Revision of the Mixed Methods Appraisal Tool (MMAT): A Mixed Methods Study

Quan Nha Hong

Department of Family Medicine

McGill University, Montréal

August 2018

A dissertation submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Family Medicine (ad hoc)

ABSTRACT

Background: Systematic mixed studies reviews (SMSRs), i.e., systematic reviews combining qualitative, quantitative and mixed methods studies, are growing in popularity owing to their potential to provide a rich and practical understanding of complex health interventions and problems. Due to the heterogeneity of the included studies, one challenging issue concerns the critical appraisal of studies. A critical appraisal tool was developed to address this challenge: the Mixed Methods Appraisal Tool (MMAT). The MMAT includes criteria for appraising the methodological quality of five categories of studies: (a) qualitative studies, (b) randomized controlled trials, (c) non-randomized studies, (d) quantitative descriptive studies, and (e) mixed methods studies. Pilot studies provided proof-of-concept for the feasibility of the MMAT and a need for further development.

Objectives: The overall objective of this project was to revise the MMAT. The specific objectives were to identify the changes that need to be made in the MMAT and the most relevant criteria that should be included in the MMAT.

Methods: A sequential exploratory mixed methods design was used. A first phase consisted in a qualitative descriptive study. Semi-structured interviews with researchers having used the MMAT were conducted to identify the strengths, limits, and areas for improvement of the tool. Then, the team composed of 12 researchers with complementary expertise in qualitative, quantitative and mixed methods research met to discuss the results and plan the next step. In a second phase, a modified e-Delphi study was performed with experts in qualitative, survey and mixed methods studies to identify the most relevant critical appraisal criteria. Consensus was reached when at least 80% of experts judged a criterion 'very' or 'extremely' relevant. In addition, a mapping of criteria from 33 existing critical appraisal tools was performed to identify the core criteria for randomized controlled trials and non-randomized studies. The results of these two phases informed the development of a revised version of the MMAT (version 2018).

Results: For the first phase, a total of 20 participants from eight different countries were interviewed. Thirteen main themes were identified and grouped into the dimensions of usefulness, i.e., utility and usability. The themes related to utility concerned the coverage, completeness, flexibility, and other utilities of the tool (educational tool). Those on usability

were related to the tool's learnability, efficiency, satisfaction and errors that could be made due to difficulties in understanding or selecting the criteria to rate. For the second phase, respectively 73 and 56 experts participated in Round-one and Round-two of the modified e-Delphi study. The experts were from 11 different countries. Consensus was reached for six qualitative criteria, eight survey criteria, and seven mixed methods criteria. The mapping of the criteria of randomized controlled trials and non-randomized studies led to add new criteria in the MMAT to covers the different categories of bias addressed in critical appraisal tools. On the basis of these results, of the 19 criteria in the MMAT (version 2011), four were removed, seven were reformulated, five were replaced, and ten new were added. Explanations were added in the user manual as well as an algorithm to help reviewers judge and select the criteria to use.

Discussion and conclusion: This project addressed the usefulness and content validity of the MMAT. A revised version of the MMAT was developed and includes 25 criteria on five categories of studies. Changes from the previous version concerned mainly the number of criteria, the user manual, and the overall scoring. This revised version will need to be pilot tested and the website will be modified. Continuous development of the MMAT is required and future research should focus on its validity, reliability, and usefulness.

RÉSUMÉ

Introduction : Les revues systématiques mixtes sont populaires, car elles permettent une compréhension approfondie de phénomènes et interventions complexes. Toutefois, la combinaison d'études quantitatives, qualitatives et méthodes mixtes pose comme défi l'évaluation de la qualité des études. Pour répondre à ce défi, un outil d'évaluation de la qualité méthodologique a été développé: le *Mixed Methods Appraisal Tool* (MMAT). Le MMAT permet d'évaluer cinq types d'études : (a) qualitatives, (b) quantitatives avec répartition aléatoire, (c) quantitatives sans répartition aléatoire, (d) quantitatives descriptives et (e) méthodes mixtes. Des études pilotes sur le MMAT ont démontré sa faisabilité et le besoin de poursuivre son développement.

Objectifs : L'objectif général de ce projet était de mettre à jour le MMAT. Les objectifs spécifiques étaient d'identifier les changements requis à apporter et les critères pertinents à inclure dans le MMAT.

Méthode : Un devis mixte séquentiel exploratoire a été utilisé. Dans une première phase, une étude qualitative descriptive a été menée. Des entrevues semi-structurées ont été réalisées avec des chercheurs qui ont utilisé le MMAT pour identifier ses forces, limites et améliorations requises. Une réunion des membres de l'équipe composée de 12 chercheurs avec des expertises complémentaires en recherche qualitative, quantitative et méthodes mixtes a été organisée pour discuter des résultats et planifier la phase suivante. Dans une deuxième phase, une étude e-Delphi modifiée a été menée avec des experts méthodologiques en recherche qualitative, sondage et méthodes mixtes afin d'identifier les critères d'évaluation jugés les plus pertinents. Un consensus était considéré atteint lorsqu'au moins 80% des experts ont jugé un critère « très » ou « extrêmement » pertinent. En outre, une analyse des critères provenant des outils d'évaluation de la qualité pour des études quantitatives avec et sans répartition aléatoire a été effectuée. Les résultats de ces phases ont servi à développer une nouvelle version du MMAT (version 2018).

Résultats : Dans la première phase, un total de 20 participants de huit pays différents ont été interviewés. Treize thèmes ont été identifiés et regroupés dans les dimensions de l'utilité et l'utilisabilité. Les thèmes liés à l'utilité concernaient la couverture, l'exhaustivité, la flexibilité et une autre utilité de l'outil (outil éducatif). Ceux sur l'utilisabilité étaient liés à la facilité

d’apprentissage, l'efficience, la satisfaction et les erreurs qui peuvent survenir en raison de difficultés dans le choix et la compréhension des critères à évaluer. Dans la deuxième phase, respectivement 73 et 56 experts provenant de 11 pays différents ont participé aux deux rondes de l'étude e-Delphi. Un consensus a été atteint pour six critères sur les études qualitatives, huit sur les sondages et sept sur les méthodes mixtes. L’analyse des critères provenant des outils d’évaluation de la qualité pour des études quantitatives avec et sans répartition aléatoire a permis d’ajouter de nouveaux critères pour couvrir les différentes catégories de biais abordées dans ces outils. Parmi les 19 critères du MMAT (version 2011), quatre ont été retirés, sept reformulés, cinq remplacés et dix ajoutés. Le manuel d’instruction a été modifié pour inclure des explications sur les critères ainsi qu'un algorithme pour aider les réviseurs à choisir les critères à utiliser.

Discussion et conclusion : Ce projet a permis d’étudier l'utilité, l’utilisabilité et la validité du contenu du MMAT. Une nouvelle version du MMAT a été développée et comporte 25 critères sur cinq catégories d'études. Les modifications apportées concernent principalement le nombre de critères, le manuel d’instruction et le score global. Cette nouvelle version sera testée et le site web sera modifié. Un développement continu du MMAT est nécessaire et les recherches futures devront étudier sa validité, fidélité, utilité et utilisabilité.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CASP	Critical Appraisal Skills Programme
CAT	Critical Appraisal Tool
CERQual	Confidence in the Evidence from Reviews of Qualitative Research
CIHR	Canadian Institutes of Health Research
COREQ	COnsolidated criteria for REporting Qualitative research
EPHPP	Effective Public Health Practice Project
EQUATOR	Enhancing the QUAlity and Transparency Of health Research Network
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HTA	Health Technology Assessment
JBI	Joanna Briggs Institute
MMAT	Mixed Methods Appraisal Tool
NICE	National Institute for Health and Clinical Excellence
NRS	Non-Randomized Study
QARI	Qualitative Assessment and Review Instrument
QATSDD	Quality Assessment Tool for Studies with Diverse Designs
QualSyst	QUALity assessment SYSTem
RCT	Randomized Controlled Trial
RoB	Risk of Bias
ROBINS-E	Risk Of Bias In Non-randomized Studies - of Exposures
ROBINS-I	Risk Of Bias In Non-randomized Studies - of Interventions
SIGN	Scottish Intercollegiate Guidelines Network
SMSR	Systematic Mixed Studies Review
STROBE	Strengthening The Reporting of OBservational studies in Epidemiology

LIST OF FIGURES

From PAPER #1

From PAPER #2

LIST OF TABLES

From PAPER #1

From PAPER #2

From PAPER #3

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Pierre Pluye, for his dedication, expertise, guidance, time, and trust. Thank you for providing me multiple training and research opportunities, and pushing me to surpass my comfort zones.

Thank you to the Canadian Institutes of Health Research (CIHR) for their scholarship that allowed me to fully dedicate my time and effort to my doctoral training.

I am grateful to the members of my thesis committee, Drs. Gillian Bartlett, Pierre Dagenais, Belinda Nicolau, and Isabelle Vedel, for their constructive advice and support. Thank you to Dr. Mathieu Ouimet (external examiner) and Dr. Melissa Park (internal examiner) for their constructive comments on this thesis.

This project would not have been possible without the generous participation of numerous researchers, professors, graduate students, research assistants and librarians. Research is very demanding and I cannot thank enough all the participants for their time and expertise.

Thank you to the MMAT developers who provided precious feedback for improving the MMAT: Drs. Gillian Bartlett, Felicity Boardman, Margaret Cargo, Pierre Dagenais, Sergi Fàbregues, Marie-Pierre Gagnon, Frances Griffiths, Belinda Nicolau, Alicia O'Cathain, Marie-Claude Rousseau, and Isabelle Vedel. In particular, my sincere gratitude to Dr. Fàbregues who generously shared his experience and results of his review on the quality of mixed methods.

Thank you to all my colleagues at the Department of Family Medicine at McGill University for their intellectual support and encouragement. In particular, my deepest thanks to Reem El Sherif, for being there whenever I needed help; Paula Bush, the best English scientific editor; Vera Granikov, my favourite librarian; Araceli Gonzalez-Reyes, second coder for the analysis of interviews; Mathieu Bujold, my NVivo mentor and second reviewer of full-texts; and Maggy Wassef, second reviewer of titles and abstracts. I will miss our Information Technology Primary Care Research Group (ITPCRG) weekly meetings led by Drs. Roland Grad and Pierre Pluye, and coordinated by Vinita D'Souza. These meetings were great opportunities to learn, explore new ideas, and discuss the challenges of research.

Last but not least, thank you to my family and friends for their moral support and understanding. I dedicate this work to my loving parents who have unconditionally supported me throughout my life and provided me the opportunity to live in a free and safe environment.

PREFACE, STATEMENT OF ORIGINALITY, AND CONTRIBUTIONS OF AUTHORS

This is a manuscript-based dissertation that includes three papers. The tool studied in this project was originally conceptualized and developed by Dr. Pierre Pluye and collaborators. As a doctoral candidate, I was responsible for conceiving and planning the different phases of this project. I significantly contributed to the literature review, data collection, analysis, and interpretation of the results, and wrote the manuscripts. All co-authors read and approved the final version of the manuscripts, and provided written permission to include the manuscripts in this thesis. Here are the contributions of the authors for each paper.

Paper #1: Hong, Q.N., & Pluye, P. (2018). A conceptual framework for critical appraisal in systematic mixed studies reviews. *Journal of Mixed Methods Research*. Advance online publication. DOI: 10.1177/1558689818770058.

> QNH conducted the literature review, conceptualized the framework, and wrote the manuscript. PP contributed to the definition of the conceptual boundaries and conceptualization of the framework.

Paper #2: Hong, Q.N., Gonzalez-Reyes, A., & Pluye, P. (2018). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). *Journal of Evaluation in Clinical Practice*, 24(3), 459-467.

> QNH conceived the research design, performed data collection, analysis and interpretation, and wrote the manuscript. AGR performed data analysis and interpretation. PP contributed to the design and data interpretation.

Paper #3: Hong, Q.N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. Improving the content validity of the Mixed Methods Appraisal Tool (MMAT): A modified e-Delphi study. *Journal of Clinical Epidemiology* (submitted).

> QNH conducted the literature review, prepared the questionnaires, identified the experts, analyzed and interpreted the data, and wrote the manuscript. All the authors contributed to the selection and clarification of items included in the modified e-Delphi study, checked the list of experts, and commented several versions of the manuscript.

CHAPTER 1. INTRODUCTION

Critical appraisal of the quality of studies is an important and challenging step in systematic reviews. This project is on a critical appraisal tool that was developed for systematic reviews including different studies designs (quantitative, qualitative, and mixed methods studies): the Mixed Methods Appraisal Tool (MMAT). This chapter presents the definition and a historical overview of systematic reviews. Then, it provides information on the importance of critical appraisal, describes the MMAT, and summarizes previous studies on this tool. It ends with the research questions addressed in this doctoral project and the structure of this dissertation.

1.1 Definition of Systematic Review

Literature review consists in synthesizing, summarizing, combining, analyzing, commenting and criticizing studies on a given subject. It is considered as secondary research in which the unit of analysis is primary studies (Hong, Pluye, Bujold, & Wassef, 2017). Traditionally, literature review has addressed broad review questions with no clear method for the selection, appraisal and synthesis of studies. This type of review was criticized for being subjective, scientifically unsound, and inefficient to extract information (Light & Pillemer, 1984). Traditional literature review is subject to several biases that affect the results produced. Table 1 presents some biases that can be found when performing a literature review (Booth, Papaioannou, & Sutton, 2012; Higgins & Green, 2008; Light & Pillemer, 1984; Martin, Renaud, & Dagenais, 2013; Whiting et al., 2016).

To limit biases, a systematic approach to literature review was developed. Systematic review is defined as the "The application of strategies that limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic" (Chalmers, Hedges, & Cooper, 2002, p. 17). It follows an explicit, transparent, and reproducible process with the following characteristics (Pluye, Hong, Bush, & Vedel, 2016):

1. Specific/focused review question(s);
2. Pre-established precise eligibility criteria;
3. Exhaustive literature search using several sources of information;

4. Comprehensive and detailed search strategy designed with specialized librarians;
5. Reliable or dependable (performed by at least two reviewers) selection of relevant studies, data extraction, and critical appraisal; and
6. Rigorous synthesis using specific qualitative, quantitative and mixed methods.

Table 1. Main Biases in Literature Reviews

Bias	Definition	Strategies to minimize this bias
Identification	Not all studies on a subject of interest are found (not exhaustive search).	• Use several sources of information and bibliographic databases. • Involve a specialized librarian to develop a comprehensive search strategy.
Selection	Arbitrary selection of studies (e.g., a reviewer voluntary include or exclude studies to support position).	• Define clear selection criteria. • Involve a least two reviewers in the selection.
Reporting	The publication of research findings is influenced by the nature and direction of results (e.g., studies with positive results are more likely to be published).	• Use different sources of information. • Include a variety of literature including the grey literature (e.g., theses and dissertations, reports, conference abstracts). • Contact authors of studies to obtain missing information.
Interpretation	The appraisal and synthesis of studies are influenced by a reviewer's subjectivity (e.g., preconceived ideas).	• Use critical appraisal tools. • Involve a least two reviewers in the synthesis and interpretation of findings.

Nowadays, systematic reviews are considered the gold standard in literature reviews. They are one of the preferred methods for collecting scientific evidence to support the development of recommendations on best practices such as clinical practice guidelines (National Health and Medical Research Council, 1998), health technology assessments (Busse et al., 2002), and research synopses (Grad et al., 2008; Pluye et al., 2012). Numerous potential impacts of systematic reviews have been described: (a) save lives by distinguishing what works and what is useless and even harmful, (b) save resources by avoiding unnecessary or unproven treatment as well as unnecessary duplication of clinical trials, and (c) improve practice, clinical quality, and policies by providing up-to-date evidence on specific topics for decision-making (Bunn et

al., 2015; Moynihan, 2004).

1.2 History of Systematic Reviews

To better understand the origin and evolution of systematic reviews, a historical overview was performed. Systematic reviews have a long history. Back in the 18th century, Dr. James Lind emphasized the importance of producing a full and impartial critical view of the existing literature on the treatment of scurvy (Lind, 1757). Chalmers et al. (2002) traced some reviews published in the early 20th century conducted in several fields such as medicine, agriculture, physics, education, and social sciences. Although several papers advocating or using systematic methods to review the literature can be found in the first half of the 20th century, it is mainly in the 1970s that the science of research synthesis began to gain prominence from the need to apply explicit, transparent, and rigorous methods to enhance the validity of reviews (Chalmers et al., 2002). From this point on, it is possible to identify three major periods in the evolution of systematic reviews: (a) foundation (1970 - 1989), (b) institutionalization (1990 - 2000) and (c) diversification (2001 -). Figure 1 highlights some salient features of each period that are related to the users, methodological influences, and technological developments. The following sections will further describe each period.

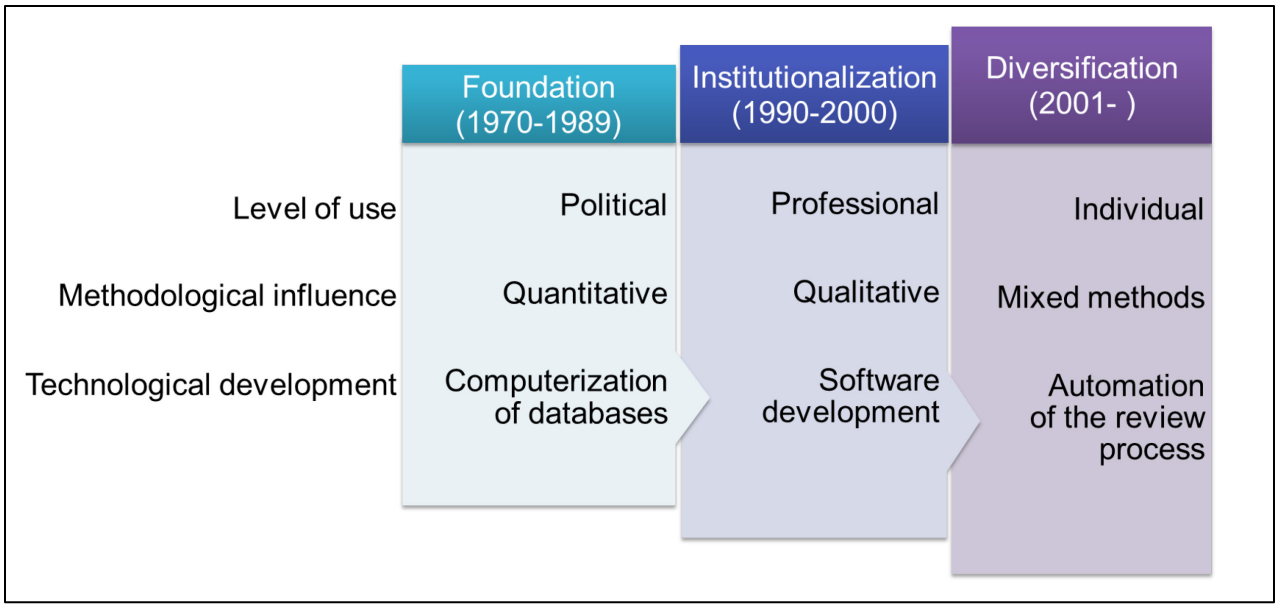


Figure 1. Main Periods in the History of Systematic Reviews

1.2.1 Foundation period (1970 - 1989)

A growing interest in the use of reviews for policy development can be seen in the late 1960s. This interest has had a major impact on the methodological development of systematic reviews. In research, literature review is an essential step for gathering information about what has been done on a subject and what results have been achieved. The results of this review can lead to identify areas of disagreement within the current state of knowledge that researchers can rely on to justify the relevance of a project. The identification of contradictory studies on a subject is not necessarily perceived as problematic by researchers. However, when applied in the political field, inconsistencies between studies can be considered as a source of confusion and a hindrance to the development of clear policies (Light & Smith, 1971). To cope with this difficulty, aggregative synthesis methods were developed to combine the results of various studies. Also, there was a need to systematize the review process to limit the arbitrary selection of studies and increase rigor in the analysis of studies.

The growing interest in supporting the development of policies based on scientific evidence had resulted, among other things, in the development of a multidisciplinary field called health technology assessment (HTA). In particular, HTA conducts systematic reviews to evaluate various dimensions of a technology (e.g., safety, effectiveness, cost-effectiveness, ethical implications) and to make policy recommendations about its use (Busse et al., 2002). Health technology is a very broad concept and can include drugs, devices, medical and surgical procedures as well as organizational and support systems within which health care is delivered (Busse et al., 2002). One of the first HTA agencies was established in 1972 in the United States (US Congressional Office of Technology Assessment) whose mandate was to advise the government on the use and application of health technologies (Banta, 2003).

Another milestone in the history of this foundation period is the computerization of bibliographic indexes that occurred primarily in the 1960s and 1970s. For example, one of the most popular bibliographic databases in medicine, the Index Medicus, was developed in 1879. Its computerized version was introduced in 1964 (MEDLARS), an online version was developed in 1971 (MEDLINE), and a free online version was made available in 1996 (PubMed) (Office of Health Technology Assessment, 1982). A similar development can be observed for other bibliographic databases in sciences such as the Sciences Citation Index inaugurated in 1963

(Office of Health Technology Assessment, 1982). This computerized access to bibliographic catalogues greatly simplified the task of searching for relevant studies and facilitated the production of systematic reviews.

During this period, there was a growth in the importance and number of randomized controlled trials (RCTs). In the 1960s and early 1970s, Dr. Donald Campbell (in social sciences) and Dr. Archie Cochrane (in medicine) advocated the use of RCTs as a standard for evaluating the effectiveness of an intervention; they considered RCTs as the most reliable source of scientific evidence (Campbell & Stanley, 1963; Cochrane, 1972). To synthesize RCTs, a statistical synthesis method to aggregate effect sizes of studies was developed. This method was named 'meta-analysis' for the first time in 1976 (Glass, 1976). Another synthesis methods developed during this period is quantitative case survey, which consists in analyzing the content of case studies by using closed-ended questions (Yin & Heald, 1975). This aggregative synthesis method was developed in political science in which case study is a commonly used methodology.

In summary, this foundational period was influenced by the need for reliable and reproducible reviews for policy development, the expansion of RCTs, and the computerization of bibliographic databases. It is also characterized by the development of aggregative synthesis methods to combine results of studies.

1.2.2 Institutionalization period (1990 - 2000)

The institutionalization period is marked by the development of the evidence movement, which was introduced in the early 1990s and advocates the "conscientious, explicit and judicious use of the current best evidence in making decisions about the care of individual patients" (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996, p. 71). Initially developed in the field of medicine (evidence-based medicine), this movement has spread to other areas such as management (Lomas, Culyer, McCutcheon, McAuley, & Law, 2005), and education (Davies, 1999). In this evidence movement, the findings from systematic reviews are not only useful at a political level, but also at a professional level such as for health-care professionals who use evidence for decision-making in their clinical practice.

During this period, several organizations dedicated to the production of systematic reviews were established to promote evidence-based practice. One of the most important organizations developed in response to Dr. Archie Cochrane's call for systematic reviews of RCTs is the Cochrane Collaboration, which was founded in 1993 (Chalmers, 1993). This non-profit, non-governmental organization works with volunteer experts around the world to conduct systematic reviews of health interventions. Currently, more than 37,000 contributors from 130 countries are involved in the Cochrane Collaboration, making it the world's largest producer of systematic reviews in health care (The Cochrane Collaboration, 2017). A similar organization, the Campbell Collaboration, was inaugurated in 2000 in honour of Dr. Donald Campbell. This organization focuses on the effect of intervention in the areas of education, crime and justice, disability and social welfare, and international development (Schuerman et al., 2002).

Part of this organizational development, several guidelines on systematic reviews and tools to facilitate their production were created. For example, critical appraisal tools have been developed to help reviewers judge the quality of RCTs. Also, levels of evidence (also named hierarchy of evidence) have been developed to help assess the scientific evidence (Evans, 2003). In this hierarchy of evidence, systematic reviews with meta-analysis are assigned the highest level of evidence. Moreover, computer software has been created to facilitate the production of reviews (e.g., software for managing bibliographic reference, for coding studies, and for performing meta-analysis). These developments have contributed to the standardization of the systematic review procedure and to the growth in the number of reviews published. For instance, fewer than 10 systematic reviews could be annually found in the 1980s in the PubMed database, while their number has steadily increased since the 1990s, reaching more than 5,000 in 2000 and more than 28,000 in 2014 (Ioannidis, 2016). Also, Ioannidis (2016) compared the annual rate of systematic reviews published from 1991 to 2014. He found that the annual publication rate of systematic reviews increased by 2,728% in PubMed compared to only 153% of increase for all PubMed indexed items.

At the methodological level, further development of meta-analysis methods can be seen during this period (Sutton, Jones, Abrams, Sheldon, & Song, 1999). For example, one development consisted of testing Bayesian meta-analysis for incorporating qualitative and quantitative evidence in the context of the uptake of childhood immunization (Jones, Dixon-

Woods, Abrams, & Fitzpatrick, 1999). Also, because RCTs are not always available and feasible in some fields, other aggregative synthesis methods have been developed such as cross-design synthesis for combining the results of studies with complementary designs (Droitcour, Silberman, & Chelimsky, 1993). At the end of the 1980s, researchers in social sciences were interested in including qualitative studies in reviews (Noblit & Hare, 1988). During this institutionalization period, several interpretive (also called configurative) synthesis approaches were developed such as meta-ethnography (Noblit & Hare, 1988), meta-synthesis (Sandelowski, Docherty, & Emden, 1997), and meta-study (Zhao, 1991). These synthesis methods aim at generating new ways of understanding a phenomenon (Gough, Thomas, & Oliver, 2012).

In summary, the period of institutionalization is marked by the establishment of various organizations promoting evidence-based practice, and developing standards and tools for producing systematic reviews. Besides from RCTs, other types of studies are considered in reviews, especially in social sciences such as qualitative studies. Both aggregative and interpretive synthesis methods were developed.

1.2.3 Diversification period (2001 -)

This period can be characterized by the diversification of synthesis methods and types of reviews as well as diversification of users. In addition to political and professional levels, the use of systematic reviews can also be seen at an individual level. For instance, consumers (e.g., patients) play a more active role in decision-making as part of person-centred care (Olsson, Jakobsson Ung, Swedberg, & Ekman, 2013). Also, during this period, more emphasis was put on knowledge translation to close the gap between practice and research (Sudsawad, 2007). This trend influenced how systematic reviews are produced and disseminated. For example, studies have been conducted on the effectiveness of dissemination strategies to facilitate the uptake of systematic reviews (Tricco et al., 2015). Also, increasing involvement of stakeholders (e.g., consumers, clinicians, and policymakers) in the production of systematic reviews has been advocated (Cottrell et al., 2014; Morley, Norman, Golder, & Griffith, 2016).

During this period, the number of scientific documents available has considerably increased. Systematic reviews are used to keep up-to-date and cope with the huge volume of scientific publications (Bastian, Glasziou, & Chalmers, 2010). For example, 10 million

documents were indexed in PubMed in 2006 while this number exceeded 20 million in 2014. This means that it took less than 10 years (2007-2014) to reach the same number of documents (10 million) that took more than 100 years to index (1865 - 2006) (Bastian et al., 2010; U.S. National Library of Medicine, 2013). The number of scientific publications worldwide was estimated at 50 million and about 2.5 million articles are published every year in peer-reviewed journals (Bjork, Roos, & Lauri, 2009; Jinha, 2010; Ware & Mabe, 2015). There are more than 33,000 peer-reviewed academic journals and this number continues to grow exponentially (Gu & Blackmore, 2016). For example, from 1986 to 2013, it was found that the growth rate of academic journals was 4.7% on average (Gu & Blackmore, 2016). This growth creates new challenges for the production of systematic reviews (e.g., increased time and resources for the selection of studies). With more literature to cover and more complex issues to consider, it is also required to produce and disseminate systematic reviews faster. New types of reviews have been developed to reduce their production time such as rapid review (Tricco, Langlois, & Straus, 2017) and review of systematic reviews (Robinson et al., 2016). Moreover, to reduce the burden of identifying and categorizing studies, researchers have been interested in studying the use of text mining tools, which involve machine learning to allow automation of specific systematic review tasks (Thomas, McNaught, & Ananiadou, 2011). Also, within this growing research industry, a new interdisciplinary discipline called meta-research has emerged (Ioannidis, 2018). It consists of research on research in order to evaluate and improve scientific methods and practices (Ioannidis, Fanelli, Dunne, & Goodman, 2015).

