifried, S. (2003). Bearing witness to suffering: The lived experience of nursing students. *Journal of Nursing Education, 42*(2), 59-67.

Goldbort, J. G. (2009). Women's lived experience of their unexpected birthing process. *MCN: The American Journal of Maternal Child Nursing, 34*(1), 57-62.

Gosain, A. K., Santoro, T. D., Havlik, R. J., Cohen, S. R., & Holmes, R. E. (2002). Midface distraction following Le Fort III and Monobloc osteotomies: Problems and solutions. *Plastic and Reconstructive Surgery, 109*(6), 1797-1808.

Hagen, K. B., Hilde, G., Jamtvedt, G., & Winnem, M. F. (2002). The Cochrane Review of advice to stay active as a single treatment for low back pain and sciatica. *Spine, 27*(16), 1736-1741.

Heetveld, M. J., Raaymakers, E. L. F. B., Luitse, J. S. K., Nijhof, M., & Gouma, D. J. (2007). Femoral neck fractures: Can physiologic status determine treatment choice? *Clinical Orthopaedics and Related Research, 461*, 203-212.

Hinck, S. M. (2007). The meaning of time in oldest-old age. *Holistic Nursing Practice, 21*(1), 35-41.

Iwama, H., Ohmizo, H., Furuta, S., Ohmori, S., Watanabe, K., Kaneko, T., & Tsutsumi, K. (2002). Inspired superoxide anions attenuate blood lactate concentrations in postoperative patients. *Critical Care Medicine, 30*(6), 1246-1249.

Jablonski, R., Reed, D., & Maas, M. (2005). Care intervention for older adults with Alzheimer's disease and related dementias: effect of family involvement on cognitive and functional outcomes in nursing homes. *Journal of Gerontological Nursing, 31*(6), 38-48.

Jais, P., Haissaguerre, M., Shah, D. C., Chouairi, S., Gencel, L., Hocini, M., & Clementy, J. (1997). A focal source of atrial fibrillation treated by discrete radiofrequency ablation. *Circulation, 95*(3), 572-576.

Jiang, R., Manson, J. E., Stampfer, M. J., Liu, S., Willett, W. C., & Hu, F. B. (2002). Nut and peanut butter consumption and risk of Type 2 diabetes in women. *JAMA, 288*(20), 2554-2560.

Jones, A. E., Brown, M. D., Trzeciak, S., Shapiro, N. I., Garrett, J. S., Heffner, A. C., & Kline, J. A. (2008). The effect of a quantitative resuscitation strategy on mortality in patients with sepsis: A meta-analysis. *Critical Care Medicine, 36*(10), 2734-2739.

Kaunonen, M. P. D., Tarkka, M.-T. P. D., Laippala, P., & Paunonen-Ilmonen, M. P. D. (2000). The impact of supportive telephone call intervention on grief after the death of a family member. *Cancer Nursing, 23*(6), 483-491.

King, W. A., Wackym, P. A., Sen, C., Meyer, G. A., Shiau, J., & Deutsch, H. (2001). Adjunctive use of endoscopy during posterior fossa surgery to treat cranial neuropathies. *Neurosurgery, 49*(1), 108-116.

Kontorinis, N., Agarwal, K., & Dieterich, D. (2005). Treatment of hepatitis C virus in HIV patients: A review. *AIDS, 19*(3), S166-S173.

Liu, L.-N., Li, C.-Y., Tang, S. T., Huang, C.-S., & Chiou, A.-F. (2006). Role of continuing supportive cares in increasing social support and reducing perceived uncertainty among women with newly diagnosed breast cancer in Taiwan. *Cancer Nursing, 29*(4), 273-282.

Lobo, S. M., Salgado, P. F., Castillo, V. G., Borim, A. A., Polachini, C. A., Palchetti, J. C., . . . de Oliveira, G. G. (2000). Effects of maximizing oxygen delivery on morbidity and mortality in high-risk surgical patients. *Critical Care Medicine, 28*(10), 3396-3404.

Lundell, J. C., Silverman, D. G., Brull, S. J., O'Connor, T. Z., Kitahata, L. M., Collins, J. G., & LaMotte, R. (1996). Reduction of postburn hyperalgesia after local injection of ketorolac in healthy volunteers. *Anesthesiology, 84*(3), 502-509.

Marcinkowski, K., Wong, V., & Dignam, D. (2005). Getting back to the future: A grounded theory study of the patient perspective of total knee joint arthroplasty. *Orthopaedic Nursing, 24*(3), 202-209.

McGillis Hall, L., Doran, D., & Pink, L. (2008). Outcomes of interventions to improve hospital nursing work environments. *Journal of Nursing Administration, 38*(1), 40-46.

Meighan, M., Davis, M., Thomas, S., & Droppleman, P. (1999). Living with postpartum depression: The father's experience. *MCN: The American Journal of Maternal Child Nursing, 24*(4), 202-208.

Mellegers, M. A., Furlan, A. D., & Mailis, A. (2001). Gabapentin for neuropathic pain: Systematic review of controlled and uncontrolled literature. *Clinical Journal of Pain, 17*(4), 284-295.

Menges, T., Konig, I. R., Hossain, H., Little, S., Tchatalbachev, S., Thierer, F., . . . Bein, G. (2008). Sepsis syndrome and death in trauma patients are associated with variation in the gene encoding tumor necrosis factor. *Critical Care Medicine, 36*(5), 1456-1462 (plus supplementary material).