In the previous periods, systematic reviews mainly focused on RCTs. However, these reviews are problematic in areas where research is dominated by non-trial quantitative evidence such as public health, rehabilitation, and primary care (Goldsmith, Bankhead, & Austoker, 2007). Also, these reviews only address effectiveness questions such as ‘Does it work?’ and ‘What works for whom?’. Besides from the effectiveness of an intervention, it was advocated to consider other issues such as its acceptability, applicability, feasibility, and transferability in different contexts (Petticrew et al., 2013; Shaw, Larkin, & Flowers, 2014). Other review questions can be asked such as ‘Why does it work?’, ‘How does it work?’, ‘In what context and when?’, and ‘What are the factors that promote or hinder the implementation of an intervention?’. These questions can be addressed by reviewing other types of studies such as

qualitative and mixed methods studies. A new type of reviews was developed to provide a more complete picture of the evidence and to address different questions: systematic mixed studies reviews (SMSRs), i.e., systematic reviews of qualitative, quantitative and mixed methods studies (Pluye & Hong, 2014). SMSRs combine the strengths of both qualitative and quantitative evidence and provide in-depth answers to review questions involving complex phenomena (Petticrew et al., 2013; Pluye, Gagnon, Griffiths, & Johnson-Lafleur, 2009; Whittemore, Chao, Jang, Minges, & Park, 2014).

In the previous periods, the synthesis methods were mainly designed for one type of data (quantitative or qualitative). In this period, new interpretive methods for synthesizing qualitative and quantitative evidence were developed such as meta-narrative synthesis (Greenhalgh et al., 2005), critical interpretive synthesis (Dixon-Woods et al., 2006), and realist synthesis (Pawson, Greenhalgh, Harvey, & Walshe, 2005). Also, several papers on the conceptualization of the synthesis in SMSRs were written (Frantzen & Fetters, 2015; Heyvaert, Maes, & Onghena, 2013b; Hong et al., 2017; Sandelowski, Voils, Leeman, & Crandell, 2012). The conceptualization of SMSRs was highly influenced by the development of the mixed methods research, which combines methods of collecting and analyzing qualitative and quantitative primary data using experimentation, observation and simulation (Abbott, 1998; Creswell & Plano Clark, 2018).

In summary, the conceptual and methodological development of systematic reviews during this diversification period was influenced by the importance of knowledge translation to all users (at political, professional and individual levels), the explosion of the number of available documents, and the development of mixed methods research. New tools have been explored to automate the review process. Also, new types of reviews and synthesis methods have been developed to address other reviews questions, to accelerate the review process, and to deal with different data.

1.3 Importance of Critical Appraisal

Critical appraisal of included studies is a core step of systematic reviews. It consists in a systematic and careful examination of studies to ensure they are trustworthy, valid and reliable (Burls, 2009; Harden & Gough, 2012). Discussion about judging the quality of included papers

in systematic reviews started mainly with the seminal paper of Glass in 1976 on the development of meta-analysis (Wortman, 1994). Glass' paper opened a debate on whether the quality of a design influenced the findings in meta-analysis (Glass, 1976). Although previous studies had found a relationship between study quality and study outcome (i.e., poorly designed studies tended to provide more positive findings, whereas well-designed studies showed no effect), his position was "I believe the difference to be so small that to integrate research results by eliminating the 'poorly done' studies is to discard a vast amount of important data" (Glass, 1976, p. 4). This position generated much debate not only on whether to exclude poor quality studies but also on how to appraise the quality of studies (Wortman, 1994). Nowadays, critical appraisal has become an integral step in systematic reviews. The rationale for performing critical appraisal is that the inclusion of studies that are methodologically flawed can lead to weaknesses in the results and conclusions drawn from a systematic review (Higgins & Green, 2008). Without a proper appraisal, the conclusions of a review might be misleading or even wrong (Higgins & Green, 2008).

Since the results from systematic reviews are used to guide decision-making, it is of prime importance that a rigorous process is used in the appraisal of studies to ensure that the conclusions properly reflect the quality of evidence reviewed. More than 40 years after Glass' seminal paper, there is still much debate on critical appraisal that mainly focuses on what should be appraised, how it should be performed, and how should the results of the appraisal be used in a review (Carroll & Booth, 2015; Glasziou, Vandenbroucke, & Chalmers, 2004). Because reviewers' judgment of a same study can greatly vary, critical appraisal tools (also named quality assessment tools and risk of bias tools) were developed to help reviewers appraise the quality of studies on a more consistent and objective basis (Wells & Littell, 2009; Wortman, 1994). A variety of critical appraisal tools (scales or checklists) has been developed to formalize the appraisal process and ensure it is done in a systematic, transparent and reproducible manner (Petticrew & Roberts, 2006). Currently, over 500 critical appraisal tools have been developed and several reviews of these tools have been conducted (Bai, Shukla, Bak, & Wells, 2012; Crowe & Sheppard, 2011; Deeks et al., 2003; Katrak, Bialocerkowski, Massy-Westropp, Kumar, & Grimmer, 2004; Santiago-Delefosse, Gavin, Bruchez, Roux, & Stephen, 2016; West et al., 2002).

The results of the critical appraisal of individual studies are used to assess the overall quality of evidence and strength of the recommendations, i.e., to judge how much confidence to place in the body of evidence (relevant results of all studies included in a systematic review). Several approaches for rating the overall quality of evidence have been developed, such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) (Guyatt et al., 2011a) and the Confidence in the Evidence from Reviews of Qualitative Research (GRADE-CERQual) (Lewin et al., 2015). In these approaches, the methodological quality of individual studies (or risk of bias) is one factor that is considered among others such as the relevance of the evidence (population, intervention, comparator, outcome) to answer the review question (indirectness), the variation across studies (inconsistency), and random error on evidence (imprecision) (Guyatt et al., 2011b; Guyatt et al., 2011c; Guyatt et al., 2011d).

This dissertation focuses on critical appraisal in SMSRs. Critical appraisal is challenging in SMSRs because this type of review combines findings from heterogeneous study designs (qualitative, quantitative, and mixed methods studies). Each study design has specific characteristics that might preclude the development of a same set of criteria to appraise the quality of all studies. Also, still very few tools have been developed for appraising the quality of studies included in SMSRs. To contribute to the appraisal stage of SMSRs, a tool was developed: the Mixed Methods Appraisal Tool (MMAT) (Pluye et al., 2009). To our knowledge, the MMAT is the only published tool for assessing the methodological quality of different study designs including mixed methods studies. Since its development, a pilot study showed that the MMAT was useful and relevant to researchers, research professionals, graduate students, and members of the Cochrane collaboration (Pace et al., 2012). Also, the interrater reliability of some criteria of the MMAT has been pilot tested showing poor to perfect score and revealed a need for further testing and refinement (Pace et al., 2012; Souto et al., 2015). The aim of this project is to revise the MMAT.

1.4 The Mixed Methods Appraisal Tool (MMAT)

1.4.1 Description of the MMAT

The MMAT was developed more than 10 years ago and was first published in 2009. It was

created from a literature review on mixed studies reviews and in line with a social constructionist worldview (Pluye et al., 2009). The first version of the MMAT included 15 criteria on four categories of studies (qualitative, quantitative experimental, quantitative observational, and mixed methods). It was presented as a proof-of-concept.

The MMAT was last updated in 2011 (see Appendix 1) (Pluye et al., 2011). The MMAT (version 2011) is available online (http://mixedmethodsappraisaltoolpublic.pbworks.com) and comes with a user manual (tutorial) in which each criterion is described and examples are provided. This version proposes two screening questions and 19 criteria for appraising the methodological quality of five categories of studies: (a) qualitative studies (4 criteria), (b) RCT (4 criteria), (c) non-randomized studies (NRS) (4 criteria), (d) quantitative descriptive studies (4 criteria), and (e) mixed methods studies (3 criteria). The screening questions are used to exclude papers that are not empirical studies and thus cannot be appraised using the MMAT. The MMAT was conceived so that one set of criteria can be used when appraising a qualitative or quantitative study. It does not follow the hierarchy of evidence approach that sees some designs as more robust than others and which grade the levels of evidence (Atkins et al., 2004).

When appraising mixed methods studies, three sets are assessed: (a) the qualitative studies set, (b) a quantitative studies set (either, the RCT, the NRS, or the quantitative descriptive studies set), and (c) the mixed methods studies set. In doing so, it acknowledges the methodological distinctive characteristics specific to each method used in mixed methods studies (i.e., quantitative, qualitative, and mixed methods). This can be particularly convenient for assessing multiphase mixed methods studies that use more than one qualitative and quantitative methods.

Each criterion is rated on a categorical scale: yes, no, and can't tell. An overall score can be calculated by counting the number of criteria rated 'yes'. For mixed methods studies, the overall score corresponds to the lowest score of one of the set assessed (either qualitative, quantitative, or mixed methods components) (Pluye et al., 2011).

1.4.2 Previous studies on the MMAT

Since its development, the MMAT has been pilot tested on four occasions. First, during the summer 2009, four reviewers tested six participatory research studies using the MMAT. This

led to some changes of the 15-item original version of the MMAT such as the rewording of criteria and the addition of criteria for assessing NRS (4 criteria) (Pace et al., 2012).

Second, the MMAT was used and discussed during four 90-minute workshops involving diverse audiences: graduate students enrolled in a mixed methods research course; researchers and research professionals with experience in qualitative, quantitative, and mixed methods research; and members of the Cochrane Collaboration (Pace et al., 2012). These workshops showed the potential usefulness of developing critical appraisal tools such as the MMAT. For example, at a Cochrane Colloquium workshop, 21 of the 23 (91%) attendees reported that having a valid and reliable mixed methods appraisal instrument, the MMAT or an equivalent, is 'essential' or 'absolutely essential'.

Third, the interrater reliability of the MMAT was assessed in 2010. A total of 32 studies (eight qualitative studies, eight RCT, six NRS, nine observational studies, and one mixed methods study) were appraised by two independent reviewers using the MMAT. The appraisal mean time per article was 14 minutes (ranging from 4 to 40 minutes). Based on Landis and Koch (1977) interpretation of observer agreement (<0.00 = poor; 0.00–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–1.00 = almost perfect), the strength of agreement between reviewers can be considered substantial with respect to the overall quality score of appraised studies (ICC = 0.72 [0.49-0.85 with 95% CI]), and poor to almost perfect regarding each criterion (Pace et al., 2012). The greatest disagreements between reviewers were observed for the qualitative studies and the NRS sets, and the reliability of the mixed methods studies set was not calculated due to lack of studies.

Fourth, in 2014, the interrater reliability of the MMAT was assessed using data from two previous SMSRs that appraised a total of 261 studies (Souto et al., 2015). In each review, two independent reviewers appraised the quality of the studies using the MMAT. The reliability of the MMAT varied based on the categories of studies: substantial agreement for criteria of mixed methods studies (n=27 papers; k=0.72), moderate agreement for criteria of RCTs (n=72; k=0.53), fair agreement for criteria of qualitative studies (n=140; k=0.29), and slight agreement for criteria of NRS (n=22; k=0.15) (Souto et al., 2015). This study showed the need for clarification of criteria in the MMAT, particularly those related to NRS and qualitative studies. Also, the criteria on quantitative descriptive studies were not assessed in this study and remained to be

tested.

In summary, the pilot studies provided proof-of-concept for the feasibility of the MMAT. These studies also identified several issues regarding its reliability especially the criteria in the NRS and the qualitative studies sets. In addition, the mixed methods, NRS and descriptive studies sets have only been tested with a limited number of studies. Moreover, the literature on mixed methods research has grown considerably over the past decade and there is a need to develop new criteria in line with recent advances in mixed methods research. This demonstrates a need for further development and testing of the MMAT.

1.5 Research Questions

The overall objective of this project was to revise the MMAT. The specific research questions were:

(1) Based on the experience of users of the MMAT, what changes need to be made in the MMAT?
(2) Based on research experts, what are the most relevant criteria that should be included in the MMAT?

1.6 Structure of This Dissertation

This dissertation is organized into eight chapters:

- Chapter 1 introduces this project by providing a definition and a historical overview of systematic reviews, and justifying the importance of critical appraisal. Then, the tool under investigation in this project (i.e., the MMAT) is described. This chapter ends with the research questions addressed in this project.
- Chapter 2 clarifies the definition of SMSRs and presents a conceptual framework of critical appraisal in SMSRs. This framework consists of Paper #1 of this manuscript-based dissertation.

- Chapter 3 provides an overview of the current state of knowledge on critical appraisal in SMSRs and critical appraisal tools. The methods and results of a literature review are presented.
- Chapter 4 addresses the methodology of this project. The overall study design consisting in a two-phase mixed methods design is described. Then, the methodology used in each phase and its justification are presented. This chapter ends with the ethical considerations.
- Chapter 5 presents the methods and results of phase 1 that aimed to identify the views and experience of MMAT users. A qualitative descriptive approach was used in this phase. This chapter consists of Paper #2 of this manuscript-based dissertation.
- Chapter 6 presents the methods and results of phase 2 that aimed to identify relevant criteria for appraising qualitative, survey and mixed methods research using a group consensus technique (modified e-Delphi technique). This chapter consists of Paper #3 of this manuscript-based dissertation.
- Chapter 7 discusses the findings of this project and presents a revised version of the MMAT. Also, the MMAT is compared with other critical appraisal tools. This chapter ends with the limitations, strengths and contributions of this project.
- Chapter 8 concludes this dissertation with a summary of the main research findings as well as some final remarks and directions for future research.

CHAPTER 2. CONCEPTUAL FRAMEWORK

In this chapter, critical appraisal and SMSRs are defined and conceptualized in a framework. This chapter is presented in the format of a manuscript and consists of the first paper of this manuscript-based dissertation. Prior to introducing the framework, a subsection on the terminology is presented to clarify why the term SMSR is used.

2.1 Clarifying the Terminology of Systematic Mixed Studies Reviews (SMSRs)

Different types of reviews have been developed over the past decades. Several terminology problems are encountered from this diversity. First, a myriad of terms are used to designate review papers. Grant and Booth (2009) proposed a typology of 14 types that differ based on the methods used for the search, appraisal, synthesis, and analysis. However, in addition to the types proposed in Grant and Booth (2009), other terms have been used to name a review such as intervention review, diagnostic test review, realist review, meta-ethnography review, and meta-narrative review (Tricco, Tetzlaff, & Moher, 2011). Second, there are different combinations of terms and the meaning of the terms is not uniformly understood and used. For example, the word 'systematic' is used in different reviews such as systematic mapping review (O'Cathain, Thomas, Drabble, Rudolph, & Hewison, 2013), systematic scoping review (Wilson, Petticrew, Calnan, & Nazareth, 2010), rapid systematic review (Fitzpatrick-Lewis et al., 2011), critical systematic review (Hartikainen, Lönnroos, & Louhivuori, 2007), mixed-methods systematic review (Lawrence & Kinn, 2012), and systematic narrative review (Powell, Rushmer, & Davies, 2009). For some, the term 'systematic' refers to a review that followed a structured process. For others, it means that the review process included a comprehensive search of the literature, critical appraisal of included papers, and two independent reviewers for the selection and appraisal of papers (Pai et al., 2003). Currently, there are different types of reviews and similar terms are used with different meanings. Third, several terms are used to designate a same type of review. For instance, review including reviews have been named umbrella review, review of reviews, overview of systematic reviews, and meta-review (Hunt, Pollock, Campbell, Estcourt, & Brunton, 2018).

Several typologies of reviews have been suggested (Cooper, 1988; Grant & Booth, 2009; Munn, Stern, Aromataris, Lockwood, & Jordan, 2018; Paré, Trudel, Jaana, & Kitsiou, 2015;

Schryen, Wagner, & Benlian, 2015). In general, six main categories of terms can be identified (Table 2).

Table 2. Categories of Terms Used to Designate a Review

Category	Examples of review names used
Purpose	Aggregative review, configurative review, critical review, integrative review, literature review, mapping review, overview of the literature, scoping review, state-of-the-art review
Review process	Rapid review, systematized review, systematic review
Synthesis methods	Meta-analysis, meta-ethnography review, framework synthesis, mixed methods review, meta-narrative review, meta-synthesis, narrative review, network meta-analysis, realist review, thematic review
Topic (review questions)	Diagnostic test accuracy review, economic evaluation review, effectiveness review, etiology review, methodological review, prognostic review, psychometric review
Types of studies	Mixed studies review, quantitative review, qualitative review
Unit of analysis in the review	Umbrella review, review of reviews, meta-review, overview of reviews, systematic review of individual patient data

The first category provides a general idea of the overall purpose for conducting the review. For example, a 'mapping review' aims to map the literature. The second category is based on the review process. For example, a 'systematic review' usually includes several elements to ensure rigor in the review process (search, appraisal, synthesis, interpretation) whereas 'systematized review' attempts to include one or some of these elements (Grant & Booth, 2009). In the third category, the synthesis methods are used to designate the review. For example, 'framework synthesis' means that a review used a framework to synthesize the included studies and develop a new framework (Carroll, Booth, Leaviss, & Rick, 2013). Fourth, reviews are named after the topic that is studied (or review questions addressed) (Munn et al., 2018). For example, an 'effectiveness review' will focus on papers related to the effectiveness of a treatment or program, and 'diagnostic test accuracy review' will assess how well a diagnostic test performs in diagnosing and detecting a particular disease. Fifth, reviews are named based on the types of studies included. For example, 'quantitative review' means that the review included

quantitative studies. A last category is linked with the unit of analysis in a review. In general, reviews will be interested in primary studies. However, some researchers have also been interested in combining raw data from included studies (individual patient data) (Stewart & Tierney, 2002). Also, with the multiplication of reviews published, there are more and more researchers interested in combining the results of several reviews (McKenzie & Brennan, 2017).

Besides from SMSR, different terms have been used to designate this type of review such as integrative review (Whittemore & Knafl, 2005), mixed approach to synthesis (Pope, Mays, & Popay, 2007), mixed research synthesis (Sandelowski, Voils, & Barroso, 2006), mixed methods research synthesis (Heyvaert et al., 2013b), mixed methods systematic review (Pearson et al., 2014), and systematic review of qualitative and quantitative evidence. The terms 'systematic mixed studies review' (SMSR) was first introduced by Pluye et al. (2009). It is based on the types of studies included in a review process. In the literature on mixed methods research, the term 'mixed' designate combining qualitative and quantitative methods and is a subset of multimethod research that is associated with any combination of methods (Hunter & Brewer, 2015). Thus, in 'systematic mixed studies review', the term 'mixed' denotes including qualitative and qualitative 'studies' in a 'review'. This terminology is limited to the types of studies included and does not advocate any synthesis methods that should be used. In contrast, the terms 'mixed methods review' could mean using quantitative and qualitative (mixed) synthesis methods in a review. Integrative review aims to integrate the results of research, methods or theories (Whittemore & Knafl, 2005).

2.2 Conceptual Framework of the Critical Appraisal Process in SMSRs

In the following paper, a conceptual framework of the critical appraisal process in SMSRs will be presented. This framework was developed based on a literature review on critical appraisal and systematic reviews. It provides a definition of SMSR and three dimensions of quality. The purposes for performing critical appraisal are also described. This paper is published in the Journal of Mixed Methods Research (Hong & Pluye, 2018).

PAPER #1: A Conceptual Framework for Critical Appraisal in Systematic Mixed Studies Reviews

Published in the Journal of Mixed Methods Research.

(DOI: 10.1177/1558689818770058)

Quan Nha Hong[1], Pierre Pluye[1]

[1]Department of Family Medicine, McGill University, Montréal, Canada

Corresponding Author:

Quan Nha Hong, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, Canada, H3S 1Z1; Tel: 1-514-398-8483, Fax: 1-514-398-4202.

Email address: quan.nha.hong@mail.mcgill.ca

ABSTRACT

The past decade has been rich with methodological advancements in systematic reviews, several of which were inspired by the literature on mixed methods research. Systematic mixed studies reviews, i.e., reviews combining qualitative and quantitative evidence, are increasingly popular as they can provide a better understanding of complex phenomena and interventions. However, they raise new challenges, especially regarding how to perform critical appraisal of the included studies that vary regarding the methodologies used. To address this challenge, conceptually clarifying critical appraisal is necessary. To this end, this paper provides a framework for critical appraisal in systematic mixed studies reviews. This framework is an essential first step toward providing clear guidance on how to perform critical appraisal.

Keywords: critical appraisal, mixed methods research, quality assessment, systematic mixed studies review, systematic review

INTRODUCTION

Systematic reviews are considered among the best sources of research evidence, are used for decision-making, and are helpful for coping with the rapidly increasing volume of scientific literature (Bunn et al., 2015; Moynihan, 2004). There has been a call to broaden the scope of systematic reviews and integrate evidence from studies with diverse designs, especially to address the complexity of interventions, implementation, and context (Anderson et al., 2013; Pluye, Hong, Bush, & Vedel, 2016). Systematic reviews that include qualitative, quantitative and mixed methods studies (hereafter, *systematic mixed studies review* [SMSR]) respond to this need (Heyvaert, Hannes, & Onghena, 2016; Pluye & Hong, 2014).

Because of the heterogeneity in the designs of included studies, SMSRs raise several new challenges related to the syntheses of qualitative, quantitative and mixed methods studies and their integration, and the critical appraisal of the quality of included studies (Gough, 2015; Harden & Thomas, 2005). Previous work on SMSRs has focused on understanding how quantitative and qualitative evidence could be synthesized and integrated (Frantzen & Fetters, 2015; Heyvaert, Maes, & Onghena, 2013b; Hong, Pluye, Bujold, & Wassef, 2017; Sandelowski, Voils, Leeman, & Crandell, 2012). However, few papers have addressed the challenges of critical appraisal when the included studies have different designs. This paper focuses on this challenge.

Critical appraisal, the systematic and careful examination of study quality, is an important step in systematic reviews (Burls, 2009; Harden & Gough, 2012). Currently, there are over 500 critical appraisal tools for various study designs (Bai, Shukla, Bak, & Wells, 2012; Deeks et al., 2003; Katrak, Bialocerkowski, Massy-Westropp, Kumar, & Grimmer, 2004; West et al., 2002), but there is no clear guidance regarding which tool and approach to use, nor how or why to use them. This is may be due to a lack of conceptual clarity of what 'critical appraisal' means and what is appraised. This paper addresses this knowledge gap by providing a conceptual framework to better understand the components of critical appraisal in SMSRs. We have organized this paper into three main parts. The first provides a definition of SMSRs. The second presents a framework illustrating the different components involved in the critical appraisal process as well as some challenges and debates encountered in SMSRs. The third addresses the implications of the framework and suggests avenues for future research.

SYSTEMATIC MIXED STUDIES REVIEWS

Definition

SMSR follows the principles of mixed methods research (Heyvaert et al., 2016; Pluye & Hong, 2014). In primary research, mixed methods research is often defined based on its core characteristics; that is, the combination of elements of qualitative and quantitative research approaches, namely research question, research design, data collection, data analysis, and results (Creswell & Plano Clark, 2011; Johnson, Onwuegbuzie, & Turner, 2007). Based on these core components, we propose the following definition: Mixed methods research is a research approach in which a researcher or team of researchers integrates (a) qualitative and quantitative research questions, (b) qualitative and quantitative research designs and methods, (c) techniques for collecting and analyzing qualitative and quantitative data, and (d) qualitative findings and quantitative results (Pluye & Hong, 2014).

Applied to secondary research (i.e., literature reviews), the same components of this mixed methods definition can be found with slight differences in the terminology: (a) qualitative and/or quantitative **review** questions, (b) qualitative and/or quantitative **synthesis designs**, and (c) techniques for **extracting** and **synthesizing** qualitative and quantitative data, and (d) qualitative findings and quantitative results **of the synthesis**. SMSR has been defined as a systematic literature review conducted by a team of researchers that includes qualitative, quantitative, and/or mixed methods studies, and uses qualitative and/or quantitative synthesis methods (Heyvaert et al., 2016; Pluye & Hong, 2014). The term 'systematic' means that the review uses an explicit, transparent, and reproducible process with (a) specific review question(s) and precise study eligibility criteria; (b) a comprehensive set of information sources, and an exhaustive search strategy designed with specialized librarians; (c) a reliable or dependable (performed by at least two researchers) selection of relevant studies, data extraction, and critical appraisal; and (d) a rigorous synthesis (Pluye et al., 2016).

Two Levels of Integration

Integration can occur at two levels in SMSRs (Figure 1) (Heyvaert et al., 2013b). The first possible level of integration occurs during the selection of studies. A SMSR focuses on

synthesizing quantitative and qualitative evidence and includes any combination of qualitative, quantitative, and/or mixed methods studies (see check marks in Figure 1).

The second possible level of integration occurs during the synthesis, i.e., when the extracted data from the included studies are brought together using synthesis methods (Mays, Pope, & Popay, 2005). In SMSRs, there are multiple synthesis method options (Hong et al., 2017). As illustrated in Figure 1, the synthesis methods in SMSRs can be qualitative, quantitative, or mixed.

Insert Figure 1 about here

SMSRs using qualitative synthesis methods will provide a summary or interpretation of data to generate outputs such as themes, concepts, or theories. Several qualitative synthesis methods have been developed such as thematic synthesis, framework synthesis, meta-narrative synthesis, meta-ethnography, and critical interpretive synthesis (Barnett-Page & Thomas, 2009). Markoulakis and Kirsh (2013) provide an example of a SMSR using qualitative synthesis. They used critical interpretive synthesis (i.e., reciprocal translational analysis, lines of argument synthesis, and refutational synthesis) to develop a theory of difficulties faced by students with mental health issues in the university setting.

SMSRs using quantitative synthesis methods will provide numerical data and summaries of variables of interest of included studies. Basic and advanced meta-analysis methods (e.g., meta-regression and Bayesian synthesis) (Sutton & Higgins, 2008) are well-known examples. Roberts, Dixon-Woods, Fitzpatrick, Abrams, and Jones (2002) provide an illustration of a Bayesian synthesis used in a SMSR of factors affecting uptake of childhood immunization. In this review, to establish prior probabilities, the authors transformed the data from the included qualitative studies into quantitative data using quantitative content analysis. Then, these prior probabilities were combined with the results of the included quantitative studies to calculate probabilities that factors might affect immunization uptake.

The synthesis is considered mixed in SMSRs when both quantitative and qualitative synthesis methods are used. For example, Thomas et al. (2004b) conducted a review on the consumption of fruits and vegetables intake among children in which they performed a meta-analysis of controlled trials of the effectiveness of interventions and a thematic synthesis of

studies about children's views. Then, the findings of both syntheses were juxtaposed in a matrix to identify interventions that matched the children's views and to further explore if these interventions were more effective.

CRITICAL APPRAISAL

Critical appraisal is usually performed in systematic reviews to identify the strengths and weaknesses of studies, to determine how much confidence to have in the findings, and to ensure that the recommendations and conclusions properly reflect the quality of evidence reviewed, using sensitivity analysis, for instance; i.e., the comparison of results of lower vs. higher quality studies (Booth, Papaioannou, & Sutton, 2012). Different terms have been used to designate this construct, such as quality appraisal, quality assessment, validity assessment, and assessment of risk of bias (Higgins & Green, 2008). Hereafter, we will use *critical appraisal* to encompass all of these terms. To better understand critical appraisal, we looked at how this construct has been defined in literature on systematic reviews and how the critical appraisal process was performed in a sample of 459 SMSRs selected in a review of SMSRs (Hong et al., 2017). We compared the different definitions to highlight the commonalities and differences, and to identify the main components. We synthesized our findings into a conceptual framework. That is, we generated a representation of the interrelated constructs that provide a comprehensive understanding of a phenomenon (Jabareen, 2009).

Figure 2 presents the conceptual framework illustrating the process of critical appraisal in SMSRs including three main components: studies, papers and review. Based on this framework, critical appraisal in SMSRs can be defined as: a process related to judging the quality of qualitative, quantitative and mixed methods studies reported in research papers. In this process, three main dimensions of quality can be appraised: methodological, conceptual, and reporting. The purposes and choice of dimensions of quality to judge will vary depending on the objectives and synthesis method(s) adopted in a given review.

Insert Figure 2 about here

The Quality of Quantitative, Qualitative and Mixed Methods Studies

The first component in the framework is *studies*, represented by spheres (Figure 2). Since the unit of synthesis in SMSRs is studies, the judgment made concerns two dimensions of quality: methodological (trustworthiness) and conceptual (insightfulness). Depending on the research designs of the included studies and the review objectives, the criteria used to appraise the methodological and/or conceptual quality will vary.

Methodological quality. Methodological quality is concerned with how a study is conducted. It is usually related to the construct of trustworthiness: Is a study good enough for the results to be trustworthy? The judgment made about the trustworthiness of a study is typically related to the methodology and methods used and how biases were minimized (Higgins & Green, 2008; West et al., 2002).

There are two main approaches to appraising methodological quality of studies. In the first approach, studies are ranked based on their designs, with the assumption that some designs produce more credible inferences than others (Wells & Littell, 2009). In this approach, the methodological quality is conceived of as excellence. That is, quality studies meet the highest methodological standards that can yield results closer to the most plausible value. This approach is named the hierarchy of evidence or design hierarchy approach, in which systematic reviews with meta-analysis and randomized controlled trials are considered the best source of evidence (Wells & Littell, 2009). This approach is problematic in SMSRs since qualitative studies are excluded from the hierarchy of designs (Dixon-Woods et al., 2006).

A second approach used to appraise methodological quality of studies is associated with the absence of threats to validity (or risk of bias); the fewer threats or risks, the more trustworthy the results of the study. This threats-to-validity approach differs from the previous one by considering the specific features of a study design rather than contrasting these features with gold standards (Wells & Littell, 2009). One challenge when using this approach in SMSRs concerns the dimensions of trustworthiness that should be appraised. Table 1 presents different dimensions of trustworthiness that can be considered. In several critical appraisal tools, methodological quality refers to the internal validity of a study (Bai et al., 2012). For example, the Cochrane Risk of Bias Tool (Higgins et al., 2011) and the Newcastle-Ottawa Scale (Wells et al., 2000) include criteria that focus on how well a study was done to minimize bias. However, some tools

suggest appraising other types of validity such as external validity (Dyrvig, Kidholm, Gerke, & Vondeling, 2014). There are still diverging views on whether the appraisal should be limited to one or several types of validity and which types are the most important to appraise in SMSRs.

Insert Table 1 about here

Another challenge in SMSRs is evaluating and comparing the quality of studies from different epistemological and methodological traditions. As presented in Table 1, the dimensions of trustworthiness differ for quantitative and qualitative research. For mixed methods studies, the qualitative and quantitative components are combined to produce an integration that is greater than the sum of each component (Fetters & Freshwater, 2015). This might preclude the use of a single critical appraisal instrument for all included studies in SMSRs. There remains a lack of consensus on how critical appraisal should be performed and what criteria should be used, especially for qualitative and mixed methods studies (Carroll & Booth, 2015). Reviews on the quality in mixed methods have identified up to 13 different checklists for appraising mixed methods studies (Heyvaert, Hannes, Maes, & Onghena, 2013a) and 19 quality criteria (Fàbregues & Molina-Azorín, 2017). From our review on SMSRs, we identified four main approaches that were used for appraising the quality of mixed methods studies. One approach is to use specific criteria for the quantitative and qualitative components of the studies. To exemplify, several SMSRs used different tools such as the CASP tool for qualitative studies (Critical Appraisal Skills Programme (CASP), 2017) and the Effective Public Health Practice Project (EPHPP) tool for quantitative studies (Thomas, Ciliska, Dobbins, & Micucci, 2004a). A second approach is to use generic criteria that could be applied to all studies, such as the assessment form suggested by Hawker, Payne, Kerr, Hardey, and Powell (2002). A third approach consists of using specific criteria for mixed methods studies. For example, some SMSRs used the Mixed Methods Appraisal Tool (MMAT) that includes qualitative, quantitative, and mixed methods criteria (Pluye et al., 2011). A final approach is to appraise only the dominant component (qualitative or quantitative) of a mixed methods study.

Conceptual quality. Conceptual quality is defined as how clearly a concept is articulated to facilitate theoretical insight (Toye et al., 2013). This dimension of quality is related to insightfulness: Does the study provide a clear, rich, and deep understanding of a phenomenon? This dimension has been explored in a study on the process of critical appraisal of qualitative

studies (Toye et al., 2013). The authors found that conceptual clarity was an important dimension of quality used by reviewers to determine the inclusion of qualitative studies in a review using meta-ethnography. This dimension is linked with clarity but also with depth of description providing rich insight into a concept (Toye et al., 2013). Some reviews have also used this dimension of quality to appraise quantitative studies. For example, Beauregard, Marchand, and Blanc (2011) were interested in clarifying a construct in their review and appraised the conceptual quality of observational longitudinal studies using two criteria (i.e., analytical breadth and depth).

Conceptual quality is usually mentioned in systematic reviews interested in generating new understanding of a phenomenon. In these reviews, authors argue that too much emphasis on methodological rigor can limit the insight that could be gained from included studies (Sandelowski, 2000). Campbell et al. (2011) observed an inverse correlation between methodological and conceptual quality (i.e., papers providing good conceptual insight are generally of low methodological quality) that they explained, in part, due to the inadequate reporting of qualitative research methods. They suggest limiting methodological quality appraisal to a few screening criteria that allow identifying and excluding fatally flawed papers and focusing on papers that are conceptually useful for the synthesis (Campbell et al., 2011; Dixon-Woods et al., 2006).

The Quality of Research Papers

The second component of the framework is *research papers*, represented by boxes (Figure 2). Research papers can take several forms such as a journal article, dissertation, or report. The quality of how a research paper reports a study (reporting quality) varies widely depending on the authors and the structure of each form (e.g., journal articles are more concise than dissertations). This influences reviewers' judgment of the methodological and conceptual quality of a study.

Reporting quality. Reporting quality is related to the extent to which a paper "provides information about the design, conduct, and analysis of a study" (Huwiler-Müntener, Jüni, Junker, & Egger, 2002, p. 2801). This quality dimension is linked with the constructs of transparency,

accuracy, and completeness (Simera et al., 2010). These constructs can be defined as the extent to which a paper provides clear, detailed, and easy to understand information about a study (transparency), provides correct and true information (accuracy), and includes sufficient information (completeness) to allow readers to understand a study (Hornby, 2000; Simera et al., 2010). Over the past decade, more than 90 guidelines have been developed to provide standards for reporting research (Simera et al., 2010). These guidelines focus on issues that might introduce bias into a given study, and thus need to be reported such that readers can judge the quality of that study (Simera et al., 2010). In our review of SMSRs, several reporting guidelines have been used for critical appraisal such as the CONSORT statement for randomized controlled trials (Moher et al., 2010), the STROBE statement for observational studies (von Elm et al., 2007), the COREQ for qualitative studies (Tong, Sainsbury, & Craig, 2007), and TREND statement for non-randomized designs (Armstrong et al., 2008).

Opposing views regarding the use of reporting quality in systematic reviews remain. On the one hand, some are against the use of reporting quality in systematic reviews, especially when used as a surrogate for appraising methodological quality (Higgins & Green, 2008; Wells & Littell, 2009). It was found that using reporting quality as a proxy measure for methodological quality could lead to the misinterpretation of study quality (Huwiler-Müntener et al., 2002). Thus, the results, recommendations and conclusions of a review should be consistent with what was appraised.

On the other hand, reviewers have argued that reporting quality and methodological quality are related since a poorly reported paper will hinder the proper assessment of the trustworthiness of a study (Carroll, Booth, & Lloyd-Jones, 2012). Reporting quality criteria are said to be easier to judge and less prone to subjectivity (Carroll et al., 2012). Carroll et al. (2012) tested the effect of excluding papers solely based on the adequacy of their reporting using four criteria (pertaining to information provided on the question and study design, selection of participants, methods of data collection, and methods of analysis). They found that excluding inadequately reported papers had no meaningful impact on the results of a review. They suggest appraising reporting quality in a first step to exclude inadequately reported papers and then appraising the methodological quality of the remaining studies.

The Purposes of Critical Appraisal in SMSRs

A third component of the framework is the *review process*, represented by a funnel (Figure 2). In the review process, studies/research papers are identified through databases and other sources, selected using clear eligibility criteria, appraised, and synthesized. In our review of SMSRs, several reasons were provided for performing critical appraisal such as to describe the quality of the papers retained, to exclude papers of low quality or fatally flawed, to do a sensitivity analysis, to guide and strengthen the interpretation of study findings, and to explain differences in study results. These results can be grouped into three main purposes for performing critical appraisal in SMSRs.

The first purpose is for the selection of papers. A threshold approach has been suggested in which only studies meeting a predefined cut-off value of quality are retained (Gough, Thomas, & Oliver, 2012). Other approaches focus on conceptual quality in order to judge the utility, relevance, worth or value of each study, and include only studies meeting minimum criteria of scientific rigor (Pawson, Greenhalgh, Harvey, & Walshe, 2005).

The second purpose is descriptive. That is, the results of the critical appraisal are used to describe the quality of the studies included in a review. This can contribute to understanding how much readers may trust the results, identifying knowledge gaps, and making recommendations for future research (Booth et al., 2012). For this purpose, the appraisal focuses mainly on methodological quality.

The third purpose is related to the synthesis and interpretation of papers. Different strategies have been suggested as alternatives to excluding low methodological quality papers. One strategy is to use a weighting approach in which less weight is given to papers of low quality during the synthesis and interpretation of results (Gough et al., 2012). Another suggested strategy is to perform a sensitivity analysis based on the results of the critical appraisal. Sensitivity analysis consists of repeating an analysis by removing the studies that failed to achieve a pre-defined quality threshold (Carroll & Booth, 2015). If results differ based on the quality of studies, the conclusions of the review should be nuanced, with more importance placed on the results from higher-quality studies.

In summary, in our analysis of the literature on critical appraisal and SMSRs, we identified three main dimensions of quality (methodological, conceptual, and reporting) that are summarized in Table 2. These dimensions are related to different components: studies and research papers. Although these two components are closely linked, we found it necessary to present them separately since they address different dimensions of quality. That is, methodological and conceptual qualities are associated with studies, whereas, reporting quality is related to research papers. Distinguishing these dimensions in a review process is important since it will influence the review results and recommendations.

Insert Table 2 about here

Different Quality Dimensions Used Based on the Objectives of SMSRs

A variety of synthesis methods have been developed for SMSRs to address different objectives (Tricco et al., 2016). We used the components of our framework to understand the differences in critical appraisal approaches used in various SMSRs. The following presents three main review objectives and the dimension of quality addressed for each (Table 3).

One objective of SMSRs can be to test hypotheses by using aggregative synthesis approach such as meta-analysis. The questions may concern, for example, understanding the magnitude of a problem, testing the effectiveness of an intervention, or highlighting the association between factors. One or several predetermined critical appraisal tools are generally used to estimate bias that could lead to drawing misleading conclusions (Gough et al., 2012). Thus, the appraisal will mainly focus on the methodological quality of studies.

A second objective consists of interpreting and arranging the results of studies to generate new ways of understanding a subject and articulate new concepts or theories. To achieve this objective, configurative synthesis (or interpretive) approaches are usually used (Gough et al., 2012). In this type of review, there is no consensus regarding how the critical appraisal should be performed. Appraisal processes range from using the tools employed in the aggregative synthesis approach to focusing on study relevance and contribution to generating new understanding, rather than the methodological quality (Barnett-Page & Thomas, 2009). For example, in critical interpretive synthesis, Dixon-Woods et al. (2006, p. 4) propose excluding papers that are deemed

fatally flawed according to five quality criteria that cover reporting (e.g., are the aims and objectives of the research clearly stated?) and methodological (e.g., is the method of analysis appropriate and adequately explicated?) quality. On the other hand, in meta-narrative synthesis, Greenhalgh et al. (2005) suggest appraising the validity and relevance of primary studies using criteria within their respective research traditions.

A third objective is found in realist synthesis that is interested in understanding narrative causation using middle range theories (Jagosh et al., 2014). This synthesis approach seeks to explore and contextualize a complex intervention in multiple social settings and to answer the following question: "What is it about this kind of intervention that works, for whom, in what circumstances, in what respects and why?" (Pawson et al., 2005, p. 25). This synthesis approach involves an ongoing iterative interpretive process that uses abductive reasoning; i.e., hunches about conditions and outcomes can be incorporated in the synthesis (Jagosh et al., 2014). The appraisal in realist synthesis is more interested in the merit of each paper for the purpose of identifying/testing the middle range theory. Papers are appraised based on the minimum criteria of relevance (i.e., whether the study contributes to theory building and/or testing) and rigor (i.e., whether the method used is credible) (Pawson et al., 2005).

Insert Table 3 about here

DISCUSSION

The critical appraisal process in SMSRs is complex due to the heterogeneity of studies designs included. We found the literature on critical appraisal to be disparate, lacking consensus, and subject to multiple debates. There are various definitions of research quality but no agreement regarding what quality is. Also, it is not always clear why critical appraisal is performed, nor is consensus on how to perform it. There exists a wide variety of critical appraisal tools and approaches as well as debate regarding the appropriate expertise required for appraising studies.

To help reviewers deal with this complexity, this paper provides a conceptual framework of critical appraisal in SMSRs in which three components and dimensions of quality are

described (Figure 2). The three dimensions of quality (methodological, conceptual, and reporting) are intertwined in the critical appraisal process. For example, inadequate reporting will preclude a proper appraisal of the methodological and conceptual qualities of studies, selective reporting can be a source of methodological bias, and clear and concise definitions of constructs are fundamental for empirical testing (Higgins & Green, 2008; Simera et al., 2010; Suddaby, 2010). This suggests that multidimensional approaches to critical appraisal could be considered when performing a SMSR. How such approaches can be used needs to be explored further.

The critical appraisal process in SMSRs can be illustrated by the analogy of a courtroom trial where the three components described in Figure 2 can be found: evidence (studies), lawyers (papers), and judge and jury (review). First, in a courtroom trial, evidence will come from various sources such as witnesses and experts. Several questions can be posed: Are the witnesses relevant to the case? Are they credible? Are they making truthful claims? Analogously, questions posed during critical appraisal of studies in a review can be likened to those listed above: Is this study relevant for the review? Are adequate methods used in a rigorous manner? Are the results of the study trustworthy? Second, in a courtroom trial, lawyers are responsible for conveying pertinent evidence of what happened, and convincing the judge and jury of their case. Similarly, studies generally become accessible to the reviewers when they are published. Researchers 'package' their work, communicating it in a way that will convince reviewers it is worthy of being published and also convince others to read and cite it. Third, once the jury and judge have heard all the evidence, they will need to reach a decision. They might have different questions: Which if the diverging accounts presented is true? Should the evidence provided by less credible witnesses be excluded or weighted? Similarly, in a review, when different studies present contradictory results, how can we explain the differences? Which studies are credible and valid? What recommendations should be made based on all the evidence gathered? This analogy illustrates the intermediate position of lawyers (research papers) to convey the evidence (studies) to the judge and jury (review). The way the evidence is 'packaged' can greatly influence the judgment made.

Several future research avenues may be pursued on the critical appraisal process in SMSRs. First, the framework needs to be validated with a group of experts to determine if other dimensions of quality are addressed in SMSRs, and to refine the dimensions. Second, there is a

need to explore the interdependencies between the methodological, conceptual, and reporting dimensions of quality and how they influence the appraisal. Third, there is much debate around appraising methodological quality. It is necessary to test which criteria (for qualitative, quantitative, and mixed methods studies) have significant impact on review recommendations and conclusions. Fourth, an analysis of how the available critical appraisal tools and approaches fit within this framework could be performed. This analysis could lead to proposing a typology of tools and approaches, which will provide guidance for reviewers in selecting the most appropriate one for their reviews. Finally, improving our understanding of how criteria differ among studies from different epistemological and methodological traditions is needed.

CONCLUSION

The lack of conceptual underpinnings of critical appraisal in SMSRs is a source of multiple debates and inconsistency in the terminology and approaches used. We focused on what critical appraisal is, why the definition of quality varies, and how the findings of critical appraisal can be used. Another important question needs to be addressed: How should the critical appraisal of quantitative, qualitative and mixed methods studies be performed? The proposed framework is an essential first step to help answer this question.

ACKNOWLEDGEMENTS

Quan Nha Hong, OT, MSc, PhD candidate, holds a Doctoral Fellowship Award from the Canadian Institutes of Health Research (CIHR). Pierre Pluye, MD, PhD, Full Professor, holds a Senior Investigator Award from the Quebec Health Research Funds (FRQS) and is the Director for Method Development at the Quebec SPOR-SUPPORT Unit, which is funded by the CIHR, the FRQS, and the Quebec Ministry of Health. The authors would like to thank Dr. Suzanne Rivard, Full Professor at HEC Montréal, and Dr. Paula Bush, Academic Associate at McGill University, for constructive feedback on previous versions of this manuscript. We are also grateful to reviewers and editors who provided helpful comments and suggestions on behalf of the journal.

REFERENCES

Anderson, L. M., Oliver, S. R., Michie, S., Rehfuess, E., Noyes, J., & Shemilt, I. (2013). Investigating complexity in systematic reviews of interventions by using a spectrum of methods. *Journal of Clinical Epidemiology, 66*(11), 1223-1229.

Armstrong, R., Waters, E., Moore, L., Riggs, E., Cuervo, L. G., & Lumbiganon, P. (2008). Improving the reporting of public health intervention research: Advancing TREND and CONSORT. *Journal of Public Health, 30*(1), 103-109.

Bai, A., Shukla, V. K., Bak, G., & Wells, G. (2012). *Quality assessment tools project report*. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health.

Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: A critical review. *BMC Medical Research Methodology, 9*(59), 1-11.

Beauregard, N., Marchand, A., & Blanc, M.-E. (2011). What do we know about the non-work determinants of workers' mental health? A systematic review of longitudinal studies. *BMC Public Health, 11*(439), 1-15.

Booth, A., Papaioannou, D., & Sutton, A. (2012). *Systematic approaches to a successful literature review*. London: SAGE Publications.

Bunn, F., Trivedi, D., Alderson, P., Hamilton, L., Martin, A., Pinkney, E., et al. (2015). The impact of Cochrane Reviews: A mixed-methods evaluation of outputs from Cochrane Review Groups supported by the National Institute for Health Research. *Health Technology Assessment, 19*(28), 1-100.

Burls, A. (2009). *What is critical appraisal?* (2nd ed.). Newmarket, UK: Hayward Medical Communications.

Campbell, R., Pound, P., Morgan, M., Daker-White, G., Britten, N., Pill, R., et al. (2011). Evaluating meta-ethnography: Systematic analysis and synthesis of qualitative research. *Health Technology Assessment, 15*(43), i-164.

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

Carroll, C., Booth, A., & Lloyd-Jones, M. (2012). Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. *Qualitative Health Research, C22*(10), 1425-1434.

Creswell, J. W., & Plano Clark, V. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Critical Appraisal Skills Programme (CASP). (2017). 10 Questions to help you make sense of qualitative research. Retrieved November 3, 2017, from http://docs.wixstatic.com/ugd/dded87_25658615020e427da194a325e7773d42.pdf.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27), i-186.

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., et al. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology, 6*(35), 1-13.

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.

Dyrvig, A.-K., Kidholm, K., Gerke, O., & Vondeling, H. (2014). Checklists for external validity: A systematic review. *Journal of Evaluation in Clinical Practice, 20*(6), 857-864.

Fàbregues, S., & Molina-Azorín, J. F. (2017). Addressing quality in mixed methods research: A review and recommendations for a future agenda. *Quality & Quantity, 51*(6), 2847-2863.

Fetters, M. D., & Freshwater, D. (2015). The 1+ 1= 3 integration challenge. *Journal of Mixed Methods Research, 9*(2), 115-117.

Frantzen, K. K., & Fetters, M. D. (2015). Meta-integration for synthesizing data in a systematic mixed studies review: Insights from research on autism spectrum disorder. *Quality & Quantity, 50*(5), 2251-2277.

Gough, D. (2015). Qualitative and mixed methods in systematic reviews. *Systematic Reviews, 4*(181), 1-3.

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews, 1*(28), 1-9.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science and Medicine, 61*(2), 417-430.

Harden, A., & Gough, D. (2012). Quality and relevance appraisal. In D. Gough, S. Oliver & J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 153-178). London: SAGE Publications.

Harden, A., & Thomas, J. (2005). Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methodology, 8*(3), 257-271.

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.

Heyvaert, M., Hannes, K., Maes, B., & Onghena, P. (2013a). Critical appraisal of mixed methods studies. *Journal of Mixed Methods Research, 7*(4), 302-327.

Heyvaert, M., Hannes, K., & Onghena, P. (2016). *Using mixed methods research synthesis for literature reviews: The mixed methods research synthesis approach*. Thousand Oaks, CA: SAGE Publications.

Heyvaert, M., Maes, B., & Onghena, P. (2013b). Mixed methods research synthesis: Definition, framework, and potential. *Quality & Quantity, 47*(2), 659-676.

Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley Online Library.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal, 343*(d5928).

Hong, Q. N., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews, 6*(61), 1-14.

Hornby, A. S. (2000). *Oxford advanced learner's dictionary of current English* (6th ed.). Oxford, UK: Oxford University Press.

Huwiler-Müntener, K., Jüni, P., Junker, C., & Egger, M. (2002). Quality of reporting of randomized trials as a measure of methodologic quality. *Journal of the American Medical Association, 287*(21), 2801-2804.

Jabareen, Y. (2009). Building a conceptual framework: Philosophy, definitions, and procedure. *International Journal of Qualitative Methods, 8*(4), 49-62.

Jagosh, J., Pluye, P., Wong, G., Cargo, M., Salsberg, J., Bush, P. L., et al. (2014). Critical reflections on realist review: Insights from customizing the methodology to the needs of participatory research assessment. *Research Synthesis Methods, 5*(2), 131-141.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

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

Markoulakis, R., & Kirsh, B. (2013). Difficulties for university students with mental health problems: A critical interpretive synthesis. *Review of Higher Education: Journal of the Association for the Study of Higher Education, 37*(1), 77-100.

Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research and Policy, 10*(Suppl 1), 6-20.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., et al. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology, 63*(8), e1-e37.

Moynihan, R. (2004). *Evaluating health services: A reporter covers the science of research synthesis*. New York: Milbank Memorial Fund.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review - A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy, 10* (Suppl 1), 21-34.

Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health, 35*, 29-45.

Pluye, P., Hong, Q. N., Bush, P. L., & Vedel, I. (2016). Opening-up the definition of systematic literature review: The plurality of worldviews, methodologies and methods for reviews and syntheses. *Journal of Clinical Epidemiology, 73*(5), 2-5.

Pluye, P., Robert, E., Cargo, M., Bartlett, G., O'Cathain, A., Griffiths, F., et al. (2011). *Proposal: A Mixed Methods Appraisal Tool for systematic mixed studies reviews*. Retrieved November 15, 2013, from http://mixedmethodsappraisaltoolpublic.pbworks.com.

Roberts, K. A., Dixon-Woods, M., Fitzpatrick, R., Abrams, K. R., & Jones, D. R. (2002). Factors affecting uptake of childhood immunisation: A Bayesian synthesis of qualitative and quantitative evidence. *Lancet, 360*(9345), 1596-1599.

Sandelowski, M. (2000). Focus on research methods - Whatever happened to qualitative description? *Research in Nursing and Health, 23*(4), 334-340.

Sandelowski, M., Voils, C. I., Leeman, J., & Crandell, J. L. (2012). Mapping the mixed methods–mixed research synthesis terrain. *Journal of Mixed Methods Research, 6*(4), 317-331.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. *BMC Medicine, 8*(24), 1-6.

Suddaby, R. (2010). Editor's comments: Construct clarity in theories of management and organization. *Academy of Management Review, 35*(3), 346-357.

Sutton, A. J., & Higgins, J. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*(5), 625-650.

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004a). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176-184.

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., et al. (2004b). Integrating qualitative research with trials in systematic reviews. *British Medical Journal, 328*(7446), 1010-1012.

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.

Toye, F., Seers, K., Allcock, N., Briggs, M., Carr, E., Andrews, J., et al. (2013). 'Trying to pin down jelly'- Exploring intuitive processes in quality assessment for meta-ethnography. *BMC Medical Research Methodology, 13*(46), 1-12.

Tricco, A. C., Antony, J., Soobiah, C., Kastner, M., MacDonald, H., Cogo, E., et al. (2016). Knowledge synthesis methods for integrating qualitative and quantitative data: A scoping review reveals poor operationalization of the methodological steps. *Journal of Clinical Epidemiology, 73*(5), 29-35.

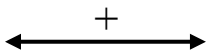
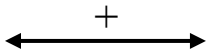
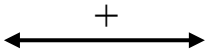
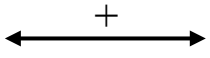
von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine, 4*(10), e296.

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., et al. (2000). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Retrieved April 16, 2016, from http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf.

Wells, K., & Littell, J. H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice, 19*(1), 52-62.

West, S. L., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., et al. (2002). *Systems to rate the strength of scientific evidence*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ).

Table 1. Dimensions of Trustworthiness and Comparison of Criteria in Quantitative, Mixed Methods, and Qualitative Research

Dimensions of trustworthiness	Type of research		
	Quantitative research	Mixed methods research	Qualitative research
Truthfulness	Internal validity		Credibility
Applicability	External validity		Transferability
Consistency	Reliability		Dependability
Neutrality	Objectivity		Confirmability

(adapted from Heyvaert et al., 2016)

Table 2. Three Dimensions of Quality in Critical Appraisal

Features	Quality dimension		
	Methodological	Conceptual	Reporting
Definition	Extent to which a study's design, conduct, and analysis have minimized selection, measurement, and confounding biases	Extent to which a concept is clearly articulated to facilitate theoretical insight	Extent to which a paper provides information about the design, conduct, and analysis of a study
Constructs	Trustworthiness	Insightfulness	Accuracy Completeness Transparency
Component	Study	Study	Research paper
Example of criteria*	Were the statistical tests used to assess the main outcomes appropriate?	Are there clear translatable concepts?	Is the hypothesis/aim/ objective of the study clearly described?

* Examples from: Downs and Black (1998) and Toye et al. (2013).