Mentes, J., & Culp, K. (2003). Reducing hydration-linked events in nursing home residents. *Clinical Nursing Research, 12*(3), 210-225.

Mercat, A., Richard, J.-C. M., Vielle, B., Jaber, S., Osman, D., Diehl, J.-L., . . . Brochard, L. (2008). Positive end-expiratory pressure setting in adults with acute lung injury and acute respiratory distress syndrome: A randomized controlled trial. *JAMA, 299*(6), 646-655.

Mignone, J., & Guidotti, T. L. (1999). Support groups for injured workers: Process and outcomes. *Journal of Occupational and Environmental Medicine, 41*(12), 1059-1064.

Motheral, B., & Fairman, K. A. (2001). Effect of a three-tier prescription copay on pharmaceutical and other medical utilization. *Medical Care, 39*(12), 1293-1304.

Peeters, M. G., Verhagen, A., de Bie, R. A., & Oostendorp, B. R. (2001). The efficacy of conservative treatment in patients with whiplash injury: A systematic review of clinical trials. *Spine, 26*(4), E64-E73.

Polanczyk, G., Zeni, C., Genro, J. P., Guimaraes, A. P., Roman, T., Hutz, M. H., & Rohde, L. A. (2007). Association of the adrenergic α2A receptor gene with methylphenidate improvement of inattentive symptoms in children and adolescents with attention-deficit/hyperactivity disorder. *Archives of General Psychiatry, 64*(2), 218-224.

Presti, C., Puech-Leao, P., & Albers, M. (1999). Superficial femoral eversion endarterectomy combined with a vein segment as a composite artery-vein bypass graft for infrainguinal arterial reconstruction. *Journal of Vascular Surgery, 29*(3), 413-421.

Rao, N., & Regalla, D. M. (2006). Uncertain efficacy of daptomycin for prosthetic joint infections: A prospective case series. *Clinical Orthopaedics and Related Research, 451*(10), 34-37.

Reid, S. A., Speedy, D. B., Thompson, J. M., Noakes, T. D., Mulligan, G., Page, T., . . . Milne, C. (2004). Study of hematological and biochemical parameters in runners completing a standard marathon. *Clinical Journal of Sport Medicine, 14*(6), 344-353.

Rubertsson, S., Grenvik, A., Zemgulis, V., & Wiklund, L. (1995). Systemic perfusion pressure and blood flow before and after administration of epinephrine during experimental cardiopulmonary resuscitation. *Critical Care Medicine, 23*(12), 1984-1996.

Salt, J., Cummings, G. G., & Profetto-McGrath, J. (2008). Increasing retention of new graduate nurses: A systematic review of interventions by healthcare organizations. *Journal of Nursing Administration, 38*(6), 287-296.

Saposnik, G., & Del Brutto, O. H. (2003). Stroke in South America: A systematic review of incidence, prevalence, and stroke subtypes. *Stroke, 34*(9), 2103-2107.

Schluter, W., Judson, F., Baro'n, A., McGill, W., Marine, W., & Douglas, J. (1996). Usefulness of Human Immunodeficiency Virus post-test counseling by telephone for low-risk clients of an urban sexually transmitted diseases clinic. *Sexually Transmitted Diseases, 23*(3), 190-197.

Singh, S., & Kumar, A. (2007). Wernicke encephalopathy after obesity surgery: A systematic review. *Neurology, 68*(11), 807-811.

Sinuff, T., Adhikari, N. K. J., Cook, D. J., Schunemann, H. J., Griffith, L. E., Rocker, G., & Walter, S. D. (2006). Mortality predictions in the intensive care unit: Comparing physicians with scoring systems. *Critical Care Medicine, 34*(3), 878-885.

Smith, A., Lew, R., Shrimpton, C., Evans, R., & Abbenante, G. (2000). A novel stable inhibitor of endopeptidases EC 3.4.24.15 and 3.4.24.16 potentiates bradykinin-induced hypotension. *Hypertension, 35*(2), 626-630.

Taleghani, F., Yekta, Z. P., Nasrabadi, A. N., & Kappeli, S. (2008). Adjustment process in Iranian women with breast cancer. *Cancer Nursing, 31*(3), 32-41.

The Tads Team. (2007). The treatment for adolescents with depression study (TADS): Long-term effectiveness and safety outcomes. *Archives of General Psychiatry, 64*(10), 1132-1143.

van Gils, E. J. M., Veenhoven, R. H., Hak, E., Rodenburg, G. D., Bogaert, D., Ijzerman, E. P., . . . Sanders, E. A. (2009). Effect of reduced-dose schedules with 7-valent pneumococcal conjugate vaccine on nasopharyngeal pneumococcal carriage in children: A randomized controlled trial. *JAMA, 302*(2), 159-167.

Whalen, C. C., Nsubuga, P., Okwera, A., Johnson, J. L., Hom, D. L., Michael, N. L., . . . Ellner, J. J. (2000). Impact of pulmonary tuberculosis on survival of HIV-infected adults: a prospective epidemiologic study in Uganda. *AIDS, 14*(9), 1219-1228.

Whitney, C. M. (2004). Maintaining the square: How older adults with Parkinson's Disease sustain quality in their lives. *Journal of Gerontological Nursing, 30*(1), 28-35.

Full mind map of research methods (see Chapter 3)