Table 3. Comparison of Critical Appraisal Based on the Objectives of Reviews

Objective of reviews	Example of synthesis methods	Type of reasoning	Purpose of the appraisal	Dimension of quality appraised
Test hypothesis	Meta-analysis	Deduction	Determine if studies are affected by significant bias	Methodological
Provide causal pathway or causal explanation	Realist synthesis	Abduction	Determine if studies are fit for purpose for theory development and/or testing	Conceptual and methodological
Develop conceptual understanding	Critical interpretive synthesis Meta-narrative synthesis	Induction	Determine the relevance, credibility, and contribution of studies	No consensus Conceptual and methodological

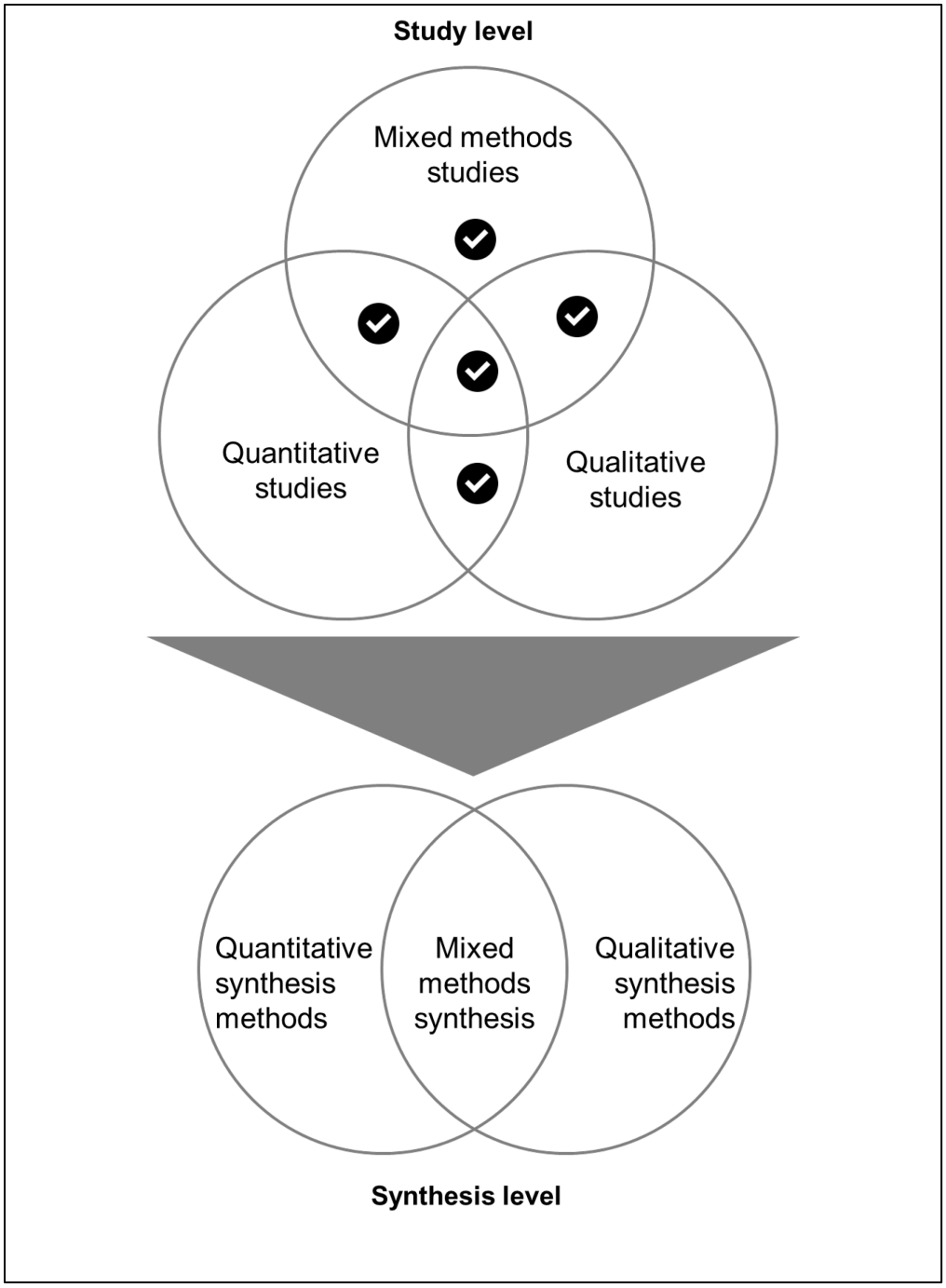


Figure 1. Integration of Studies and Integration of Synthesis Methods in Systematic Mixed Studies Reviews

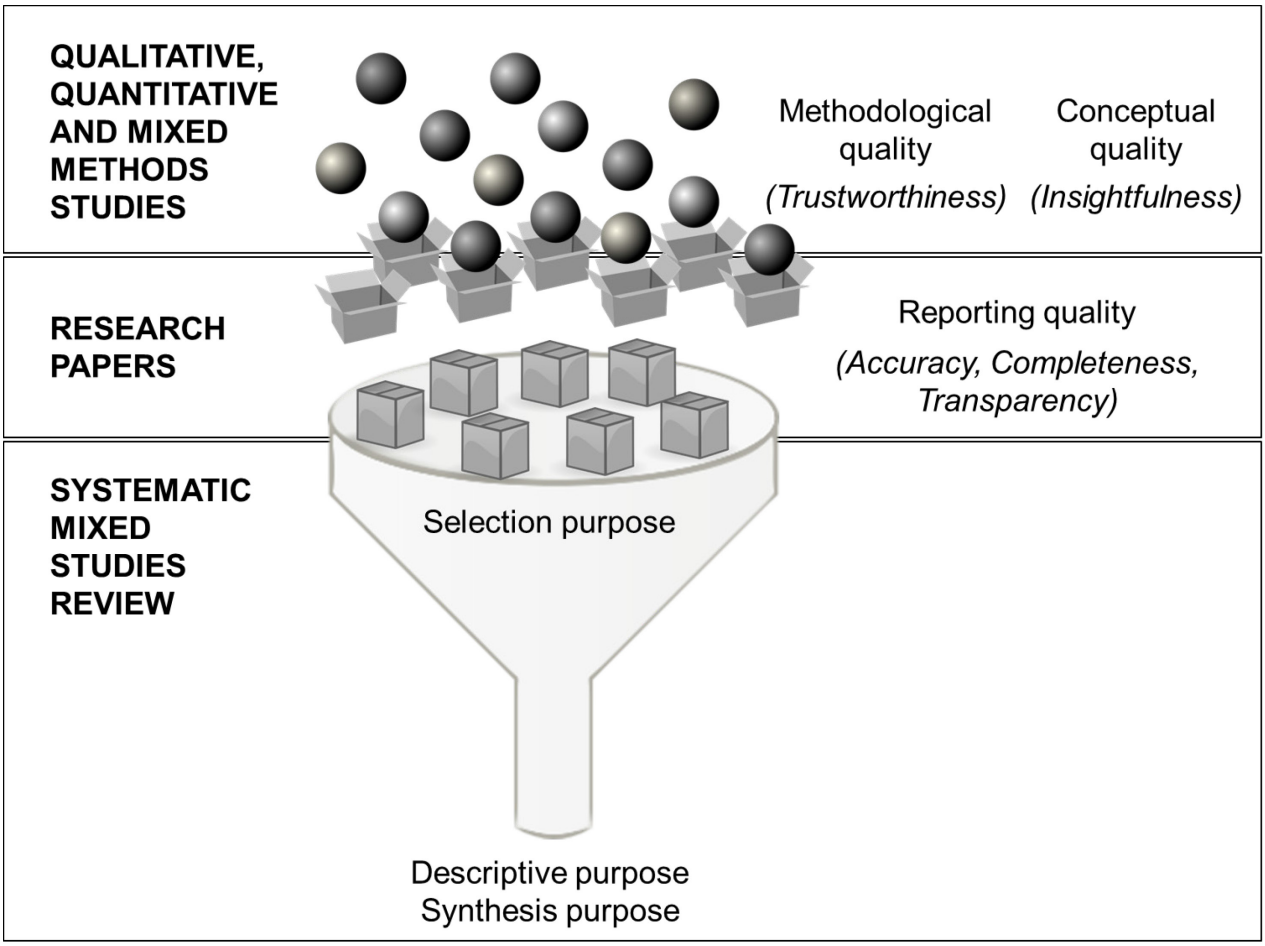


Figure 2. Framework of the Different Components Involved in the Critical Appraisal Process in Systematic Mixed Studies Reviews

CHAPTER 3. LITERATURE REVIEW

This chapter first presents the questions addressed in this literature review. Then, details on the methods used to identify, select, and synthesize the literature are provided. Finally, this chapter ends with the results found and a discussion of the main findings.

3.1 Review Questions

For this literature review, two review questions were asked:

1. How is critical appraisal performed in SMSRs?
2. What are the existing critical appraisal tools for assessing the methodological quality of primary studies that had been subject to validity and reliability testing?

3.2 Methods

3.2.1 Sources

To answer the first question, a literature review on SMSRs was carried out in 2015. Details on the methods used in this review is available in Hong et al. (2017) (Appendix 2). In summary, SMSRs were searched in six databases (MEDLINE, PsycINFO, Excerpta Medica database (Embase), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Allied and Complementary Medicine Database (AMED), and Web of Science) from inception of each database until December 8, 2014. The search strategy included free text keywords on reviews, syntheses, and mixing qualitative and quantitative studies, methods or data. The search strategy was drafted by the first author and checked by two specialized librarians.

To answer the second question, two sources were used: the literature review on SMSRs and published literature reviews on critical appraisal tools (CATs). Up to 2015, several reviews on CATs have been performed. These reviews were identified from citations tracking of CATs found in the review of SMSRs and sources known to the authors. The reviews that provided information on the measurement properties of the tools were used to identify the CATs that had been tested for validity and/or reliability. Forward citation tracking of the identified CATs was performed in Google Scholar to check if new papers were published after the date of the retained

reviews. The search and selection were performed by one reviewer.

3.2.2 Selection criteria

For the first review question, SMSRs published in English or French were retained if they included either (a) qualitative, quantitative and/or mixed methods studies; (b) qualitative and mixed methods studies; (c) quantitative and mixed methods studies or; (d) only mixed methods studies. More detail on the selection criteria of SMSRs is provided in Appendix 2. Two reviewers were independently involved in the screening of titles and abstracts as well as in the selection of full-text papers. All CATs mentioned in the included SMSRs were analyzed to have a better understanding of how critical appraisal was performed in these reviews.

For the second review question, the CATs analyzed in the retained reviews and used in SMSRs were listed in an Excel spreadsheet. A CAT could be a scale or checklist in which a list of criteria and domains are suggested to appraise the quality of a study. Different terms are used to designate these tools such as risk of bias tools, quality assessment instruments, validity assessment tools, and quality appraisal tools. Only the tools that were subject to validity and reliability testing were considered for inclusion. Also, CATs were retained if they included methodological quality. Tools limited to reporting quality criteria of studies such as those listed on the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network website (http://www.equator-network.org/) were excluded. For example, popular tools exist for reporting systematic reviews (e.g., Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement (Moher, Liberati, Tetzlaff, & Altman, 2009)), RCT (e.g., COnsolidated Standards Of Reporting Trials (CONSORT) statement (Moher et al., 2010), qualitative studies (e.g., COnsolidated criteria for REporting Qualitative research (COREQ) (Tong, Sainsbury, & Craig, 2007)), and observational studies (e.g., Strengthening The Reporting of OBservational studies in Epidemiology (STROBE) statement (von Elm et al., 2008)). Moreover, only CATs appraising primary research involving human subjects were retained. Thus, CATs were excluded if developed for the quality assessment of systematic reviews (e.g., Risk of Bias in Systematic Reviews (ROBIS) tool (Whiting et al., 2016)) or guidelines (e.g., Appraisal of Guidelines for Research & Evaluation (AGREE) instrument (Brouwers et al., 2010)). Also, tools for appraising animal studies or limited to external validity were excluded.

Table 3 summarizes the inclusion and exclusion criteria used.

Table 3. Eligibility Criteria of Critical Appraisal Tools

Criteria	Inclusion	Exclusion
Critical appraisal tool	• List of criteria to judge the quality of studies.	• Grading approaches
Quality dimension	• Tools including methodological quality criteria.	• Tools limited to reporting quality criteria.
Type of studies	• Tools for primary research involving human subjects.	• Tools not appraising primary research such as systematic reviews or guidelines. • Tools for animal studies. • Tools limited to external validity.
Validity (at least one of these criteria)	• The tool development included consultations with experts (e.g., Delphi study, survey). • The tool was compared with other existing tools or expert judgment. • The tool was pilot tested with experts/users, and results were used to refine the tool. • Factor analysis was performed. • Correlations with related or unrelated constructs were calculated.	• No information provided on validity.
Reliability (at least one of these criteria)	• Correlations between the items of the tool were performed (internal consistency). • Two reviewers or more appraised studies with the tool and the ratings were compared (interrater reliability). • Studies were rated twice by the same reviewers at an X time interval and the ratings were compared (test-retest reliability).	• No information provided on reliability.
Language	• English or French.	• Other languages than English or French.

3.2.3 Data extraction and synthesis

A descriptive synthesis was performed by one reviewer. For the first review question, the following data were extracted: number of qualitative, quantitative and mixed methods studies retained in the SMSRs, appraisal tool(s) used, number of reviewers involved in the critical appraisal process, presentation format of results of the appraisal, and purpose for performing critical appraisal.

For the second review question, the retained appraisal tools identified were classified based on the study design assessed (e.g., RCT, NRS, descriptive studies, qualitative, mixed methods studies). Also, information on the validity and reliability of each tool was extracted. Moreover, the following information of the characteristics of the CATs were collected: number of criteria, scale used, time to complete (when available), and user guide available.

3.3 Results

3.3.1 Critical appraisal in SMSRs

For the first review question, a total of 459 SMSRs were retained. The number of papers included in the SMSRs ranged from 2 to 295. Figure 2 presents the number of SMSRs for each combination of studies that can characterize SMSRs (as defined in Figure 1 in Paper #1, section 2.2).

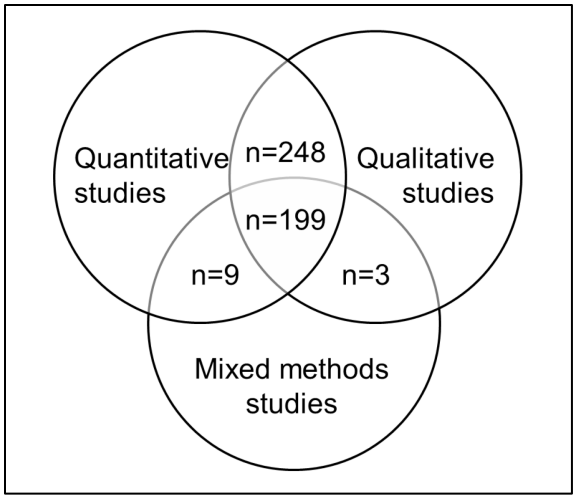


Figure 2. Number of Systematic Mixed Studies Reviews Including Qualitative, Quantitative and Mixed Methods Studies

3.3.1.1 Purposes for performing critical appraisal in SMSRs

Among the 459 SMSRs, 20 did not provide any results of the critical appraisal. The main purpose for performing critical appraisal was descriptive (n=378). The description of the results of the critical appraisal greatly varied from one sentence on the overall mean score of the included studies to a detailed section on the methodological limitations of studies in the results or discussion section. More than a third of the SMSRs presented the results of the critical appraisal in the table of characteristics of included studies (n=172). Also, nearly a third of SMSRs provided results of the critical appraisal in tables or figures (n=150) such as detailed ratings for each item of the CATs used. Nearly 60% of SMSRs (n=267) provided an overall score that could be numerical (e.g., percentage) or textual (e.g., high, moderate, low).

The second purpose for performing critical appraisal was for the exclusion of studies (n=65). In these SMSRs, the reviewers usually determined a priori a minimum threshold to include papers. Studies not meeting this threshold were considered of low/poor quality or with substantive flaws.

The third purpose was to influence the synthesis and interpretation of findings of the review (n=46). These SMSRs mentioned using the results of the critical appraisal to strengthen and guide the interpretation of study findings, to determine the robustness of the synthesis, to explain differences between study results, and to provide a level of evidence for each recommendation. Also, some SMSRs (n=21) performed sensitivity analysis to assess the impact of quality variation or weight the findings according to the results of higher quality studies.

3.3.1.2 Number of reviewers involved in critical appraisal in SMSRs

Among the 459 SMSRs, 352 provided information on the number of reviewers involved in critical appraisal. Most SMSRs had two or more reviewers performing independently the appraisal of studies (n=260) or a second reviewer independently appraising a random number of studies (n=23). In a smaller number of SMSRs, the critical appraisal was performed by only one reviewer (n=45), or by a second reviewer counterchecking the appraisal of a first reviewer (n=24)

3.3.1.3 Critical appraisal tools used in SMSRs

Among the 459 SMSRs, it was possible to collect information on the tools used of 424 reviews since 35 had missing data (e.g., only mentioned having conducted a critical appraisal without detailing how it was performed). In 152 SMSRs, the authors mentioned that they developed their own criteria or adapted criteria from several existing tools (bespoke tools). The sources on which they rely to develop their critical appraisal criteria were clearly stated in 102 SMSRs. Also, 53 SMSRs mentioned using evidence levelling approaches to judge the quality of the included studies such as those proposed by GRADE (Guyatt et al., 2011a), GRADE-CERQual (Lewin et al., 2015), NHMCR levels of evidence hierarchy (National Health and Medical Research Council, 2000), Joanna Briggs Institute (JBI) Levels of Evidence (Joanna Briggs Institute, 2017d), American Association of Critical-Care Nurses (AACN) levels of evidence (Armola et al., 2009), and Daly et al. (2007). Most of these SMSRs used an evidence levelling approach in complement to CATs. However, nine SMSRs did not use any CATs and only considered the level of evidence for each study.

A total of 124 CATs mentioned in 315 SMSRs were identified. Appendix 3 presents the tools that were identified and the number of SMSRs that used each tool for the appraisal of quantitative, qualitative, and mixed methods studies as well as those that used them for several types of studies. This list includes the tools that the authors mentioned using or adapting. The tools used in SMSRs appraised different dimensions such as reporting, relevance, or methods. The number of tools used in a SMSR ranged from one to six.

For quantitative studies, 65 tools addressing either one specific design or several designs were used. Most tools have been used in a very small number of reviews. The most often used tools for quantitative studies were the Cochrane Risk of Bias (RoB) tool for RCT (n=21), the Quality Assessment Tool for Quantitative Studies from the Effective Public Health Practice Project (EPHPP) (n=19), the STROBE statement for reporting observational studies (n=8), and the Newcastle Ottawa Scale for NRS (n=7), the Downs & Black for RCT and NRS (n=6), and the Jadad tool (n=6). Six SMSRs mentioned that they did not appraise the quality of other quantitative studies than RCT (especially surveys).

For qualitative studies, 36 tools were used. Compared to tools for quantitative studies, tools for qualitative studies have not been developed for specific designs. The most common

tools used were the Critical Appraisal Skills Programme (CASP) for qualitative research (n=50) followed by far by the quality appraisal checklist for qualitative studies from the National Institute for Health and Clinical Excellence (NICE) (n=7), the COREQ (n=6), and the JBI-Qualitative Assessment and Review Instrument (QARI) (n=5). Some reviews mentioned that they did not appraise the quality of qualitative studies due to the lack of valid tools and of consensus (n=9).

A total of 211 SMSRs included mixed methods studies. These studies were assessed using either only criteria from the most dominant component (only qualitative or quantitative criteria), both types (e.g., using the tools chosen for qualitative and quantitative studies and awarding the highest quality rating), or only using specific criteria for mixed methods studies. For the latter, only one tool was mentioned: the MMAT (n=7).

Among the SMSRs, 46 different tools were used for different designs. The most common tools found were the JBI-QARI and JBI-Meta-Analysis of Statistics Assessment and Review Instrument (MAStARI) (n=20), the CASP (n=19), and, and the McMaster Critical Review Form – Quantitative and Qualitative Studies (n=9). In addition to these tools, others have been specifically developed to assess the quality of diverse designs. The most popular ones were the MMAT (n=20), Quality assessment system (QualSyst) (n=17), Hawker's appraisal tool (n=16), Dixon-Woods's appraisal prompts (n=3), and Quality Assessment Tool for Studies with Diverse Designs (QATSDD) (n=2).

From the analysis of the different CATs and critical appraisal process, it is possible to identify four main categories of tools: (a) generic tools; (b) generic tools including specific criteria; (c) specific tools for categories of studies; and (d) specific tools for a study design. A first category is to use a generic set of criteria for the appraisal of all included studies. For example, the Hawker's appraisal tool includes nine criteria related to the abstract/title, introduction/aims, methods and data, sampling, data analysis, ethic and bias, findings/results, transferability/generalizability and implication/usefulness (Hawker, Payne, Kerr, Hardey, & Powell, 2002). When using generic tools in SMSRs, only one tool is necessary since the criteria can be applied to any study. The second category is to use generic criteria and add specific criteria for qualitative and quantitative studies. For example, this approach is used in the QATSDD which includes 14 generic criteria, two criteria specific to qualitative studies and two

criteria specific to quantitative studies (Sirriyeh, Lawton, Gardner, & Armitage, 2012). Also, some SMSRs adapted existing tools. For example, the Medical Education Research Study Quality Instrument (MERSQI) was initially developed for appraising quantitative studies (Reed et al., 2007). In two SMSRs, the authors mentioned that they also used this tool to appraise the quality of qualitative studies by omitting or adapting some criteria that were not applicable. The last two categories of tools suggest specific criteria. Some developed criteria for a category of studies such as the QualSyst that has 10 qualitative criteria and 10 quantitative criteria (Kmet, Lee, & Cook, 2004). Other tools focused on one specific design such as RCT for the Cochrane RoB Tool (Higgins et al., 2011). Figure 3 presents the number of CATs used in SMSRs for each category of tools.

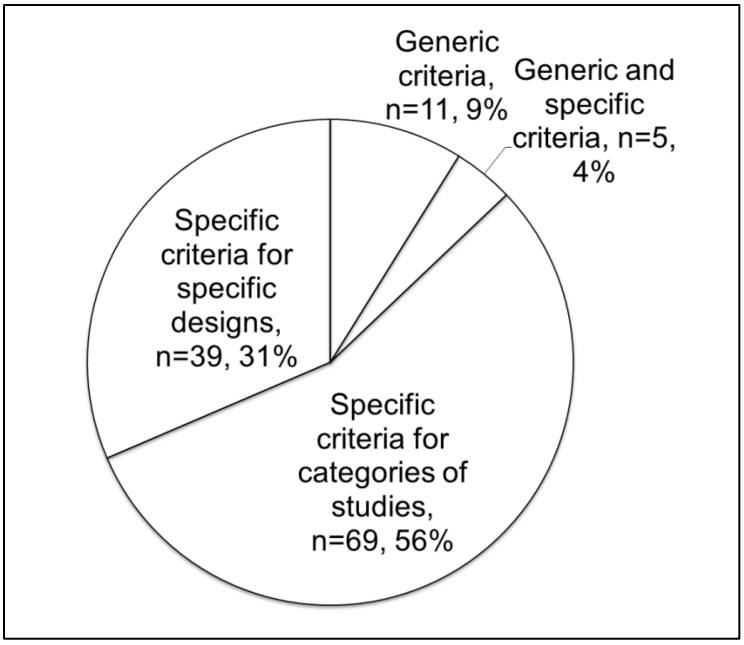


Figure 3. Distribution of the Critical Appraisal Tools Used in Systematic Mixed Studies Reviews (n=124) Among Four Categories of Tools

3.3.2 Critical appraisal tools with validity and reliability testing

For the second review question, a total of 17 reviews on CATs were identified in this literature review, of which 12 provided information on the measurement properties of the

included CATs. A detailed list of retained reviews is provided in Appendix 4. The reviews were published between 1995 and 2015. The number of CATs retained in the reviews ranged from 8 to 267. Most of the reviews searched for CATs in databases such as MEDLINE and PubMed. One was a review of systematic reviews of CATs. The majority of reviews retained were for CATs on quantitative studies (12 out of 17); one was specific to qualitative studies, one to mixed methods studies, and three included CATs for different types of study designs.

From the reviews on CATs and the review on SMSRs, a total of 508 CATs were identified of which only 52 CATs provided information on validity and reliability testing (Figure 4).

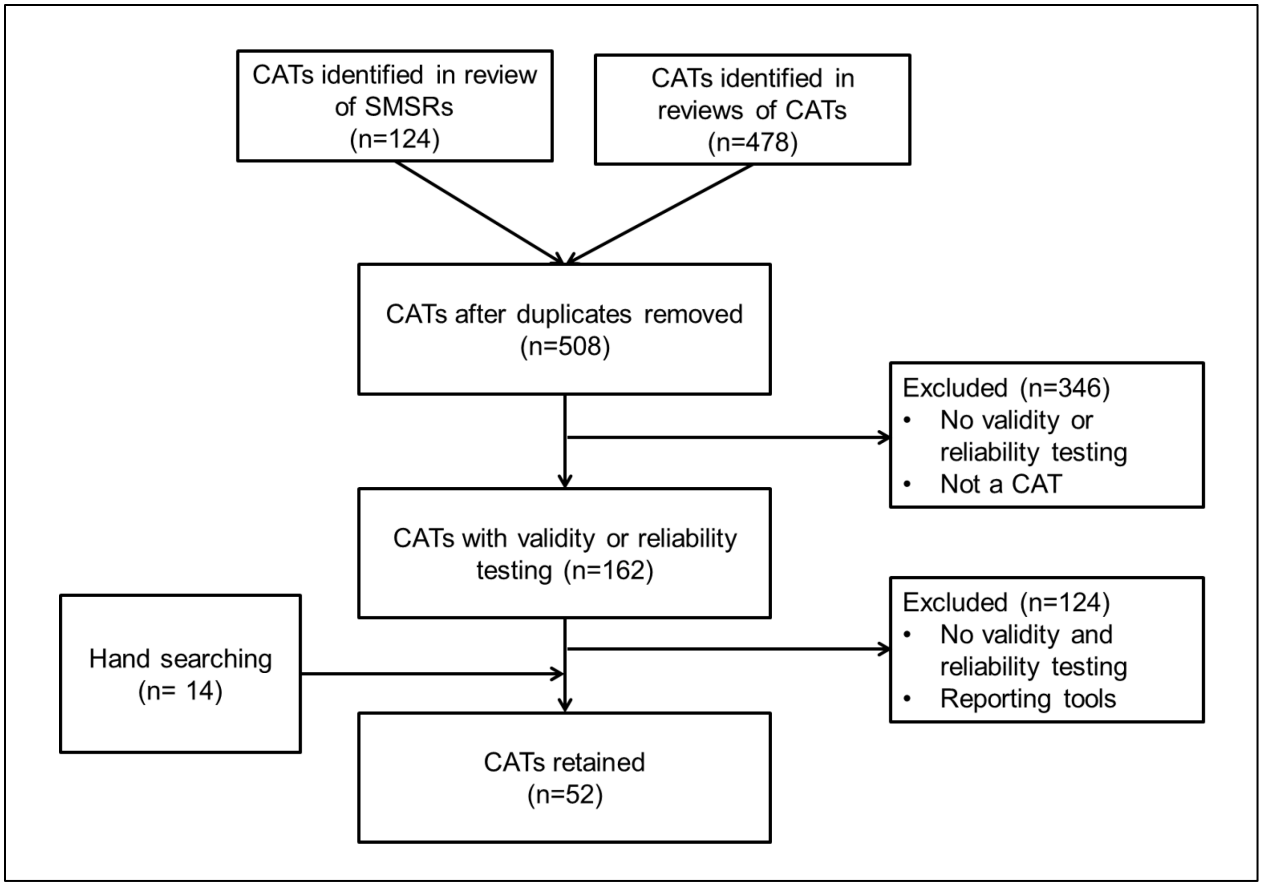


Figure 4. Flowchart of the Review on Critical Appraisal Tools (CATs)

A description of each retained CAT is provided in Appendix 5. These 52 CATs were developed between 1972 and 2016. The large majority of the identified CATs were developed for a specific study design (n=27) or for a category of studies (n=20). Several CATs were for specific quantitative designs: RCT (n=14), cohort study (n=1), prognostic study (n=1), single case experiment design (n=1), prevalence study (n=2), and case series (n=2). Other tools for quantitative studies could be applied to several designs such as for observational studies or NRS (n=6), for RCT and NRS (n=5), for intervention studies (n=4), for quantitative studies in general (n=3), and for case-control and cohort studies (n=1). Only three CATs for qualitative studies were retained. Six tools can be used for several study designs but only one of these tools included criteria specific to mixed methods studies. Other tools were developed for specific types of studies such as studies on measurement properties of health instruments (n=1), studies on diagnostic reliability (n=1), and studies on diagnostic accuracy (n=1).

The 52 retained CATs were subject to interrater reliability testing. The content validity was clearly described in 45 CATs. It consisted either of consultations with experts (e.g., Delphi study, survey) and/or an extensive literature review with pilot testing. Criterion and construct validity were tested in respectively 25 and 20 CATs. Finally, a smaller number of tools investigate the test-retest reliability (n=8) or internal consistency (n=9).

3.4 Summary

In summary, this literature review analyzed 459 SMSRs and 17 reviews of CATs to describe how critical appraisal was performed in SMSRs and identify the existing CATs that were subject to validity and reliability testing. Four main approaches to critical appraisal in SMSRs and a very large variety of tools were found. This large variety and the small number of reviews using the same tools show that there is a lack of consensus on how to appraise studies included in SMSRs. Some researchers preferred using different tools chosen based on the designs of included studies, while others used only one tool that covers different designs.

In spite of the very large number of CATs that are used, very few tools were tested for validity and reliability. This finding is corroborated in several existing reviews on CATs (Crowe & Sheppard, 2011; Deeks et al., 2003; Jarde, Losilla, & Vives, 2012; Katrak et al., 2004; Saunders, Soomro, Buckingham, Jamtvedt, & Raina, 2003; Shamliyan, Kane, & Dickinson,

2010; Wendt & Miller, 2012; West et al., 2002). In addition, the hand searching was useful to identify 14 more recent CATs (published after 2010), which might indicate that a more rigorous process of tool development is being advocated and implemented (Whiting, Wolff, Mallett, Simera, & Savović, 2017).

The 52 identified CATs were mainly for quantitative studies, especially for experimental and observational studies. Some tools were found for quantitative descriptive studies that appraised prevalence studies and case series. However, no tool for survey research was found. Among the retained CATs, few were specific to qualitative studies. Historically, interest in reviewing quantitative studies, especially RCT, started in the 1960s and 1970s from the seminal works of Donald Campbell (Campbell & Stanley, 1963) and Archie Cochrane (Cochrane, 1972). Interest in qualitative reviews emerged more in the late 1980s with the development of the meta-ethnography synthesis method (Noblit & Hare, 1988). Yet, it is mainly in the years 2000s that greater interest and methodological development on systematic reviews of qualitative studies have been seen. For example, the Cochrane Collaboration published the first review of qualitative evidence in 2013 (Gulmezoglu, Chandler, Shepperd, & Pantoja, 2013). A more recent review published in 2016 on the quality of qualitative studies identified 133 guidelines and retained 58 that were further analyzed with a group of experts (Santiago-Delefosse et al., 2016). Based on the titles of these guidelines, several have been developed for providing advice for manuscript submissions. It is likely that several of these guidelines would have not met the eligibility criteria of this review since reporting tools were excluded as well as those that did not have validity or reliability testing.

Similarly to qualitative studies, few tools for mixed methods studies were found. Most SMSRs appraised mixed methods studies by using different criteria for qualitative and quantitative studies and some used generic criteria that can be applied to all study designs. However, besides from the MMAT, none included criteria specific to mixed methods studies. There is still no consensus regarding the critical appraisal of mixed methods studies, but research is developing rapidly (Bryman, Becker, & Sempik, 2008; Burrows, 2013; Fàbregues, Paré, & Meneses, 2018; Heyvaert, Hannes, Maes, & Onghena, 2013a; Long, Godfrey, Randall, Brettle, & Grant, 2002; O'Cathain, Murphy, & Nicholl, 2008; Sale & Brazil, 2004). A recent review analyzed 64 papers on the quality in mixed methods studies and provided recommendations on

the importance of empirical publications on quality, the necessity for greater consistency in the quality terminology, and the need to reach agreement on core quality criteria (Fàbregues & Molina-Azorín, 2017). Also, mixed methods studies is increasingly used and valued in health science (Ostlund, Kidd, Wengstrom, & Rowa-Dewar, 2011). For example, in the review of SMSRs, more than 45% of SMSRs had included at least one mixed methods study (Hong et al., 2017). This shows a clear need to pursue the development of research on critical appraisal of mixed methods studies.

The MMAT was used in 27 SMSRs retained in this review. The majority of SMSRs used the MMAT for appraising different types of studies. However, seven SMSRs used the MMAT only to appraise mixed methods studies and other CATs were used to appraise the qualitative and quantitative studies. It would be interesting to further investigate why these authors chose to use other tools. Different hypotheses can be put forward. For example, the name of the tool might be misleading; some might think that the 'Mixed Methods Appraisal Tool' is limited to mixed methods studies. Also, the authors might consider that the MMAT is not appropriate for appraising only qualitative or quantitative studies and that other CATs are more suitable for studies included in their reviews.

CHAPTER 4. METHODOLOGY

This chapter focuses on the study design and methodology used in this project. The first part addresses the overall study design and justifies the reasons for performing a two-phase mixed methods project. Then, the description and justification of the methodology used for each of the two phases of the project are presented. The last part is on ethical considerations.

4.1 Study Design: Sequential Exploratory Mixed Methods Design

This project consisted in a mixed methods research in which both qualitative and quantitative data collection and analysis were performed. Mixed methods research is defined as "a research approach in which a researcher or team of researchers integrates (a) qualitative and quantitative research questions, (b) qualitative methods and quantitative research designs, (c) techniques for collecting and analyzing qualitative and quantitative data, and (d) qualitative findings and quantitative results" (Pluye & Hong, 2014, p. 30). Besides from the combination of qualitative and quantitative components, other core characteristics of mixed methods research include the use of a specific research design to organize the procedures and the integration of phases, data and results of both components (Creswell & Plano Clark, 2018). The following paragraphs will address these two characteristics.

In general, three core mixed methods designs can be identified: convergent, sequential explanatory, and sequential exploratory (Creswell & Plano Clark, 2018; Pluye & Hong, 2014). These designs differ based on their intent: converging qualitative and quantitative results to enhance understanding of a phenomenon (convergent design); explaining quantitative significant or nonsignificant results with qualitative data (sequential explanatory design); or exploring a phenomenon by developing and applying quantitative measures or intervention grounded on qualitative data (sequential exploratory design) (Creswell & Plano Clark, 2018). Another difference concerns the sequencing, also named timing, which refers to the temporal relationship between the data collection and analysis of the qualitative and quantitative components (Plano Clark & Ivankova, 2015). In the convergent design, the data collection and analysis of both components are usually (but not necessarily) performed concomitantly and are usually (but not necessarily) independent from each other. In sequential designs, the sequencing is considered dependent since the results of one phase are used to inform the following phase. Also, in these

latter designs, the order of the phases will determine the choice of the design: in the sequential explanatory design, the quantitative phase is performed first to inform the qualitative phase, whereas in the sequential exploratory design, the quantitative phase will build on the results of the qualitative phase. A third difference pertains to the point of interface, which refers to when the integration between the qualitative and quantitative components occurred (Creswell & Plano Clark, 2018). In a convergent design, the point of interface occurs during or after the data collection and analysis of both components. In sequential designs, the point of interface occurs between and after the phases; after the data collection and analysis of one phase have been completed. Table 4 provides a summary of the main characteristics of each mixed methods design.

Table 4. Three Core Research Designs in Mixed Methods Research

Mixed methods designs	Intent	Sequencing	Point of interface
Convergent	• Converge qualitative and quantitative results	• Independent	• During or after qualitative and quantitative data collected and analyzed
Sequential explanatory	• Explain the initial results in more depth	• Dependent • Quantitative then qualitative	• Between and after phases
Sequential exploratory	• Explore a phenomenon	• Dependent • Qualitative then quantitative	• Between and after phases

Integration is another core component of mixed methods research and is defined as the "explicit interrelating of the quantitative and qualitative component in a mixed methods study" (Plano Clark & Ivankova, 2015, p. 40). Mixed methods research is more than the sum of individual qualitative and quantitative components as expressed by the following equation: 1 + 1 = 3 (Fetters & Freshwater, 2015). Integration is crucial to justify the added value for performing a mixed methods research. Several strategies have been developed to help researchers carry out integration in mixed methods research (Fetters, Curry, & Creswell, 2013; Guetterman, Fetters, & Creswell, 2015). Pluye, Garcia Bengoechea, Granikov, Kaur, and Tang (2018) analyzed 93 health-related mixed methods studies published in 2015 and identified three main types of

integration and nine specific strategies (Table 5). These strategies are concerned with integrating qualitative and quantitative phases, results and data.

Table 5. Types and Strategies of Integration Used in Mixed Methods Research

Types of integration*	Integration strategies
1. Connection of phases	1.1 Connecting the results of the qualitative phase to data collection of the quantitative phase
	1.2 Connecting the results of the quantitative phase to data collection of the qualitative phase
	1.3 Following a thread
2. Comparison of results	2.1 Comparing qualitative and quantitative results obtained from separate data collection and analysis
	2.2 Comparing qualitative and quantitative results obtained from interdependent data collection and analysis
	2.3 Comparing divergences of qualitative and quantitative results
3. Assimilation of data	3.1 Transforming qualitative data into quantitative data (quantitizing)
	3.2 Transforming quantitative data into qualitative data (qualitizing)
	3.3 Merging qualitative and quantitative data

*from Pluye et al. (2018)

As illustrated in Figure 5, the types of integration differ based on what is being integrated. In the first type, the integration occurs by connecting the phases; between the results of phase 1 and data collection phase 2. In the second type, the integration consists of comparing the results of the qualitative and quantitative components, and occurs once the data of both components have been collected and analyzed. In the third type (assimilation of data), the integration occurs at the level of the data. Once the data have been collected, the data of one component are transformed and then combined with those from the other component.

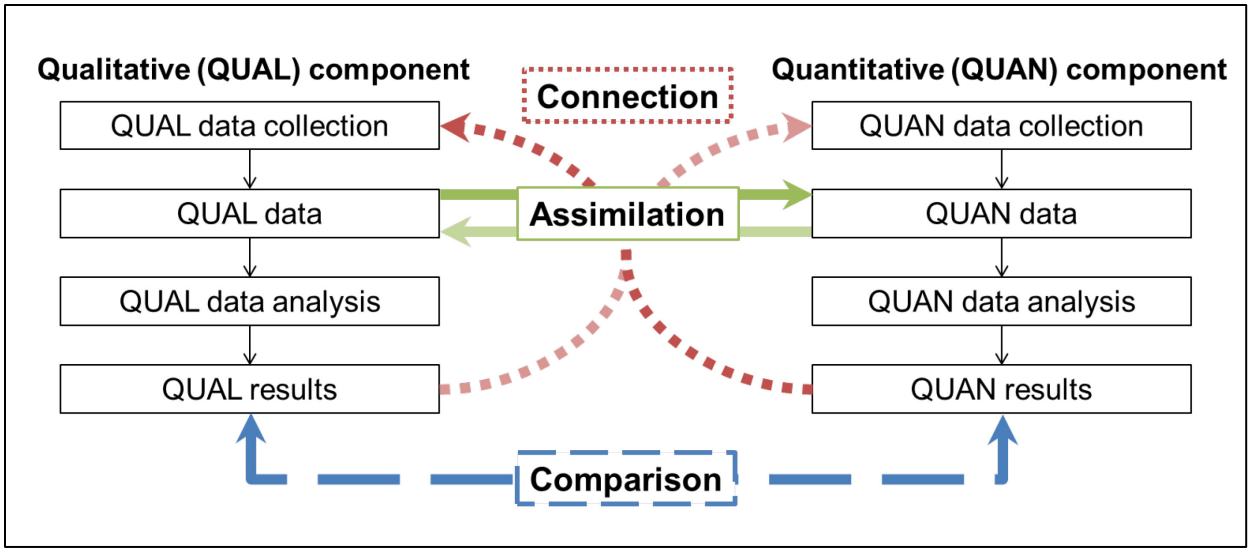


Figure 5. Three Types of Integration in Mixed Methods Research

In this project, the design used was sequential exploratory mixed methods. This design was used because it best fits the aim of this project that was to revise a critical appraisal tool. Sequential exploratory design is often advocated for tool development and assessment, and is suitable for exploration of a phenomenon since the first phase is qualitative (Creswell & Plano Clark, 2018). To achieve the project's aim, a qualitative phase was first needed to provide a general understanding of problems encountered by MMAT users and identify areas for improvement, which then informed the quantitative phase that addressed one of these problems.

Figure 6 presents an overview of the study design of this project. First, a qualitative descriptive study was conducted to collect data on the experience of MMAT users. Next, the team met to discuss the results and plan the following phase. From the results of phase 1, it was decided to focus on the criteria that were considered more difficult to judge by MMAT users, and those that were scarcely studied in the literature and lack of consensus: the criteria on qualitative, survey, and mixed methods research. In a second phase, a modified e-Delphi study was carried out with a group of experts to identify relevant criteria for appraising qualitative, survey, and mixed methods studies. The results of both two phases were used to revise the MMAT.

The type of integration used in this project consisted of connection of phases (strategy 1.1 in Table 5), where the phases were aligned sequentially and the results of the first qualitative phase informed the data collection of the second quantitative phase (Pluye et al., 2018). In this project, the combination of qualitative and quantitative data provided complementary information for the revision of the MMAT: phase 1 identified areas for improvement needed and phase 2 focused on one area of improvement and identified criteria that need to be modified, removed or added.

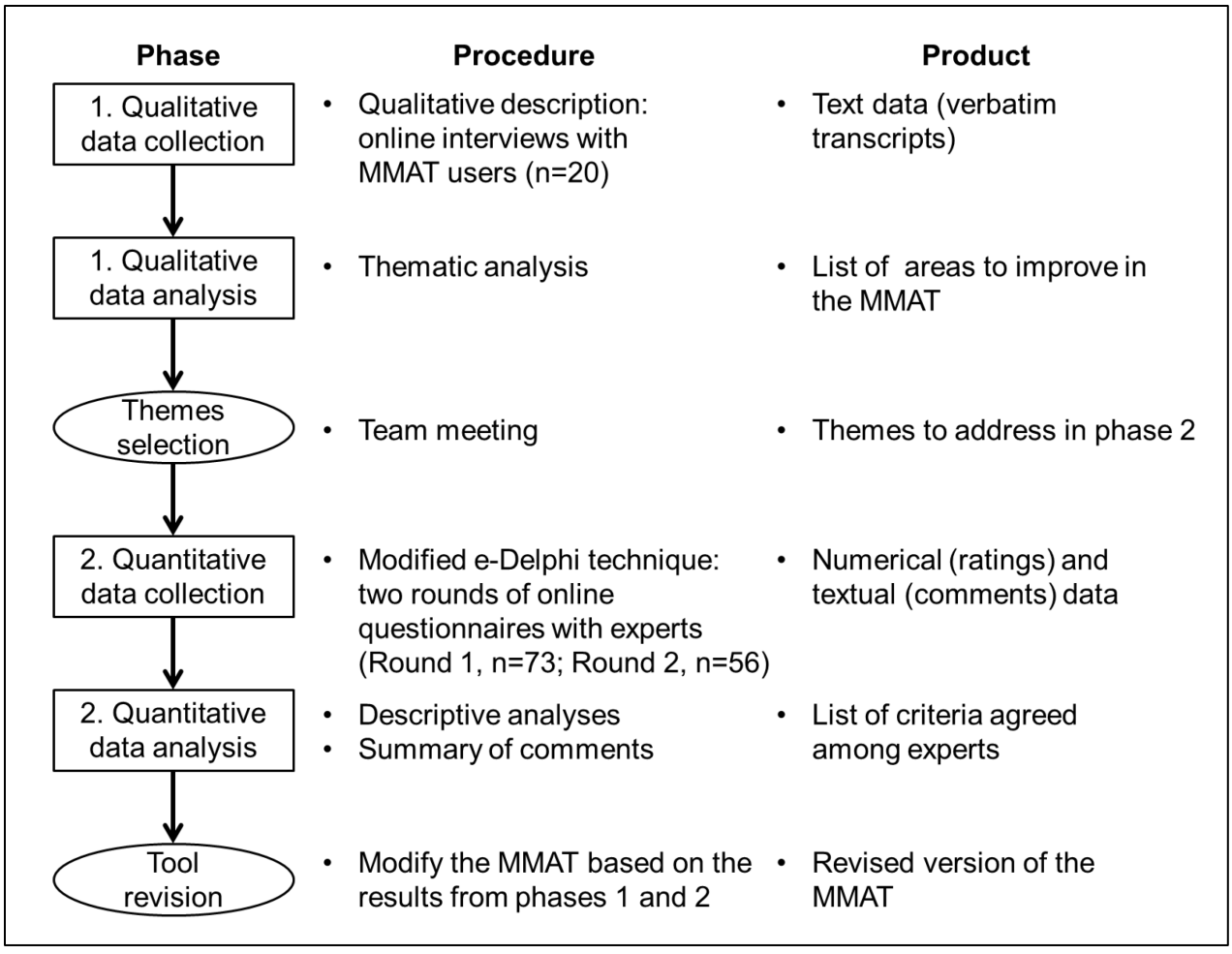


Figure 6. Diagram of the Overall Design of the Project

4.2 Methodology of Phase 1: Qualitative Descriptive Study

In phase 1, a qualitative descriptive study was conducted with MMAT users. Qualitative description is a qualitative approach aiming to provide a rich and straight description of experience or event in the language close to the participants' language (Neergaard, Olesen, Andersen, & Sondergaard, 2009). The seminal papers on this approach were written by Sandelowski (2000, 2010) at the beginning of the 21st century. This approach fitted well with the aim of this phase that was to identify the views and experiences of researchers regarding the use of the MMAT, which was useful to identify the changes to be made in the MMAT. Qualitative description is particularly relevant for collecting factual responses to questions and providing information directly from the perspective of the participants experiencing the phenomenon under investigation (Bradshaw, Atkinson, & Doody, 2017; Colorafi & Evans, 2016; Neergaard et al., 2009). Also, qualitative description has been advocated in a mixed methods research for the development and refinement of questionnaires or interventions as well as for needs assessment (Neergaard et al., 2009).

Qualitative description is still rarely presented in qualitative research reference books. When referring to qualitative research, the five most typical approaches stated are phenomenology, ethnography, grounded theory, narrative research, and case study (Creswell, 2013). Although not often included in reference books, qualitative description is one of the most popular qualitative approaches used in health sciences (e.g., in nursing). In a review of studies published in eight nursing research journals in 2005 and 2006, Polit and Beck (2014) identified that more than half (52%) of the qualitative studies were qualitative description, which was more frequently used than other qualitative approaches such as phenomenology (20%), grounded theory (11%), and ethnography (1%). Qualitative description differs from other qualitative approaches in that its aim is to seek to provide a comprehensive summary of experience or event (Sandelowski, 2000). It stays close to the data and events (data-near) and involves low-inference interpretation (Sandelowski, 2000). The other qualitative approaches will usually involve higher level of interpretation to achieve other aims such as understanding the essence of individuals' experience (phenomenology), developing a theory (grounded theory), providing an in-depth understanding of cases (case study), exploring an individual's life (narrative research), and interpreting the shared patterns of a group culture (ethnography) (Creswell, 2013). Qualitative

description also differs from qualitative interpretive description approach. The latter uses constant comparative methods, and aims to provide a coherent conceptual description that considers the thematic patterns and commonalities of a phenomenon (Thorne, 2008).

Five main features have been highlighted to characterize qualitative description (Kim, Sefcik, & Bradway, 2017; Neergaard et al., 2009; Sandelowski, 2000):

1. Philosophical orientation: qualitative description lies within a naturalistic inquiry. The phenomenon is studied in its 'natural' context, i.e., in the setting of the participants experiencing the phenomenon (not in a laboratory or a controlled environment). There is no pre-selection or manipulation of variables as well as no a priori commitment to one theoretical view.
2. Sampling: purposeful sampling techniques are typically employed, such as maximum variation sampling, to obtain rich information and broad insights.
3. Data collection: usually uses individual and/or focus group interviews with minimally structured or semi-structured open-ended interview questions. Data collection can also be undertaken through observations of the events, and examination of documents and other relevant materials.
4. Data analysis: a common strategy used is qualitative content analysis. Some will also perform 'quasi-statistical analysis' by providing numbers to summarize data such as frequencies.
5. Outcomes: provides a straight description of data organized in a way that fits the data and in a language similar to the participants' language.

Details about the participants' recruitment, data collection, sampling and data analysis are provided in Chapter 5. In summary, the method used in this phase 1 was the following:

- Participants: the participants were MMAT users, i.e., researchers who had used the MMAT for appraising the quality of studies. They were recruited from those who had published a systematic review in which they used the MMAT, or those who had contacted the developer for questions or permission to use the MMAT.
- Data collection: 20 individual semi-structured interviews.

- Sampling: maximum variation sampling, ensuring that the participants were from different institutions, countries, and occupations.
- Data analysis: thematic analysis, including first an inductive analysis approach of the data to identify the main themes and then a deductive approach to organize the themes within an existing framework on system acceptability.

The results of the interviews with MMAT users were discussed during a two-hour meeting of the MMAT developers. The MMAT developers are researchers that participated in the initial development of the MMAT: Drs. Pierre Pluye, McGill University, Montréal, Canada; Gillian Bartlett, McGill University, Montréal, Canada; Felicity Boardman, Warwick Medical School, Coventry, England; Margaret Cargo, University of Canberra, Canberra, Australia; Marie-Pierre Gagnon, Université Laval, Québec, Canada; Frances Griffiths, Warwick Medical School, Coventry, England; Belinda Nicolau, McGill University, Montréal, Canada; Alicia O'Cathain, University Sheffield, Sheffield, England; and Marie-Claude Rousseau, INRS–Institut Armand-Frappier Research Centre, Laval, Canada. In addition, three new members were added for their expertise in mixed methods research and HTA: Drs. Pierre Dagenais, Université de Sherbrooke, Sherbrooke, Canada; Sergi Fàbregues, Universitat Oberta de Catalunya, Barcelona, Spain; and Isabelle Vedel, McGill University, Montréal, Canada.

During the meeting, the planning of the next phase was discussed. The themes pertaining to the comprehensiveness and clarity of the criteria were identified among the most problematic and required further investigations. Also, as found in the literature review (section 3.3.2), most CATs were developed for quantitative studies, mainly RCT and NRS. Since several recent CATs for RCT and NRS were developed with experts, it was judged unnecessary to perform further investigation for these designs. Several studies using Delphi technique or surveys with RCT and NRS experts can be identified in the literature (Hayden, Côté, & Bombardier, 2006; Sindhu, Carpenter, & Seers, 1997; Slim et al., 2003; Verhagen et al., 1998; Yates, Morley, Eccleston, & Williams, 2005). Also, international committees of experts on RCT and NRS have been created in the Cochrane Collaboration to conduct empirical research on the impact of biases on systematic reviews and on how to identify and address biases in systematic reviews: the Cochrane Bias Methods Group (http://methods.cochrane.org/bias/) and the Cochrane Non-

Randomized Studies Methods Group (http://methods.cochrane.org/nrsi/). It was suggested to map the criteria in the existing tools and compare them with the MMAT to check if it captures all the dimensions that are internationally considered important. The team members were more concerned about the paucity of tools and lack of consensus on the other categories of the MMAT, i.e., qualitative, quantitative descriptive, and mixed methods studies. The literature review did not find studies that attempted to reach a consensus on specific criteria for these types of studies. Thus, the team decided to focus on these three categories of studies in the phase 2 of the project. Among the different quantitative descriptive study designs, the team decided to focus on survey research because of the lack of consensus studies and rigorous CATs. Also, survey research is often included in SMSRs and are among the most commonly used methods in mixed methods research (Bryman, 2006).

4.3 Methodology of Phase 2: Modified e-Delphi Technique

The aim of phase 2 was to identify the most relevant methodological criteria for appraising the quality of qualitative, survey, and mixed methods research. To achieve this aim, a Delphi technique was used. The Delphi technique consists of a group technique using an iterative multi-staged survey with experts to achieve consensus on important issues with no previous agreement (Keeney, Hasson, & McKenna, 2011). It was developed in the 1950s by the Rand Corporation in California as a forecasting tool to estimate the effect of atomic warfare as part of the defense plan (Pill, 1971). The Delphi technique is based on the assumption that group judgments are considered more valid and reliable than individual opinions (Keeney et al., 2011). This technique is useful for topics with contradictory or insufficient evidence (Hasson, Keeney, & McKenna, 2000). This is especially the case for qualitative, survey and mixed methods research where they are still few validated CATs and no clear consensus on how to perform quality appraisal. The Delphi technique has been used for the development of several CATs (Downes, Brennan, Williams, & Dean, 2016; Hayden et al., 2006; Mokkink et al., 2010; Pincus et al., 2011; Sindhu et al., 1997; Verhagen et al., 1998; Yang et al., 2009; Yates et al., 2005). Also, for the content validation of assessment tools, it is recommended to involve experts at different stages of the tool development such as in the definition of constructs, generation and selection of items, and evaluation of the tool (Haynes, Richard, & Kubany, 1995).

Many variations of the Delphi technique can be found in the literature. Hasson and Keeney (2011) described 10 different types of Delphi techniques (i.e., classical, modified, decision, policy, real time/consensus conference, e-Delphi, technological, online, argument, and disaggregative policy) that differ based on the aim of the study, the number of rounds needed, the administration requirement, the target participants, and the method used in round-one. In this study, two types were chosen: modified and e-Delphi. These types differ from the classical type on two main points. First, the classical type usually employs three or more rounds of questionnaires (Hasson & Keeney, 2011). The first round uses an open-ended set of questions to generate a list of items. Then, the subsequent rounds use structured questionnaires to reduce the number of items. In the modified Delphi, the first round uses pre-selected items that can be drawn from different sources such as focus groups, interviews or literature reviews (Hasson & Keeney, 2011). In this study, the items included in the questionnaire for the first round were identified from the literature review of CATs. Second, traditionally, the classical type is administered through postal services (Hasson & Keeney, 2011). In this study, an e-Delphi was used, meaning that the questionnaires were administered online using LimeSurvey, a web-based tool designed for creating surveys, and available and hosted on the McGill University server. The potential participants were recruited through email. Email is one of the main communication means of the targeted experts in this study (i.e., researchers) and is convenient to joint an international group of experts. Compared to postal mail, web-based survey is more time efficient and cost-effective (Hoonakker & Carayon, 2009).

Despite the variety of types, four main features can characterize all Delphi studies (Rowe, Wright, & Bolger, 1991; von der Gracht, 2012):

1. Anonymity: the participants are not aware of who are participating in the study. The whole process is coordinated by a facilitator who is in charge of developing the questionnaires and analyzing the responses. This feature can have the advantage of reducing the effect of dominant participants and avoiding social pressure. It can also lead to higher response rates.
2. Iteration: a series of rounds is needed (at least two rounds). This process is usually repeated until stability in the responses is attained. It allows the participants to decide

whether to modify previous answer or to remain with their initial opinion based on the feedback received from the group.

3. Controlled feedback: after each round, each participant is provided a summary of the opinions of the other participants and a reminder of their responses. The feedback provided to the participants is controlled by the facilitator; i.e., the facilitator is responsible for defining the type of feedback and its presentation.
4. Statistical group responses: the group responses are presented numerically and/or graphically and can include measures of central tendency, dispersion and frequencies. This is used to encourage the convergence of opinions and as an indicator of the strength of consensus.

Details on the participants, recruitment, data collection and analysis of the modified e-Delphi study are presented in Chapter 6. Here is a summary of the method used in phase 2:

- Experts: the experts were researchers with an academic or research position who have research interests on the methodological development of qualitative, survey, or mixed methods research.
- Sampling: the sampling was purposeful. A total of 196 experts were invited to participate in this study.
- Questionnaire development: the questionnaire in Round-one was developed from a literature review of CATs. The questionnaires included a list of criteria and an open-ended box for commentaries.
- Scale: a 5-point ordinal Likert scale was used (1=not at all relevant, 2=slightly relevant, 3=moderately relevant, 4=very relevant, and 5=extremely relevant).
- Number of rounds: two rounds of questionnaires were performed.
- Data analysis: an agreement index was calculated for each item, which was computed by dividing the number of experts giving a rating of 4 (very relevant) or 5 (extremely relevant) by the total number of experts. The content of open-ended boxes was summarized.
- Consensus: a cut-off score of 0.80 was used to retain criteria.

4.4 Statement of Ethics

This project was approved by the Institutional Review Board (IRB) of the Faculty of Medicine at McGill University on May 14, 2015 (project # A05-E26-15B; Appendix 6). All participants from the two phases of the project completed a consent form prior to their participation (Appendix 6). The consent form was made available online using LimeSurvey that was hosted on the McGill University server. The participants received a link to the online consent form by email and were asked to read and approve it prior to the interview (phase 1) or before completing the questionnaire (phase 2). Also, in the modified e-Delphi study (phase 2), the participants were asked in the consent form if they accepted to have their names stated in the acknowledgment section of publications. This is common practice in studies conducted with experts.

In phase 1, the interviews were recorded using a digital voice recorder as well as the GoToMeeting recording function (for interviews that used this software). Prior to starting the recording, all participants were asked if they had any questions on the study and if they accepted to be recorded. A number was allotted to the participants and all names mentioned during the interviews were removed during the transcription.

In phase 2, the responses were kept anonymous to the panel and the members of the panel only received group's responses and a reminder of their responses. The data collected was only accessible to the persons directly involved in this study. No personally identifiable information was presented in the data file used for the analysis and reporting of results.

CHAPTER 5. METHODS AND RESULTS OF PHASE 1 – QUALITATIVE DESCRIPTIVE STUDY

This chapter focuses on the qualitative phase of the project that was conducted with researchers who have used the MMAT. The objective of this study was to explore the views and experiences of the researchers on the MMAT. A total of 13 themes were identified and classified into the dimensions of usefulness (i.e., utility and usability). Results from this study helped to formulate recommendations to revise the MMAT. The methods and results are presented in the following manuscript (Paper #2 of this manuscript-based dissertation). This manuscript is published in the Journal of Evaluation in Clinical Practice (Hong, Gonzalez-Reyes, & Pluye, 2018). The invitation email and interview guide for this phase 1 are provided in Appendix 7.

PAPER #2: Improving the Usefulness of a Tool for Appraising the Quality of Qualitative, Quantitative and Mixed Methods Studies, the Mixed Methods Appraisal Tool (MMAT)

Published in the Journal of Evaluation in Clinical Practice
(https://doi.org/10.1111/jep.12884)

Quan Nha Hong, OT, MSc, PhD candidate, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, H3S 1Z1, Canada, quan.nha.hong@mail.mcgill.ca

Araceli Gonzalez-Reyes, MSc, PhD candidate, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, H3S 1Z1, Canada, araceli.gonzalezreyes@mail.mcgill.ca

Pierre Pluye, MD, PhD, Full professor, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, H3S 1Z1, Canada, pierre.pluye@mcgill.ca

Corresponding Author:
Pierre Pluye, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, Canada, H3S 1Z1; Tel: 1-514-398-8483, Fax: 1-514-398-4202.
Email address: pierre.pluye@mcgill.ca

Running title: Usefulness of the Mixed Methods Appraisal Tool

Keywords: qualitative research, quality appraisal, mixed studies reviews, systematic review, usefulness, usability, utility.

ABSTRACT

Rationale, aims and objectives: Systematic reviews combining qualitative, quantitative, and/or mixed methods studies are increasingly popular due to their potential for addressing complex interventions and phenomena, specifically for assessing and improving clinical practice. A major challenge encountered with this type of review is the appraisal of the quality of individual studies given the heterogeneity of the study designs. The Mixed Methods Appraisal Tool (MMAT) was developed to help overcome this challenge. The aim of this study was to explore the usefulness of the MMAT by seeking the views and experiences of researchers who have used it.

Methods: We conducted a qualitative descriptive study using semi-structured interviews with MMAT users. A purposeful sample was drawn from the researchers who had previously contacted the developer of the MMAT, and those who have published a systematic review for which they had used the MMAT. All interviews were transcribed verbatim and analyzed by two coders using thematic analysis.

Results: Twenty participants from eight countries were interviewed. Thirteen themes were identified and grouped into the two dimensions of usefulness, i.e., utility and usability. The themes related to utility concerned the coverage, completeness, flexibility, and other utilities of the tool. Those regarding usability were related to the learnability, efficiency, satisfaction and errors that could be made due to difficulties understanding or selecting the items to appraise.

Conclusions: On the basis of the results of this study, we make several recommendations for improving the MMAT. This will contribute to greater usefulness of the MMAT.

INTRODUCTION

Systematic reviews combining qualitative, quantitative, and/or mixed methods studies, are more and more popular due to their potential for addressing complex evaluation questions that matter in clinical practice[1, 2]. Indeed, including different types of studies in a review can provide a richer understanding of the impact of contextual factors, help focusing on outcomes that are important for patients, and explore the diversity of effect across studies[3]. These reviews have various labels such as systematic mixed studies reviews[4], mixed methods research synthesis[5], and integrative review[6]. The first label refers to combining qualitative, quantitative and/or mixed methods studies while the second one can also refer to combining qualitative and quantitative methods (such as thematic synthesis and meta-analysis)[7]. Hereinafter, we will use the term 'systematic mixed studies reviews' to designate this type of review. While they are increasing popular[7], these reviews present several challenges given the heterogeneous nature of study designs, including the critical appraisal of the quality of individual studies. Critical appraisal is a core step of systematic reviews and consists of a systematic and careful examination of studies to ensure they are trustworthy[8, 9].

Critical appraisal tools have been developed to formalize the quality appraisal process and ensure it is done in a systematic, transparent, and reproducible manner[10]. A large variety of these tools exists and are, for most part, checklists and scales of quality appraisal items[11]. For example, authors of literature reviews have identified 94 tools for randomized controlled trials (RCT)[12], 194 for non-randomized studies[13], 13 for mixed methods studies[14], and 58 for qualitative research[15]. The wide variety makes it difficult for reviewers to choose the most appropriate one(s). This is particularly true for systematic mixed studies reviews since the heterogeneity in the designs of the included studies requires that reviewers search for, select, and learn how to use several tools. Also, there is a lack of agreement regarding the most appropriate critical appraisal tools and approaches to use[11-13]. Many tools were not developed using rigorous development process including sound validation and reliability testing[16-18]. To address this, Whiting et al.[19] recently proposed a framework for developing quality assessment tools, which includes three key stages: initial steps (including identifying needs and scope for a new tool), tool development, and dissemination.

The Mixed Methods Appraisal Tool (MMAT) allows for the critical appraisal of quantitative, qualitative, and mixed methods studies and was developed to address the challenges of critical appraisal in systematic mixed studies review. The MMAT is rooted in a literature review on systematic mixed studies reviews conducted in 2006[4]. To provide proof-of-concept of the feasibility of the MMAT, the research team conducted a pilot study and subsequent studies of interrater reliability. These studies showed that it is relevant to researchers and decision/policymakers and feasible for them to use[20], and that there is a variability of agreement of the items ranging from poor to perfect and a need for further testing and refinement of this tool[20, 21]. To further the development and testing of the MMAT, more research is needed with researchers who had used this tool.

Since its development, the MMAT has been cited in more than one hundred systematic reviews, and its website[22] has been visited more than 20,000 times. This widespread use made it possible to explore the views and experiences of researchers who have used the MMAT and were not directly involved in its initial development (hereinafter 'MMAT users'). Our research question was: What are the views and experiences of researchers regarding the use of the MMAT? The results of this study with users contributed to identifying the key areas for improvement that is required in the MMAT.

METHODS

A qualitative descriptive method[23, 24] was employed with MMAT users. This method fits well with the aim of this study that focused on describing the experience of MMAT users. This method stays close to the data and focuses on reporting the manifest content of data, rather than being highly interpretive and conceptual[23, 24]. Qualitative description is appropriate in mixed methods research for the development and refinement of questionnaires or interventions[25].

Description of the Mixed Methods Appraisal Tool (MMAT)

The latest version of the MMAT (version 2011) includes two screening questions and 19 items for appraising the methodological quality of five categories of studies: qualitative studies (4 items), randomized controlled trials (4 items), non-randomized studies (4 items), quantitative

descriptive studies (4 items), and mixed methods studies (3 items). The screening questions are used to exclude non-empirical studies from the appraisal stage, i.e., research that is not based on experience (e.g., observation, experiment, or simulation) such as reviews and theoretical papers[26]. The MMAT was conceived so that one set of items can be used when appraising a qualitative or quantitative study. When appraising mixed methods studies, three sets of items are assessed: the qualitative set, a quantitative set (either, the randomized controlled trial, non-randomized studies, or the quantitative descriptive studies), and the mixed methods set. Each item is rated on a categorical scale (yes, no, and cannot tell) and the number of items rated 'yes' are counted to provide an overall score (see supplementary file). The MMAT is available online (http://mixedmethodsappraisaltoolpublic.pbworks.com) and comes with a user manual (tutorial) in which each item is described, and examples and references are provided. For each category of studies, examples of common study designs are provided (see supplementary file).

Study Participants

A purposeful sample of researchers with experience using the MMAT was generated by two means. First, forward citation tracking of three papers on the MMAT[4, 20, 22] was performed on September 6, 2015 in Google Scholar. These references had been cited, respectively, 51, 156, and 54 times. From these citations, we selected the systematic reviews published after 2011 (year of the latest version of the MMAT) that included more than 10 studies and collected the name and email address of the first authors. Second, the primary developer of the MMAT had a list of 81 researchers who had contacted him over the years requesting permission to use it for research or training purposes, clarification on how to use it, or requesting for the latest version. A maximum variation sampling was used to account for the different institutions, countries, and occupations of these researchers. An email was sent to 72 researchers inviting them to participate in an interview in English or French regarding their experience using the MMAT.

Data Collection

We conducted semi-structured interviews with MMAT users either through Skype or GoToMeeting. During the interview, a semi-structured guide was used to collect information

pertaining to their (a) research experience (e.g., fields of interest, number of years of research experience, research methods experience, and occupation), and (b) experience using the MMAT (e.g., number of papers appraised using the MMAT, study designs of the papers appraised, perceived utility of the MMAT). The interview guide was developed to elicit MMAT users' perspectives and experiences with different parts of the tool, i.e., the items, the scale, the tutorial, and the five study design sets. Five questions were posed: (a) What do you like about the MMAT and why?, (b) What do you dislike about the MMAT and why?, (c) Did you encounter any problems when using the MMAT?, (d) Did you make any changes to the tool during your project?, and (e) Were you able to use the MMAT to appraise all the papers included in your reviews? The interviews were recorded using a digital recorder.

The interview guide was piloted with three students who had used the MMAT in their master's research. This pilot test aimed to verify the clarity of the questions and their order, to estimate the time of the interview, and to test different communication media (phone, Skype, and in-person) and the recording quality.

Data Analysis

A professional transcriber transcribed the interviews and the interviewer checked the verbatim transcripts for accuracy. The transcripts were analyzed using thematic analysis[27]. Two coders independently coded the transcripts using a specialized software program (NVivo 11). Initially, they used open coding, reading and re-reading the transcripts to generate a preliminary list of codes. After analyzing three interviews, the two coders met to compare and discuss their codes and establish a codebook. This process was iterative and repeated until no substantive new codes were identified. The codes were then analyzed and combined into meaningful groups to identify initial themes. At this stage, the themes were grouped into three broad categories: strengths of the MMAT, difficulties encountered when using the MMAT, and changes made or suggested in the MMAT. Once the themes were identified, the team met to discuss how to organize them coherently and meaningfully. Discussions among the team led to using the framework on system acceptability to organize the themes (Figure 1). Developed in the field of human-computer interaction, this framework presents the main dimensions required to ensure that a system is good enough to satisfy the users' needs and requirements[28]. Within this

framework, a system overall acceptability is composed of its social and practical acceptability. To analyze the practical acceptability of a system, several dimensions can be considered such as its cost, reliability and usefulness. In this study, we focused on the usefulness dimension that is defined as whether the system can achieve its desired goal[28]. We considered that the MMAT is a system that users will use to achieve the intended goal of appraising the quality of qualitative, quantitative, and mixed methods studies. All the themes identified in the open coding were interpreted using, and grouped into, the dimensions of usefulness in this framework, i.e., utility and usability (Figure 1).

Insert Figure 1 about here

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of the Faculty of Medicine of McGill University (project # A05-E26-15B). All participants completed a consent form prior to the interview. Participants were numbered and no identifying information was presented in the data file used for the analysis.

RESULTS

A total of 72 invitation emails were sent between November 2015 and March 2016, of which 20 resulted in interviews. The reasons for non-participation were: did not respond to the invitation email (n=42), had invalid email address or out-of-office message (n=4), had not used the MMAT (n=3), was not available during the period of the interviews (n=1), was not interested (n=1), and used the MMAT too long ago to remember (n=1). The interviews were conducted in English (n=16) or French (n=4), and lasted between 21 and 48 minutes.

The 20 participants were affiliated with institutions from eight different countries. They were mostly female (n=17) and affiliated with a university (n=19). Their research areas were predominantly in health sciences (including nursing, public health, global health, community health, palliative care, primary care, cardiovascular, oncology, and gerontology). Nearly half of the participants were doctoral candidates. Most were mixed methods researchers (n= 9); whereas the others identified themselves as primarily qualitative (n=5) or quantitative (n=6) researchers (Table 1). With the exception of one participant who used the MMAT in a journal club, all had

used it in a systematic review. Participants used the MMAT results to describe the quality of included studies (n=14), exclude studies from the review (n=3), justify the quality criteria extracted from studies (n=1), make recommendations (n=1), and compare with the appraisal of other critical appraisal tools (n=1).

Insert Table 1 about here

A total of 13 themes were identified and grouped into the two dimensions of usefulness, i.e., utility and usability (Figure 1).

Utility

Utility is defined as whether or not the tool can function as needed[28]. Five themes were found regarding the utility of the MMAT; two addressed its coverage, and one each for completeness, flexibility, and other utilities.

Coverage

Two themes were related with the scope of designs covered by the MMAT.

Theme 1 – Comprehensive tool: The MMAT users appreciated that the tool can be applied to several study designs (qualitative, quantitative, and mixed methods studies):

> *The thing that I liked about it it's an all in one package. […] There's so much to write, there's so much to analyze, if you have 2 different or 3 different tools to use. I can see that the development of this tool was also based on previous work of critical appraisal and all of those. But for me, what is very… what I really liked the most about it, it's there, it's all in one. You can use it… yeah… You don't have to use any other tool. (P16)*

Theme 2 - Study designs that could not be appraised with the tool: Some MMAT users mentioned that the items in the MMAT were less relevant for some study designs such as cost-effectiveness studies, political analysis, transcultural adaptation, and pragmatic trials:

> *Like studies in political science, political analysis, policy development process, they did not really fit with the MMAT. A second type of studies that I had difficulty assessing with the MMAT were studies in economics, cost-effectiveness studies. (P04)*

Completeness

One theme addressed concerns about the completeness of the tool. The completeness refers to the degree to which all important items to appraise the quality of studies are included in the MMAT.

Theme 3 - Concerns about completeness of the tool: Because the MMAT includes four items for each research design set, MMAT users were concerned that the tool might be 'too simple', 'superficial', 'global', and would not discriminate 'good' and 'bad' studies. Some MMAT users mentioned that items were missing in the tool such as those concerning conflict of interest, quality of reporting, confounding variables, selective reporting bias, sample size, external validity, theoretical underpinnings, publication bias, triangulation, data analysis, and ethics.

> *So my concern, initially when I first started to use it and when comparing with other types of appraisal tools, I was afraid that it might be missing some appraisal items. At the time it had been… the pilot study had been done for validation, so that was reassuring, but at the same time that would have been just concerns about completeness. (P18)*

Flexibility

One theme pertained to the need to adapt the MMAT. We interpreted this to be about the flexibility of the tool, which refers to its ability to be modified based on the research topic or study design.

Theme 4 – Need to adapt the tool to the topic of the review: Some users suggested having a more flexible tool that could be tailored to the topic of their review. For example, they suggested providing more weight to certain items or adding optional items they judged important in their field. Also, some MMAT users questioned the utility of the two screening questions and suggested that they be removed when the selection criteria are limited to empirical studies. Moreover, they suggested having cut-off values in the items that could be adapted to their field.

> *And also in the observational ones, we wanted to be able to discriminate or give a bit of a better weighting to perspective of longitudinal studies. So within the justification of measurements, we also rated it high, we also gave an extra point if it was longitudinal perspective compared to cross-sectional. (P17)*

Other utilities

In addition to appraising the quality of studies, some users mentioned that the MMAT can have additional utility.

Theme 5 – Educational tool: The MMAT users liked that the tool was helpful to learn about study designs and that it was a relevant resource for graduate students:

> *And it's a really nice resource for students particularly, because we want to encourage them to think broadly when they think of systematic review and not to just think of the quantitative systematic review of intervention studies or the meta-analysis kind of reviews. And that gets really overwhelming. So this tool kind of consolidates a lot of ways of thinking about the quality of your studies into a single document that's useful for them to think through. (P08)*

Usability

Usability is defined as how well users can use the tool[28]. Compared with utility where no attribute is specified in the system acceptability framework, five usability attributes are defined: learnability, efficiency, memorability, errors and satisfaction (Figure 1)[28]. In this study, eight themes on usability were found and were related to four of these attributes.

Learnability

Learnability refers to how the tool is easy to learn[28]. Two themes were found on this attribute.

Theme 6 – Easy to use: The MMAT users liked that the tool was easy to understand, rate, and use:

> *[…] it was really clearly explained how you can include and exclude, how you're supposed to evaluate the studies, it was really well laid out. Easy for someone who's never done this kind of thing before to follow. The instructions are really good. (P19)*

Theme 7 – Improvements needed in the tutorial: Several comments were made on the tutorial. The MMAT users found the tutorial helpful to refer to. They appreciated the list of study designs and the explanations of the items. However, they mentioned that some explanations provided did not match the items. Some MMAT users suggested expanding the study designs list to include,

for instance, interpretive description, comparative studies, and survey. They also suggested adding information in the tutorial to facilitate the use of the tool, such as a title page, the explicit purpose of the MMAT, and an algorithm. Many mentioned that is was unclear how to score the 'cannot tell' response category and some suggested modifying the scale. Moreover, MMAT users suggested adding more examples of how to rate items, and clarifying how to compute an overall score and how to present the results of the appraisal:

> *The left-hand box is really useful. That's good because it helps you to classify the type of qualitative, what you've got. The right-hand side, where you're asking the questions, possibly give more specific examples maybe of what there is there. Because in all cases, I would say that would be relevant really. Like on the left-hand side definitely that's fine, I would leave it there. But maybe add some more examples on the right-hand side. (P21)*

Efficiency

Efficiency is defined as allowing for a high level of performance once the users have learned to use the tool[28]. One theme addressed this attribute.

Theme 8 – Short and quick: The MMAT users liked that the tool was simple, short, and allowed for completing study appraisal quickly:

> *I liked that it's simple and it's not too long. It's not an enormous task to go through. It's very clear to see which bits are going to be relevant to what I need. I can just go straight in there and see which areas I need to look at. (P14)*

Errors

Errors are defined as actions that do not accomplish the intended goal[28]. Given the goal of the MMAT is to appraise the quality of qualitative, quantitative, and mixed methods studies, we included in this attribute two themes on difficulties understanding the items or selecting of the items to appraise.

Theme 9 - Items not clear or difficult to judge: The MMAT users provided comments on items that were difficult to understand and rate. Four subthemes were identified.

Subtheme 9.1- Qualitative and mixed methods studies subject to interpretation: Several comments were made concerning items in the qualitative and mixed methods studies item sets

that were considered more difficult to judge, more open to interpretation or less precise compared with the quantitative study designs items.

> *There was some… and I think this is acknowledged in the template, some of the criteria were a little bit difficult to interpret, particularly around kind of the qualitative items about researchers' influence and the context, which were difficult to establish.[…] The mixed methods was similar to the qualitative ones in that they were a bit open to interpretation compared with the quantitative items. (P15)*

Subtheme 9.2- Several concepts in one item: MMAT users commented on the fact that some items include several concepts and suggested clarifying or modifying these items.

> *I think the ideas are quite clear. However, there are several concepts in the same question. So here, I think that was what I found difficult. Take question 3.2: 'are measurements appropriate regarding the exposure, control', etcetera. You see that in the parentheses there are a lot of concepts and each of these concepts could be a sub-question. (P11)*

Subtheme 9.3 – Missing information in papers: MMAT users pointed out that some items were considered more difficult to judge because of missing information in the papers appraised.

> *But 'is appropriate consideration given to how findings relate to the context?', that's very hard. And then 1.4 'is consideration given to how the researchers' influence or the interaction with the participants?'. Because of this word limit of publications and because qualitative… When you write a qualitative paper, you're already struggling for space, because most health science journals only allow you 3000 words. It's already a struggle to put in your citations and everything and everything counts, so I don't… I can't remember I have read a paper that goes into detail about how the researcher might have influenced the findings etcetera. (P07)*

Subtheme 9.4 – Unclear distinction between some items: MMAT users mentioned that the distinction between some items is subtle.

> *The one that I probably used least often and the one that I had the most questions about, - but again, I'm not using it all that often - is the RCT, the difference between the complete*

outcome data of 80% and the low withdrawal rate of 20%. That's a very fine line in my mind of what's the differentiation. (P13)

Theme 10 - Difficulty classifying the studies: MMAT users mentioned that they had difficulty deciding if they should use the non-randomized or the descriptive sets:

One of the things… quantitative non-randomized… quantitative descriptive… We had problems trying to classify some of the studies. We didn't have specific enough, sufficient details for you to be able to tell what type of study it is. So we ended up classifying the majority of studies as quantitative descriptive mainly because we didn't have sufficient information from the studies themselves (P09).

Satisfaction

Satisfaction refers to how pleasant the tool is to use[28]. Three themes were related to this attribute.

Theme 11 - Accessible online: The MMAT users liked that the tool was available online:

Another thing that I really liked about the tool is that it's online and everybody can get access to it. […] So when people ask me about that, I said "I can send you a link but it's right there online, you can just go in and look at it". And it's really really helpful for people. (P05)

Theme 12 - Website not user-friendly: The MMAT users provided comments on the navigation of the website:

I do remember being on your website and your website might be just a little bit tricky to navigate. (P06)

Theme 13 – Missing rating sheet: MMAT users proposed providing a rating sheet, such as an Excel document, that could be used to compute the ratings and calculate an overall score:

The only thing was that it was not available in a document that you can write in.[…] Yes like the Excel sheet I showed you. I don't know if that would be helpful. I just made it myself. (P10)

DISCUSSION

The development of the MMAT followed the framework for developing quality assessment tools[19]: initial steps (e.g., identify a need for an appraisal tool for systematic mixed methods studies), tool development (e.g., literature review, pilot testing, reliability testing), and dissemination (e.g., workshops, website, publications). However, this process is not linear and should include feedback loops to revise and refine the tool. To contribute to the revision of the MMAT, we explored the views and experiences of researchers who have used it. We identified 13 themes and classified them according to the dimensions of usefulness (utility and usability) as suggested by Nielsen[28]. Table 2 presents a summary of the themes. Regarding utility, our results pertain to the coverage, completeness, flexibility, and other utility of the MMAT. In term of usability, our findings point to issues of learnability, efficiency, errors, and user satisfaction. Some themes suggest potential areas for improvement in the MMAT (see * in Table 2).

Insert Table 2 about here

The MMAT users appreciated that the tool was easy to use, comprehensive, quick, short, and accessible online. These themes are considered strengths of the MMAT that should be maintained in subsequent revision of the MMAT. Having pre-defined items can be helpful to ensure that the key methodological aspects are examined in a systematic and transparent manner using a common approach for all included studies[10]. Since systematic mixed studies reviews can include a wide range of study designs, these tools can be particularly appealing to graduate students and researchers who are unfamiliar with certain study designs.

The results of this study can contribute to improve the ecological validity of the MMAT. Ecological validity is a subset to external validity and refers to the transferability of findings from an experimental context to the real-world environment[29, 30]. Interviewing other users that were not involved with the development of the MMAT can provide different and some more impartial views of the MMAT.

Recommendations for Improving the MMAT

On the basis of our results, six recommendations can be put forward for the MMAT.

First, the MMAT includes criteria for five broad categories of study designs and specific criteria for each design. Yet, our results show that choosing items is difficult for some studies, in particular for cross-sectional and single group studies. This difficulty could be addressed by clarifying the study design categories and adding a selection algorithm such as those developed and tested in Hartling et al. [31], and Seo et al. [32] for classifying quantitative study designs.

Second, the MMAT is focused on appraising methodological quality. In this study, the MMAT users underscored an important usability issue: poor reporting hinders the appraisal of some MMAT items. Inadequate reporting precludes adequate appraisal of how a study was conducted and its results[33]. Moreover, lack of reporting about a methodological criterion does not mean it was not met in the study[34, 35]. To address this issue, the MMAT has a 'cannot tell' response category and it is suggested to contact the researchers to obtain additional information. This approach has been critiqued since it can lead to risk of overly positive answers (i.e., tendencies of providing positive answers that do not necessarily reflect the reality of a study)[36]. Given that the reliability of the information provided may be questionable, some have recommended limiting the appraisal to published material and matching the quality of reporting with the level of information needed to appraise the methodological quality of a study[37]. This recommendation is an avenue to explore for the revised version of the MMAT. Items to include in the MMAT could be chosen on the basis of information that is typically reported. Another potential avenue is a two-step approach where inadequately reported papers are excluded on the basis of an initial reporting quality appraisal, and methodological quality of the remaining papers is subsequently appraised[38]. Carroll et al.[38] tested this approach and found that excluding inadequately reported papers does not influence the overall results of the synthesis in qualitative systematic reviews, although it might lead to exclusion of particular disciplines/perspectives. In a recent review, Verhage and Boels[39] concurred with Carroll et al.[38], but mentioned that, although the exclusion of inadequately reported papers does not affect the number and nature of the themes identified, it may influence the degree of nuance and the richness of the themes.

Third, the current version of the MMAT has four items per category of study design, which is few compared to other critical appraisal tools. Although our results show that the short and comprehensive nature of the MMAT is appreciated, they also indicate concerns about its utility due to its lack of completeness and missing items. The MMAT developers chose to focus

on efficiency, including only the most important items for judging the methodological quality of a study. Yet, in tool development, it is necessary to ensure that the tool adequately covers the construct is meant to assess (i.e., the methodological quality of studies in the case of the MMAT). This is related to the content validity of the tool[40] and will need to be further explored with methodological experts.

Fourth, our results suggest that the qualitative and mixed methods studies items are difficult to judge. These items were considered more subject to interpretation and less precise than the quantitative items. Several reasons could explain this difficulty such as the lack of reporting (e.g., unclear description and lack of details) precluding a proper appraisal, and the unfamiliarity of the reviewers with these types of studies. Also, in the MMAT, only one set of items were developed for qualitative and mixed methods studies while there are three different sets of items for quantitative studies (RCT, non-randomized, and descriptive). There is a need to provide more explanations and examples about how to interpret and rate these items in the MMAT. Also, further studies could explore the need to add items regarding specific qualitative approaches (e.g., qualitative descriptive, grounded theory, ethnography, and phenomenology).

Fifth, our results suggest making the MMAT more flexible by, for instance, adding optional, weighting items or modifying the cut-off values when judged necessary by the reviewers. This could improve the utility of the tool and help tailor it to the needs of the users. This is in line with Santiago-Delefosse et al.[15] who promote a flexible list of criteria for qualitative research based on their study with 46 participants. They found that consensus can be reached only for general criteria and that there was a lack of consensus on the definition of criteria and their weights. In addition to having core criteria for each design, the MMAT could include a list of validated items from which the researchers can choose to meet the specific needs of their review.

Sixth, the users' satisfaction when using a tool is another important usability issue that needs to be considered when developing a critical appraisal tool. Users who are pleased with the tool tend to recommend it to others[28]. Complementary materials such as a user manual or website can enhance users' satisfaction. On the basis of our results, concrete improvements to enhance users' satisfaction with the MMAT should be made such as improving the website navigation, providing more examples of rating in the tutorial, and adding a rating sheet.

Strengths and Limitations

We interviewed 20 MMAT users. Similar themes were mentioned by the MMAT users and data saturation was reached; further interviews would probably not have added new information to the overall results[41]. After the 8th interview, no new code emerged. The addition of interviews helped to provide more information on the themes. While our sample was heterogeneous with participants from several countries, working on a wide range of research topics mainly in health care, and having different expertise, almost all participants worked in university settings. Other potential MMAT users, such as health technology assessment professionals, were not reached. Also, nearly half of the participants were doctoral candidates, which can be representative of the main MMAT users. Indeed, systematic review is a method increasingly used at the graduate level. Some even suggest that systematic reviews be mandatory in doctoral programs[42].

Two authors of this study are familiar with the MMAT. The interviews were performed by the first author who was a doctoral candidate at the time of this study. She has gained experience with the MMAT one year prior to the interviews by collaborating as a second reviewer on systematic reviews. The last author is one of the developers of the MMAT and has been working on this tool since 2006. Their preconceptions of the MMAT could have influenced the interviews and analyses. Care was taken to make sure the data collected and analyzed represent the experience of the MMAT users such as involving a second coder that was not familiar with the MMAT, having independent coding, and developing a codebook. The coders did not encounter difficulties in reaching a consensus since the level of interpretation of data was low (analysis of the manifest content of interviews).

CONCLUSION

As systematic mixed studies reviews are gaining in popularity, appraisal tools that can be used to assess different study designs are needed. This study with MMAT users is a first important step in the improvement of its usefulness. The 13 themes identified and grouped into the system acceptability framework may be useful for developers of other critical appraisal tools.

ACKNOWLEDGEMENTS

The research team is grateful to the 20 participants who generously shared their time, perspectives and experiences. Also, the authors would like to thank the two anonymous reviewers and Dr. Paula Bush, Academic Associate at McGill University, for their constructive comments that helped to improve and clarify this manuscript.

Quan Nha Hong holds a Doctoral Fellowship Award from the Canadian Institutes of Health Research (CIHR) (#301011). Araceli Gonzalez-Reyes holds a Doctoral Fellowship Award from the Fonds de recherche du Québec – Santé (FRQS) (#28715). Pierre Pluye holds a Senior Investigator Award from the FRQS (#29308).

REFERENCES

1. Heyvaert M, Maes B, Onghena P. Mixed methods research synthesis: Definition, framework, and potential. Qual Quant. 2013;47(2),659-676. doi:10.1007/s11135-011-9538-6.
2. Pluye P, Hong QN. Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. Annu Rev Public Health. 2014;35,29-45. doi:10.1146/annurev-publhealth-032013-182440.
3. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Integrative Approaches to Qualitative and Quantitative Evidence. London: Health Development Agency; 2004.
4. Pluye P, Gagnon MP, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. Int J Nurs Stud. 2009;46(4),529-546. doi:10.1016/j.ijnurstu.2009.01.009.
5. Heyvaert M, Hannes K, Onghena P. Using Mixed Methods Research Synthesis for Literature Reviews: The Mixed Methods Research Synthesis Approach. Thousand Oaks, CA: SAGE Publications; 2016.
6. Whittemore R, Knafl K. The integrative review: Updated methodology. J Adv Nurs. 2005;52(5),546-553. doi:10.1111/j.1365-2648.2005.03621.x.
7. Hong QN, Pluye P, Bujold M, Wassef M. Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. Syst Rev. 2017;6(61),1-14. doi:10.1186/s13643-017-0454-2.
8. Harden A, Gough D. Quality and relevance appraisal. In Gough D, Oliver S, Thomas J, eds., An Introduction to Systematic Reviews London: SAGE Publications; 2012:153-178.
9. Burls A. What is Critical Appraisal? Newmarket, UK: Hayward Medical Communications; 2009.
10. Petticrew M, Roberts H. How to appraise the studies: An introduction to assessing study quality. In Petticrew M, Roberts H, eds., Systematic Reviews in the Social Sciences: A Practical Guide Padstow, UK: Wiley-Blackwell; 2006:125-163.
11. West SL, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ); 2002.
12. Bai A, Shukla VK, Bak G, Wells G. Quality Assessment Tools Project Report. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health; 2012.
13. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27),i-186. doi:10.3310/hta7270.
14. Heyvaert M, Hannes K, Maes B, Onghena P. Critical appraisal of mixed methods studies. J Mix Methods Res. 2013;7(4),302-327. doi:10.1177/1558689813479449.
15. Santiago-Delefosse M, Gavin A, Bruchez C, Roux P, Stephen S. Quality of qualitative research in the health sciences: Analysis of the common criteria present in 58 assessment guidelines by expert users. Soc Sci Med. 2016;148(1),142-151. doi:10.1016/j.socscimed.2015.11.007.
16. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol. 2004;4(22),1-11. doi:10.1186/1471-2288-4-22.

17. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. J Clin Epidemiol. 2011;64(1),79-89. doi:10.1016/j.jclinepi.2010.02.008.
18. Sanderson S, Tatt ID, Higgins JP, Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3),666-676. doi:10.1093/ije/dym018.
19. Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Syst Rev. 2017;6(1),204. doi:10.1186/s13643-017-0604-6.
20. Pace R, Pluye P, Bartlett G, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud. 2012;49(1),47-53. doi:10.1016/j.ijnurstu.2011.07.002.
21. Souto RQ, Khanassov V, Hong QN, Bush PL, Vedel I, Pluye P. Systematic mixed studies reviews: Updating results on the reliability and efficiency of the Mixed Methods Appraisal Tool. Int J Nurs Stud. 2015;52(1),500-501. doi:10.1016/j.ijnurstu.2014.08.010.
22. Pluye P, Robert E, Cargo M, et al. Proposal: A Mixed Methods Appraisal Tool for systematic mixed studies reviews. 2011. Retrieved November 15, 2013, from http://mixedmethodsappraisaltoolpublic.pbworks.com.
23. Sandelowski M. Focus on research methods - Whatever happened to qualitative description? Res Nurs Health. 2000;23(4),334-340. doi:10.1002/1098-240X(200008)23:4<334::AID-NUR9>3.0.CO;2-G.
24. Sandelowski M. What's in a name? Qualitative description revisited. Res Nurs Health. 2010;33(1),77-84. doi:10.1002/nur.20362.
25. Neergaard MA, Olesen F, Andersen RS, Sondergaard J. Qualitative description – The poor cousin of health research? BMC Med Res Methodol. 2009;9(52),1-5. doi:10.1186/1471-2288-9-52.
26. Porta MS, Greenland S, Hernán M, dos Santos Silva I, Last JM. A Dictionary of Epidemiology. New York: Oxford University Press; 2014.
27. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods. 2006;5(1),80-92. doi.org/10.1177/160940690600500107.
28. Nielsen J. Usability Engineering. San Francisco, CA: Morgan Kaufmann; 1994.
29. Khorsan R, Crawford C. How to assess the external validity and model validity of therapeutic trials: A conceptual approach to systematic review methodology. Evid Based Complement Alternat Med. 2014;2014(ID 694804),1-12. doi:10.1155/2014/694804.
30. Schmuckler MA. What is ecological validity? A dimensional analysis. Infancy. 2001;2(4),419-436. doi:10.1207/S15327078IN0204_02.
31. Hartling L, Bond K, Santaguida PL, Viswanathan M, Dryden DM. Testing a tool for the classification of study designs in systematic reviews of interventions and exposures showed moderate reliability and low accuracy. J Clin Epidemiol. 2011;64(8),861-871. doi:10.1016/j.jclinepi.2011.01.010.
32. Seo H-J, Kim SY, Lee YJ, et al. A newly developed tool for classifying study designs in systematic reviews of interventions and exposures showed substantial reliability and validity. J Clin Epidemiol. 2016;70(2),200-205. doi.org/10.1016/j.jclinepi.2015.09.013.

33. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR Network. BMC Med. 2010;8(24),1-6. doi:10.1186/1741-7015-8-24.
34. Mhaskar R, Djulbegovic B, Magazin A, Soares HP, Kumar A. Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. J Clin Epidemiol. 2012;65(6),602-609. doi:10.1016/j.jclinepi.2011.10.016.
35. Sandelowski M, Barroso J. Reading qualitative studies. Int J Qual Methods. 2002;1(1),74-108. doi:10.1177/160940690200100107.
36. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: Wiley Online Library; 2008.
37. Faggion CM, Jr. Risk of bias assessment should not go beyond reporting assessment. J Clin Epidemiol. 2016;72(4),126-127. doi:10.1016/j.jclinepi.2015.11.014.
38. Carroll C, Booth A, Lloyd-Jones M. Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. Qual Health Res. 2012;C22(10),1425-1434. doi:10.1177/1049732312452937.
39. Verhage A, Boels D. Critical appraisal of mixed methods research studies in a systematic scoping review on plural policing: Assessing the impact of excluding inadequately reported studies by means of a sensitivity analysis. Qual Quant. 2016;51(4),1449-1468. doi:10.1007/s11135-016-0345-y.
40. Haynes SN, Richard D, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychol Assess. 1995;7(3),238-247. doi.org/10.1037/1040-3590.7.3.238.
41. Francis JJ, Johnston M, Robertson C, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health. 2010;25(10),1229-1245. doi:10.1080/08870440903194015.
42. Olsson C, Ringner A, Borglin G. Including systematic reviews in PhD programmes and candidatures in nursing - 'Hobson's choice'? Nurse Educ Pract. 2014;14(2),102-105. doi:10.1016/j.nepr.2014.01.005.

Table 1. Profile of Participants

Characteristics		Number
Countries	Australia	4
	Canada	5
	Denmark	1
	England	3
	France	1
	The Netherlands	2
	New Zealand	1
	United States	3
Occupation	Doctoral candidate	9
	Post-doctoral fellow	4
	Professor/lecturer	5
	Research associate	1
	Librarian	1
Gender	Female	17
	Male	3
Setting	Public health agency	1
	University	19
Main research methodology used	Qualitative	5
	Quantitative	6
	Mixed methods	9
Main research areas of interest	Architecture	1
	Education	1
	Health sciences	14
	Information sciences	1
	Physical activity	2
	Psychology	1
Year of experience in research, mean (SD)		6.7 (3.7) years

Table 2. Themes Identified

Dimensions	Attributes	Themes*
Utility	Coverage	1 – Comprehensive tool 2 – Study designs that could not be appraised with the tool*
	Completeness	3 – Concerns about the completeness of the tool*
	Flexibility	4 – Need to adapt the tool to the topic of the review*
	Other utility	5 – Educational tool
Usability	Learnability	6 – Easy to use 7 – Improvement needed in the tutorial*
	Efficiency	8 – Short and quick
	Errors	9 – Items not clear or difficult to judge* 9.1 – Qualitative and mixed methods studies subject to interpretation 9.2 – Several concepts in one item 9.3 – Missing information in papers 9.4 – Unclear distinction between some items 10 – Difficulty classifying the studies*
	Satisfaction	11 – Accessible online 12 – Website not user-friendly* 13 – Missing rating sheet*

* Themes suggesting potential areas for improvement

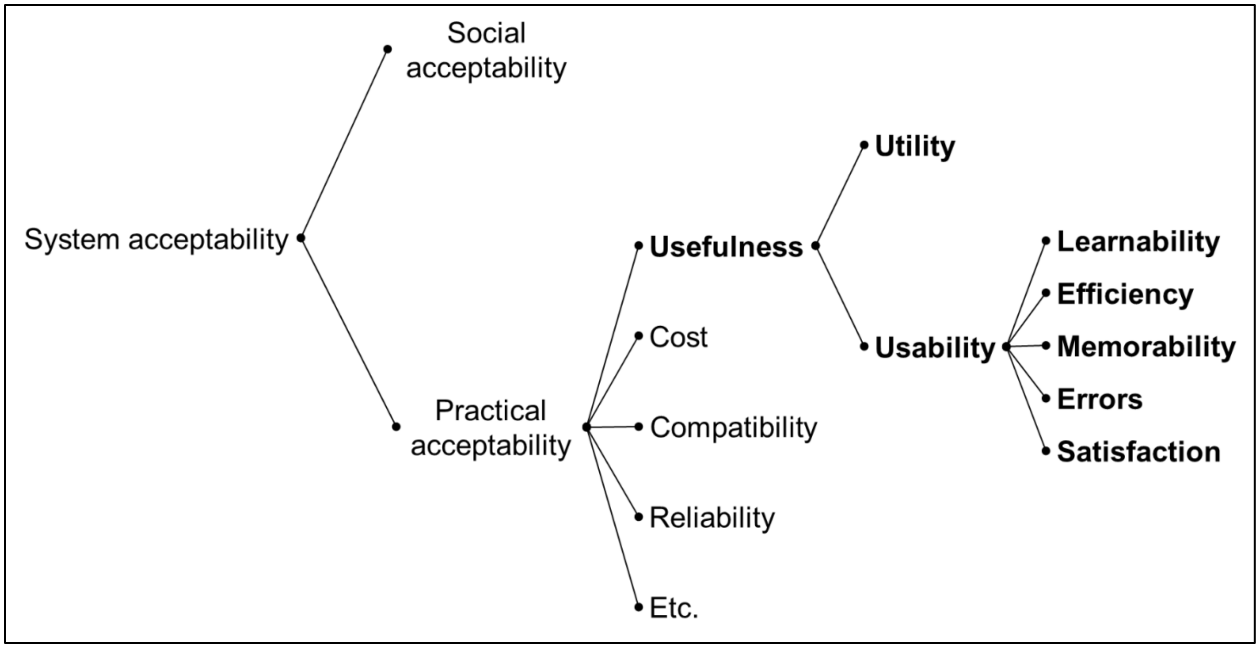


Figure 1. Framework on System Acceptability

Reprinted from Nielsen J. Usability Engineering. San Francisco, CA: Morgan Kaufmann; 1994, with permission from Elsevier.

CHAPTER 6. METHODS AND RESULTS OF PHASE 2 – MODIFIED E-DELPHI STUDY

This chapter provides detailed information on the quantitative phase of the project that aimed to identify relevant criteria for appraising the quality of qualitative, survey, and mixed methods studies. A modified e-Delphi study was conducted with methodological experts. Respectively, 73 and 56 experts participated in Round-one and Round-two of the modified e-Delphi. A consensus was reached for six qualitative criteria, eight survey criteria, and seven mixed methods criteria. The methods and results of this study are presented in the following manuscript (Paper #3 of this manuscript-based dissertation). This manuscript was submitted for publication in the Journal of Clinical Epidemiology. The invitation emails and questionnaires are presented in Appendix 8.

PAPER #3: Improving the Content Validity of the Mixed Methods Appraisal Tool (MMAT): A Modified e-Delphi Study

Submitted to the Journal of Clinical Epidemiology

Quan Nha HONG[a], Pierre PLUYE[a,*], Sergi FÀBREGUES[b], Gillian BARTLETT[a], Felicity BOARDMAN[c], Margaret CARGO[d], Pierre DAGENAIS[e], Marie-Pierre GAGNON[f], Frances GRIFFITHS[c], Belinda NICOLAU[g], Alicia O'CATHAIN[h], Marie-Claude ROUSSEAU[i], Isabelle VEDEL[a]

[a]Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, H3S 1Z1, Canada

[b]Department of Psychology and Education, Universitat Oberta de Catalunya, Rambla del Poblenou, 156, 08018, Barcelona, Spain

[c]Warwick Medical School – Division of Health Sciences, University of Warwick, Coventry, CV4 7AL, England

[d]Health Research Institute, University of Canberra, Canberra, ACT 2601, Australia

[e]Faculté de médecine et des sciences de la santé, Université de Sherbrooke, 3001, 12[e] Avenue Nord, Sherbrooke, QC, J1H 5N4, Canada

[f]Faculté des sciences infirmières, Université Laval, 1050, avenue de la Médecine, Québec, QC, G1V 0A6, Canada

[g]Faculty of Dentistry, Division of Oral Health and Society Research, McGill University, 2001 McGill College, suite 500, Montréal, QC, H3A 1G1, Canada

[h]Medical Care Research Unit, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, S1 4DA, England

[i]INRS–Institut Armand-Frappier Research Centre, 531, boulevard des Prairies, Laval, QC, H7V 1B7, Canada

***Corresponding Author:** Pierre Pluye, Department of Family Medicine, McGill University, 5858 Côte-des-Neiges, Suite 300, Montréal, QC, Canada, H3S 1Z1; Tel: 011-514-398-8483, Fax: 011-514-398-4202. Email address: pierre.pluye@.mcgill.ca

ABSTRACT

Objective: The Mixed Methods Appraisal Tool (MMAT) was developed for critically appraising different study designs. This study aimed to improve the content validity of three of the five categories of studies in the MMAT by identifying relevant methodological criteria for appraising the quality of qualitative, survey, and mixed methods research.

Study Design: First, we performed a literature review to identify critical appraisal tools and extract methodological criteria. Then, we conducted a two-round modified e-Delphi technique. We asked three method-specific panels of experts to rate the relevance of each criterion on a five-point Likert scale.

Results: A total of 383 criteria were extracted from 18 critical appraisal tools and a review on the quality of mixed methods research, and 60 were retained. In the first and second rounds of the e-Delphi, respectively 73 and 56 experts participated. Consensus was reached for six qualitative criteria, eight survey criteria and seven mixed methods criteria. These results led to modifications of 8 of the 11 MMAT (version 2011) criteria. Specifically, we reformulated two criteria, replaced four and removed two. Moreover, we added six new criteria.

Conclusion: The results of this study helped to revise the MMAT (version 2011) and improve its content validity.

Keywords: critical appraisal; quality assessment; Delphi technique; systematic reviews; qualitative research; survey research; mixed methods research.

WHAT IS NEW?

- The Mixed Methods Appraisal Tool (MMAT) is a critical appraisal tool for assessing the methodological quality of various study designs, including mixed methods studies.
- We further developed the MMAT based on experts' opinions to improve its content validity.
- This study adds to the literature on the quality of qualitative, survey, and mixed methods research, which is still sparse and lacking consensus.
- A new content validated version of the MMAT was developed and can be useful for the critical appraisal process in systematic reviews combining qualitative and quantitative evidence.

INTRODUCTION

Systematic reviews are considered among the best available sources of research evidence and are increasingly relied upon to inform decision-making [1]. The past 40 years have seen increasingly rapid methodological advances in the field of systematic reviews and research synthesis. Initial developments mainly focused on meta-analysis for addressing questions on the effectiveness of interventions and the emphasis was on randomized controlled trials (RCTs) [2, 3]. Since the early 2000s, researchers have shown a growing interest in systematic mixed studies reviews (SMSRs), which combine quantitative, qualitative and mixed methods studies to address other types of review questions concerned with, for instance, the acceptability of an intervention, participants' satisfaction, or barriers to implementation [4-6]. SMSRs are particularly useful for providing in-depth answers to complex clinical problems and practical concerns. Several challenges, however, are encountered in SMSRs due to the heterogeneity of included study designs. One of these challenges pertains to the critical appraisal of included studies.

Critical appraisal is an essential step in systematic reviews to ensure that their recommendations and conclusions reflect the quality of the evidence reviewed [7]. Since reviewers' judgment of a same study can vary greatly, critical appraisal tools have been developed to help reviewers appraise study quality in a more consistent, transparent, and reproducible way [8-10]. A critical appraisal tool (also named quality assessment tool or risk of bias tool) is a scale or checklist in which a list of criteria/domains is suggested to appraise the quality of a study. Extant reviews of critical appraisal tools have identified over 500 tools [11-31]. Most of these tools are specific to a particular research design or method. It is, thus, complex and time consuming to conduct SMSRs as reviewers must search for and learn how to use several different tools in order to complete the critical appraisal of the qualitative, quantitative, and mixed methods studies included in each review.

To address the challenge of critical appraisal in SMSRs, a unique tool for assessing the quality of different study designs was developed: the Mixed Methods Appraisal Tool (MMAT) [32]. The MMAT was developed in 2006 and has five sets of criteria for: (a) qualitative studies, (b) RCTs, (c) non-randomized studies, (d) quantitative descriptive studies, and (e) mixed methods studies. When appraising mixed methods studies, three sets of criteria are assessed: (a) the qualitative set, (b) a quantitative set (either RCT, non-randomized or quantitative descriptive

studies), and (c) the mixed methods set. In doing so, the MMAT acknowledges the methodological distinctive characteristics specific to each component used in mixed methods studies (i.e., qualitative, quantitative, and mixed methods) [33].

Previous studies on the interrater reliability of the MMAT reported that agreement scores ranged from poor to perfect [34, 35]. This suggests the need for clarification of some criteria in the MMAT, particularly those related to qualitative and non-randomized studies, for which lower agreement was observed. In addition, in interviews conducted with MMAT users to explore their views and experiences of the MMAT, concerns were raised about whether the tool included enough criteria to judge the quality of studies and criteria that were difficult to judge, in particular the criteria for qualitative and mixed methods studies [36]. This suggests a need to improve the content validity of the MMAT. The content validity of an assessment tool is defined as the degree to which criteria are relevant to, and representative of their targeted construct [37]. In the MMAT, the targeted construct is the methodological quality of studies appraised in SMSRs. In systematic reviews, the methodological quality of identified studies relates to how well a study was conducted and whether this was good enough for the results to be trustworthy [7, 30].

Currently, the existing literature on critical appraisal has focused, for the most part, on RCTs, cohort studies, and case-control studies, and several validated tools can be found for these study designs [12, 16, 20, 23, 29, 31]. This literature will inform the RCT and non-randomized criteria to revise in the MMAT. However, for other designs, such as qualitative, survey, and mixed methods, critical appraisal is more challenging since validated tools are rare and there is no clear consensus on how their quality assessment should be performed [19, 28, 38].

The objective of this study was to improve the content validity of the MMAT by identifying the most relevant methodological criteria for appraising the quality of qualitative, survey, and mixed methods studies. This study focused on these three categories of studies because of the scarcity of literature and lack of consensus.

METHODS

Two phases were conducted: (a) a literature review to identify existing criteria and (b) a modified e-Delphi technique to reach consensus on the criteria. The Delphi technique is used to reach consensus among a group of experts [39], and is particularly suitable to build consensus on issues that have limited or contradictory evidence [40]. It has been used for the development of other critical appraisal tools for different types of studies such as prognostic studies [41], case series studies [42, 43], cross-sectional studies [44], studies on measurement properties [45], and RCTs [46-48]. The Delphi technique is characterized by two or more rounds of questionnaires with controlled feedback, statistical group response, and anonymity [49]. There are different types of Delphi designs [50]. We used a modified e-Delphi, meaning that the Delphi was administered via an online web survey and used pre-selected methodological criteria in the first round.

Phase 1: Literature review

To identify methodological criteria, we performed a literature review of critical appraisal tools for qualitative, quantitative descriptive studies including surveys, and mixed methods research.

Sources

Two main literature sources were used. The first was a review of systematic reviews combining qualitative and quantitative evidence that was carried out in 2015 [4]. The second was 15 reviews on critical appraisal tools identified from citation tracking of critical appraisal tools found in the first source and from reviews known to the authors of this paper [11, 12, 15, 16, 19, 20, 22-24, 26, 28-30, 51, 52].

Selection Criteria

Critical appraisal tools assessing methodological quality were retained, while tools limited to the quality of reporting of studies, such as the COnsolidated criteria for REporting Qualitative research (COREQ), were excluded. We only retained appraisal tools that provided a clear description of their development with a group of experts, or that had been subject to

validity or reliability testing. We also considered popular tools in systematic reviews which were developed by leading international institutions such as tools from the Critical Appraisal Skills Programme (CASP) [53], the Joanna Briggs Institute (JBI) [54] and the National Institute for Health and Clinical Excellence (NICE) [55].

Identification of items

For each retained appraisal tool, all the criteria were extracted and entered in a spreadsheet. Two team members (QNH, PP) first screened the list to include methodological quality criteria. The following were excluded: criteria limited to the quality of reporting (e.g., the response rate is reported); generic criteria, i.e., criteria that could be applied to any study design (e.g., ethical issues are adequately considered); and criteria that were too specific to a topic (e.g., the ethnic composition of the population studied is recorded). Duplicate and similar criteria were removed. The preliminary list was sent to all members of the research team (authors of this paper) who had backgrounds in qualitative, epidemiology, and mixed methods research. They were asked to review the list, identify the criteria that were unclear, and suggest modifications, if necessary. They were also asked to suggest criteria they felt were missing from the list.

Phase 2: Two-round modified e-Delphi Study

Three method-specific panels of experts were asked to complete two rounds of e-Delphi questionnaires to identify the most relevant methodological criteria for critical appraisal. Relevance was defined as the appropriateness of the elements to the targeted construct [37]. In this study, the targeted construct was the methodological quality of studies.

Sample

For each panel, a purposeful sample of international experts was constituted. An expert is defined as an individual with knowledge and skills in a specific area [56]. For the purposes of this e-Delphi, the experts were researchers working in an academic or research institution with research interests in the methodological development of either qualitative, survey, or mixed methods research. To identify the experts, the lead author performed a search of books and methodological papers in Google Scholar, the McGill Library catalogue, and Amazon. Then, the biographies of publications' authors were consulted on the worldwide web to verify their

research design expertise (e.g., by checking their research interest and expertise, courses taught, and scientific publications). The lead author compiled the list of experts, categorized by research design and submitted it to the full research team asking members to add any missing experts. A total of 196 experts (i.e., potential participants) were retained.

Data Collection and Analysis

The questionnaires were put online using the LimeSurvey software hosted on the McGill University server. Pilot testing of the online questionnaires was conducted with one professor, two graduate students, and one research associate to obtain feedback regarding the clarity of the instructions, ease of completing the questionnaires, technical difficulties encountered, and to estimate the time needed to complete the task. Modifications to the questionnaires were made, accordingly.

In Round-one, the experts were asked to rate the relevance of each criterion. A 5-point Likert scale was used, ranging from 1=not at all relevant to 5=extremely relevant. Space was included at the end of the questionnaire for participants to provide comments and suggestions. A one-month turnaround time was given for panel members to complete the questionnaire. Based on the comments provided in Round-one, some criteria were modified and new criteria were added. A summary table of the results including group ratings and comments obtained in this round was prepared. This table was used to provide controlled feedback and statistical group response to participants, two important characteristics of the Delphi technique [57].

For Round-two, each participant was sent the summary table including a reminder of their responses and a new questionnaire to complete. The participants were asked to (re)rate all criteria using the same 5-point Likert scale. In addition, a 'cannot answer' response category was added (at the request of participants). Space was provided at the end of each question for comments and suggestions. The data of Round-two were summarized by calculating an agreement index. For each item, the number of experts rating criteria as very relevant or extremely relevant was divided by the total number of experts. For each item, we considered that consensus had been reached if the agreement index was 0.80 or more.

We used the agreement indexes and the comments from Round-two as well as the literature on quality of qualitative, survey, and mixed methods research to inform the revision of

the MMAT. Specifically, we verified if the criteria in the current version of the MMAT (version 2011) were among those with an agreement score ≥ 0.80. If not, we considered how they could be modified or replaced with new ones on similar concepts. Experts' comments were used to reformulate some criteria.

Ethics Statement

This project was approved by the Institutional Review Board of the Faculty of Medicine Research and Graduate Studies Offices from McGill University (ethics certificate number # A05-E26-15B). An electronic consent form was included in the questionnaire of Round-one. All experts provided informed consent to participate in this study and to be acknowledged in this paper. The responses were kept anonymous to the panel and no personally identifiable information was presented in the data file used for the analysis.

RESULTS

Phase 1 – Literature review

A total of 18 critical appraisal tools were retained: nine for qualitative studies [55, 58-65], seven for surveys, including cross-sectional and prevalence studies [66-72], and two for multiple study designs [32, 73]. Since only one tool with criteria specific to mixed methods studies was retained [32], the results of a recent literature review performed by a member of our research team on the quality of mixed methods studies were used [52]. In this latter review, the authors analyzed 64 papers on the quality of mixed methods studies and identified 46 criteria [52].

Overall, 383 criteria were extracted from the included literature (238 for qualitative studies, 99 for surveys, and 46 for mixed methods studies), of which 286 (75%) were removed because they were either duplicate, generic, topic related, or limited to reporting quality. The remaining 97 criteria were presented to the research team to assess their comprehensiveness and clarity; 38 were removed because they were not clear or similar to other criteria, and one criterion on the content validity for surveys was added. The 60 retained criteria included 20 for qualitative studies, 20 for quantitative descriptive studies, and 20 for mixed methods studies.

Phase 2 - Modified e-Delphi

Table 1 presents the number of participants in each round of the modified e-Delphi and for each of the three panels. A total of 73 experts from 11 different countries participated in Round-one: Australia (n=2), Belgium (n=3), Canada (n=11), England (n=9), Estonia (n=1), Germany (n=1), the Netherlands (n=4), Norway (n=1), Spain (n=1), Switzerland (n=1), and the United States of America (n=39).

Insert Table 1 around here

Based on the results of Round-one, of the initial 60 criteria, seven criteria were removed, 25 were reformulated, and eight new were added: two qualitative research criteria were removed, 15 modified and two added; three survey research criteria were removed, nine modified and three added; two mixed methods research criteria were removed, one modified and three added. Thus, the Round-two questionnaires included 62 criteria: 21 criteria for qualitative research, 20 criteria for survey, and 21 criteria for mixed methods research. The new questionnaires were sent to the 73 participants from Round-one, 56 of whom completed Round-two (Table 1). Consensus was reached for six qualitative research criteria, eight survey research criteria, and seven mixed methods research criteria. The results of Round-two are presented in tables 2 to 4.

Insert tables 2, 3 and 4 around here

Update of the Mixed Methods Appraisal Tool (MMAT)

In light of the results, 8 of the 11 criteria in the MMAT (version 2011) were modified: two were reformulated, four replaced, and two removed. Moreover, six new criteria were added. Table 5 presents the initial and new criteria.

Insert Table 5 around here

Qualitative studies criteria

Two criteria included in the MMAT (version 2011) were not considered among the most relevant criteria to appraise in this modified e-Delphi: criterion 1.3 on the influence of the context and criterion 1.4 about of researchers' reflexivity. Some experts considered that this

latter criterion might not always be reported given space limitations in journal publications. Inadequate reporting in qualitative studies is an important barrier to critical appraisal [74]. Based on these results, the research team decided to replace criteria 1.3 and 1.4 by three new criteria that reached high level of consensus in Round-two: one on the relevance of the qualitative approach to address the research question (Table 2, criterion #1), one on the coherence between data sources, collection, analysis and interpretation (Table 2, criterion #14), and one on the interpretation of results (Table 2, criterion #19).

Two criteria concerning the interpretation of results achieved a high level of consensus (Table 2, criteria #18 and 19). In Round-one they were combined, but experts requested they be separated because they address two different constructs (plausibility of finding vs. sufficient substantiation of findings). The latter criterion was retained for the new version of the MMAT because the agreement index was slightly higher than the former and plausibility might be more difficult to judge.

In addition, modifications were made to the first two qualitative criteria. The word 'interviews' was added to criterion 1.1, and the word 'relevant' was replaced by 'adequate'. Criterion 1.2 on analysis was reformulated and the word 'objective' was removed (see Table 5).

Survey criteria

Experts reached consensus on eight criteria. Some of these criteria addressed similar constructs and were thus combined. For example, to judge if a sample is representative of the target population (Table 5, criterion 4.2 in the MMAT), the target population needs to be clearly defined (Table 3, criterion #1) and the study participants and setting need to be described in detail (Table 3, criterion #2).

Concerning measurement bias, we included six criteria in the questionnaire but none achieved consensus. Several experts mentioned that the criteria on measurement could be useful in some circumstances, but not all. In the literature, measurement error is an important aspect to consider when conducting a survey [75]. Thus, no change was made to criterion 4.3 in the MMAT.

The original MMAT criterion on response rate was replaced with one on nonresponse bias (Table 5, criterion 4.4). The appropriateness of the response rate for surveys is often requested in appraisal tools. Some will use a cut-off (e.g., 60%). However, the experts mentioned that the cut-off value is arbitrary and that less emphasis should be put on a norm. Instead the focus should be placed on nonresponse bias. This concurs with studies reporting a weak association between response rate and nonresponse bias [76].

One criterion on the appropriateness of statistical analysis reached consensus for relevance by the experts (Table 3, criterion #14) and was added to the MMAT. Also, criterion #16 on confounding factors being accounted for in the analysis achieved consensus among the experts. This criterion was not added in the section quantitative descriptive studies of the MMAT since it is mainly applicable for analytical surveys. Analytical studies are addressed in another section of the MMAT.

Mixed methods studies criteria

All three MMAT criteria pertaining to mixed methods were replaced (Table 5). The first criterion on the relevance of research design (Table 5, criterion 5.1 in the MMAT) was replaced with a criterion on rationale (Table 4, criterion #2).

The second criterion on integration (Table 5, criterion 5.2 in the MMAT) was reformulated. Several criteria on integration reached consensus (Table 4, criteria #6, 9, 11). For the MMAT, we retained the criterion #6 (quantitative and qualitative components of the study are effectively integrated) because it was also mentioned in other studies as among the most prevalent criterion for assessing the quality of mixed methods studies [19, 77]. Also, some experts suggested avoiding the reference to qualitative and quantitative components in the formulation of the criteria. We replaced ‘quantitative and qualitative components’ by ‘different components’. In mixed methods studies, integration can be considered at different levels (e.g., philosophical, methodology, methods, data collection and analysis techniques) and one expert suggested being more precise on what is being integrated. In a review on mixed methods studies, Pluye et al. [78] identified nine strategies for integrating phases, results or data. Also, Fetters, Curry and Creswell [79] identified three integration levels (design, methods and

interpretation/reporting). Since integration can vary depending on how the study was conducted, no further information was added in the criterion to keep it comprehensive.

The third criterion on limitations in mixed methods studies (Table 5, criterion 5.3) was replaced with one on meta-inferences (Table 4, criterion #13) and one on divergences (Table 4, criterion #12). Several experts mentioned that the term 'meta-inference' was unclear. This criterion was reformulated as follows: The outputs of the integration of qualitative and quantitative components are adequately interpreted (Table 5).

One criterion was added about the trustworthiness of the qualitative and quantitative components (Table 4, criterion #18). Yet, the use of the term 'trustworthiness' did not reach consensus among the experts (some considered this term to be associated with qualitative research). Other terms were suggested such as legitimation, validity, credibility and integrity. To avoid entering into a semantic debate, we decided to reformulate this criterion based on the work of Fàbregues, Paré and Meneses [77]: The different components adhere to the quality criteria of each tradition of the methods involved. As mentioned earlier, the MMAT was conceived as a building block. Thus, the appraisal of the quality of each component in mixed methods studies is done using the criteria from the other sets in the MMAT.

DISCUSSION

We used a modified e-Delphi technique to identify the most relevant criteria for appraising the quality of qualitative, survey, and mixed methods studies. Consensus was reached for six criteria related to qualitative studies, eight for surveys, and seven for mixed methods studies. The results of this study informed the revision of the MMAT. In the previous version, the MMAT had four criteria for each category of studies. Based on our results, two of these criteria were reformulated, four were replaced, two were removed, and six were added. Thus, the revised version of the MMAT is composed of five criteria for each of the three above categories of studies.

A framework for developing assessment tools has been proposed in which three main stages are defined (initial steps, tool development and dissemination) [80]. This study is situated in the tool development stage by generating and seeking for consensus on criteria for three of the five study categories included in the MMAT (qualitative, survey, and mixed methods studies).

The literature on the quality of RCTs and non-randomized studies will inform the revision of the two other categories of studies in the MMAT. Once all the five sets of criteria are determined, we will update the MMAT's user manual, and pilot the tool [80].

There is a need to further content validate the criteria identified in this study, particularly for survey research. In this study, no criteria related to measurement and response rate biases in surveys made consensus (Table 3). This might be due to the fact that diverse sources can influence measurement errors (questionnaire, data collection method, interviewer, and respondent) [75] and can vary from one study to the other. As for measurement error, different indicators can be used to judge nonresponse bias such as identifying the reasons for nonresponse, determining if the respondents and non-respondents differ on the survey variable of interest, and weighting for nonresponse [75]. Although no specific criteria on measurement and response rate reached high level of consensus, the research team decided not to exclude these two biases from the MMAT since they are often mentioned in the survey research literature [75, 81, 82]. Further content validation work is needed to refine these criteria. Also, in the MMAT version 2011, surveys are included in the broad 'quantitative descriptive studies' category. We focused on survey research because they are often included in SMSRs, the existing tools have not been developed with experts, and surveys are among the most commonly used methods in mixed methods research [83]. Subsequent research should verify if the new criteria are applicable to other quantitative descriptive study designs.

Developing clear critical appraisal criteria is challenging. Experts provided several comments regarding the terms used in the criteria. For example, terms like 'relevant', 'adequate', and 'appropriate' were considered ambiguous. These terms are often used in critical appraisal tools of qualitative research [38]. Compared to reporting quality criteria, methodological quality criteria are more difficult to interpret because the reviewers need to judge whether the results of a study that are reported are trustworthy [84]. Also, criteria may be interpreted differently depending on the topic and context of the study.

The MMAT differs from other critical appraisal tools in several ways. To assess the quality of mixed methods studies, O'Cathain [33] suggested three different approaches: (a) generic research approach, (b) individual component approach, and (c) mixed methods approach. According to our review, the MMAT is the only tool that includes specific criteria for mixed

methods research [85]. With its five different sets of criteria, the MMAT uses a combination of individual component and mixed methods approaches. Other tools used in SMSRs approach critical appraisal differently. For example, Hawker et al. [86] and Crowe and Sheppard [15] use a generic approach by proposing one set of criteria that could be applied to any design. Others, such as those from the CASP [53], JBI [54] and NICE [55] propose one tool for each different study design (individual component approach). Also, some tools such as the QATSDD [73] and SPIDER [87] use a combination of generic and individual component approaches, with generic criteria applicable to several designs and specific criteria for qualitative and quantitative studies.

In addition, the MMAT is distinct from the other tools in that it focuses on methodological quality criteria and consists of a small number of items. Similar to other risk of bias tools [88], the MMAT focuses on the core criteria that may hinder the validity of the findings of a study. Some criteria (such as information on ethical considerations), though essential in a research process, may have less impact on the validity of a study compared to other methodological criteria (such as appropriate measurement).

Strengths and Limitations

Since 15 reviews analyzing more than 500 critical appraisal tools were found, we considered that an overview of these reviews was an efficient approach to meet our objectives. Thus, while it is possible that we did not identify all critical appraisal tools tested for validity or reliability, the pool of items we identified included over 75% generic, reporting quality, and duplicate criteria. This suggests that our sample included the main criteria.

The number of experts on the three panels in Round-two ranged from 15 to 21. There is no rule regarding the required sample size for a Delphi. Some authors suggest a panel of 8 to 12 participants, while others recommended 300 to 500 [39]. One important factor to take into consideration when determining the size is the composition of the sample (homogeneous or heterogeneous). Usually, a smaller sample, such as 10 to 15 participants, is considered sufficient for homogeneous samples [39]. Similarly, there is no clear recommendation regarding the number of experts needed for content validation. Lynn [89] suggested that five experts could be sufficient. Polit, Beck and Owen [90] recommended having 8 to 12 experts for the first round.

Given this, since our samples were relatively homogenous in terms of experts' methodological expertise, their sizes may be considered acceptable.

The list of potential experts was not exhaustive. Also, not all those who conduct systematic reviews are researchers with methodological expertise. Perhaps our study could have benefited from including such individuals in our panels of experts. For instance, the experience of health technology assessment practitioners or other health professionals with systematic reviews could have contributed to identifying relevant criteria to appraise. Future research and pilot testing of the MMAT could include this population.

The decision to use an agreement index threshold of 0.80 used in this study was arbitrary. There is no standard threshold for determining consensus in a Delphi study. Studies have used values varying from 0.50 to 0.80 [39]. In a previous study, it was found that criteria with an index of 0.78 or higher were indicative of good content validity [89]. Since the aim of this study was to identify core sets of criteria for validity content purpose, it was decided to use a high threshold.

Likert scales may have some limitations related to central tendency and desirability biases [91, 92]. To limit this bias, we calculated frequencies (instead of means) and considered two ratings (very relevant and extremely relevant) to compute the agreement index.

CONCLUSION

The MMAT can facilitate the critical appraisal process in SMSRs by providing, within a single tool, methodological quality criteria for different designs. This modified e-Delphi sought experts' consensus on the methodological quality criteria of qualitative, survey, and mixed methods research. The results led to replacing and clarifying the criteria of three of the five categories of studies in the MMAT, and improving its content validity. Additional validation research on the MMAT is still needed. In particular, the discriminatory validity of the MMAT still needs to be tested along with the interrater reliability.

ACKNOWLEDGEMENTS

The research team would like to acknowledge and sincerely thank all the participants for their contributions to the Delphi panel. Here are the names of those who accepted to be

acknowledged for their participation in this study: Lesley Andres (University of British Columbia, Canada); Theodore Bartholomew (Purdue University, United States); Pat Bazeley (Research Support/UNSW, Australia); Jelke Bethlehem (Leiden University, The Netherlands); Paul Biemer (RTI International, United States); Jaak Billiet (University of Leuven, Belgium); Felicity Bishop (University of Southampton, England); Jörg Blasius (University of Bonn, Germany); Hennie Boeije (Utrecht University, The Netherlands); Jonathan Burton (Understanding Society, England); Kathy Charmaz (Sonoma State University, United States); Benjamin Crabtree (The State University of New Jersey, United States); Elizabeth Creamer (Virginia Tech University, United States); Edith de Leeuw (Utrecht University, The Netherlands); Claire Durand (Université de Montréal, Canada); Joan Eakin (University of Toronto, Canada); Michèle Ernst Stähli (Université de Lausanne, Switzerland); Michael Fetters (University of Michigan Medical School, United States); Nigel Fielding (University of Surrey, England); Rory Fitzgerald (University of London, England); Floyd Fowler (University of Massachusetts, United States); Dawn Freshwater (University of Western Australia, Australia); Jennifer Greene (University of Illinois at Urbana-Champaign, United States); Christina Gringeri (University of Utah, United States); Greg Guest (FHI 360, United States); Timothy Guetterman (University of Michigan Medical School, United States); Muhammad Hadi (University of Leeds, England); Elizabeth Halcomb (University of Wollongong, United States); Carolyn Heinrich (Vanderbilt University, United States); Sharlene Hesse-Biber (Boston College, United States); Mieke Heyvaert (University of Leuven, Belgium); John Hitchcock (Indiana University Bloomington, United States); Nataliya Ivankova (University of Alabama at Birmingham, United States); Laura Johnson (Northern Illinois University, United States); Paul Lavrakas (University of Chicago, United States); Marilyn Lichtman (Virginia Tech University, United States); Geert Loosveldt (University of Leuven, Belgium); Peter Lynn (University of Essex, England); Mary Ellen Macdonald (McGill University, Canada); Claire Howell Major (University of Alabama, United States); Maria Mayan (University of Alberta, Canada); Sharan Merriam (University of Georgia, United States); José Molina-Azorín (University of Alicante, Spain); David Morgan (Portland State University, United States); Peter Nardi (Pitzer College, United States); Katrin Niglas (Tallinn University, Estonia); Karin Olson (University of Alberta, Canada); Antigoni Papadimitriou (Johns Hopkins University, United States); Michael Quinn Patton (Independent organizational development and program evaluation consultant, United States); Rogério Meireles

Pinto (Columbia University School of Social Work, United States); Vicki Plano Clark (University of Cincinnati, United States); David Plowright (University of Hull, England); Blake Poland (University of Toronto, Canada); Rodney Reynolds (California Lutheran University, United States); Gretchen B. Rossman (University of Massachusetts Amherst, United States); Erin Ruel (Georgia State University, United States); Michael Saini (University of Toronto, Canada); Johnny Saldaña (Arizona State University, United States); Joanna Sale (Li Ka Shing Knowledge Institute, Canada); Karen Schifferdecker (Dartmouth College, United States); David Silverman (University of London, England); Ineke Stoop (Netherlands Institute for Social Research, The Netherlands); Sally Thorne (University of British Columbia, Canada); Sarah Tracy (Arizona State University, United States); Frederick Wertz (Fordham University, United States).

Quan Nha Hong, OT, MSc, PhD candidate, holds a Doctoral Fellowship Award from the Canadian Institutes of Health Research (CIHR). Pierre Pluye, MD, PhD, Full Professor, holds a Senior Investigator Award from the Fonds de recherche du Québec – Santé (FRQS) and is the Director of the Methodological Development Platform of the Quebec-SPOR SUPPORT Unit, which is funded by the CIHR, the FRQS, and the Quebec Ministry of Health.

REFERENCES

[1] Bunn F, Trivedi D, Alderson P, Hamilton L, Martin A, Pinkney E, Iliffe S. The impact of Cochrane Reviews: A mixed-methods evaluation of outputs from Cochrane Review Groups supported by the National Institute for Health Research. Health Technol Assess 2015, 19(28):1-100. doi:10.1186/2046-4053-3-125.

[2] Glass GV. Primary, secondary, and meta-analysis of research. Educ Res 1976, 5(10):3-8. doi:10.3102/0013189X005010003.

[3] Cochrane AL. Effectiveness and efficiency: Random reflections on health services. London: Nuffield Provincial Hospitals Trust; 1972.

[4] Hong QN, Pluye P, Bujold M, Wassef M. Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. Syst Rev 2017, 6(61):1-14. doi:10.1186/s13643-017-0454-2.

[5] Petticrew M, Rehfuess E, Noyes J, Higgins JP, Mayhew A, Pantoja T, Shemilt I, Sowden A. Synthesizing evidence on complex interventions: How meta-analytical, qualitative, and mixed-method approaches can contribute. J Clin Epidemiol 2013, 66(11):1230-43. doi:10.1016/j.jclinepi.2013.06.005.

[6] Pluye P, Hong QN, Bush PL, Vedel I. Opening-up the definition of systematic literature review: The plurality of worldviews, methodologies and methods for reviews and syntheses. J Clin Epidemiol 2016, 73(5):2-5. doi:10.1016/j.jclinepi.2015.08.033.

[7] Higgins JP, Altman DG. Assessing risk of bias in included studies. In: Higgins JP, Green S, editors. Cochrane handbook for systematic reviews of interventions. Chichester, UK: The Cochrane Collaboration and John Wiley & Sons Ltd; 2008, p. 187-242.

[8] Wells K, Littell JH. Study quality assessment in systematic reviews of research on intervention effects. Res Soc Work Pract 2009, 19(1):52-62. doi:10.1177/1049731508317278.

[9] Wortman PM. Judging research quality. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russel Sage Foundation; 1994, p. 97-109.

[10] Petticrew M, Roberts H. How to appraise the studies: An introduction to assessing study quality. In: Petticrew M, Roberts H, editors. Systematic reviews in the social sciences: A practical guide. Padstow, UK: Wiley-Blackwell; 2006, p. 125-63.

[11] Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004, 4(22):1-11. doi:10.1186/1471-2288-4-22.

[12] Armijo-Olivo S, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: A systematic review. Phys Ther 2008, 88(2):156-75. doi:10.2522/ptj.20070147.

[13] Bai A, Shukla VK, Bak G, Wells G. Quality assessment tools project report. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health; 2012.

[14] Ciliska D, Thomas H, Buffett C. An introduction to evidence-informed public health and a compendium of critical appraisal tools for public health practice. Hamilton, ON: National Collaboration Center for Methods and Tools; 2012.

[15] Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. J Clin Epidemiol 2011, 64(1):79-89. doi:10.1016/j.jclinepi.2010.02.008.

[16] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. Health Technol Assess 2003, 7(27):i-186. doi:10.3310/hta7270.

[17] Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. GMS Health Technol Assess 2010, 6:Doc07 (20100614). doi:10.3205/hta000085.

[18] Dyrvig A-K, Kidholm K, Gerke O, Vondeling H. Checklists for external validity: A systematic review. J Eval Clin Pract 2014, 20(6):857-64. doi:10.1111/jep.12166.

[19] Heyvaert M, Hannes K, Maes B, Onghena P. Critical appraisal of mixed methods studies. J Mix Methods Res 2013, 7(4):302-27. doi:10.1177/1558689813479449.

[20] Jarde A, Losilla J-M, Vives J. Methodological quality assessment tools of non-experimental studies: A systematic review. An Psicol 2012, 28(2):617-28. doi:10.6018/analesps.28.2.148911.

[21] Lim SM, Shin ES, Lee SH, Seo KH, Jung YM, Jang JE. Tools for assessing quality and risk of bias by levels of evidence. J Korean Med Assoc 2011, 54(4):419-29. doi:10.5124/jkma.2011.54.4.419.

[22] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 1995, 16(1):62-73. doi:10.1016/0197-2456(94)00031-W.

[23] Sanderson S, Tatt ID, Higgins JP, Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. Int J Epidemiol 2007, 36(3):666-76. doi:10.1093/ije/dym018.

[24] Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. West J Nurs Res 2003, 25(2):223-37. doi:10.1177/0193945902250039.

[25] Seehra J, Pandis N, Koletsi D, Fleming PS. Use of quality assessment tools in systematic reviews was varied and inconsistent. J Clin Epidemiol 2016, 69:179-84.e5. doi:10.1016/j.jclinepi.2015.06.023.

[26] Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol 2010, 63(10):1061-70. doi:10.1016/j.jclinepi.2010.04.014.

[27] Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A *et al*. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ) Methods Guide for Comparative Effectiveness Reviews; 2012.

[28] Walsh D, Downe S. Appraising the quality of qualitative research. Midwifery 2006, 22(2):108-19. doi:10.1016/j.midw.2005.05.004.

[29] Wendt O, Miller B. Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. Educ Treat Children 2012, 35(2):235-68. doi:10.1353/etc.2012.0010.

[30] West SL, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, Lux L. Systems to rate the strength of scientific evidence. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ); 2002.

[31] Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, Niu Y, Du L. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: A systematic review. J Evid Based Med 2015, 8(1):2-10. doi:10.1111/jebm.12141.

[32] Pluye P, Robert E, Cargo M, Bartlett G, O'Cathain A, Griffiths F, Boardman F, Gagnon MP, Rousseau MC. Proposal: A Mixed Methods Appraisal Tool for systematic mixed studies reviews, http://mixedmethodsappraisaltoolpublic.pbworks.com; 2011 [accessed November 15, 2013].

[33] O'Cathain A. Assessing the quality of mixed methods research: Towards a comprehensive framework. In: Tashakkori A, Teddlie C, editors. Handbook of mixed methods in social and behavioral research. Thousand Oaks, CA: SAGE Publications; 2010, p. 531-55.

[34] Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, Seller R. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012, 49(1):47-53. doi:10.1016/j.ijnurstu.2011.07.002.

[35] Souto RQ, Khanassov V, Hong QN, Bush PL, Vedel I, Pluye P. Systematic mixed studies reviews: Updating results on the reliability and efficiency of the Mixed Methods Appraisal Tool. Int J Nurs Stud 2015, 52(1):500-1. doi:10.1016/j.ijnurstu.2014.08.010.

[36] Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). J Eval Clin Pract 2018, 24(3):459-67. doi:10.1111/jep.12884.

[37] Haynes SN, Richard D, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychol Assess 1995, 7(3):238-47. doi:10.1037/1040-3590.7.3.238.

[38] Santiago-Delefosse M, Gavin A, Bruchez C, Roux P, Stephen S. Quality of qualitative research in the health sciences: Analysis of the common criteria present in 58 assessment guidelines by expert users. Soc Sci Med 2016, 148(1):142-51. doi:10.1016/j.socscimed.2015.11.007.

[39] Keeney S, Hasson F, McKenna H. The Delphi technique in nursing and health research. Chichester, UK: Wiley Online Library; 2011.

[40] Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000, 32(4):1008-15.

[41] Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006, 144(6):427-37.

[42] Guo B, Moga C, Harstall C, Schopflocher D. A principal component analysis is conducted for a case series quality appraisal checklist. J Clin Epidemiol 2016, 69(1):199-207. doi:10.1016/j.jclinepi.2015.07.010.

[43] Yang AW, Li CG, Da Costa C, Allan G, Reece J, Xue CC. Assessing quality of case series studies: Development and validation of an instrument by herbal medicine CAM researchers. J Altern Complement Med 2009, 15(5):513-22. doi:10.1089/acm.2007.0806.

[44] Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016, 6(12):e011458. doi:10.1136/bmjopen-2016-011458.

[45] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. Qual Life Res 2010, 19(4):539-49. doi:10.1007/s11136-010-9606-8.
[46] Sindhu F, Carpenter L, Seers K. Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. J Adv Nurs 1997, 25(6):1262-8. doi:10.1046/j.1365-2648.1997.19970251262.x
[47] Yates SL, Morley S, Eccleston C, Williams ACdC. A scale for rating the quality of psychological trials for pain. Pain 2005, 117(3):314-25. doi:10.1016/j.pain.2005.06.018.
[48] Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, Knipschild PG. The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 1998, 51(12):1235-41. doi:10.1016/S0895-4356(98)00131-0.
[49] Vernon W. The Delphi technique: A review. Int J Ther Rehabil 2009, 16(2):69-76. doi:10.12968/ijtr.2009.16.2.38892.
[50] Hasson F, Keeney S. Enhancing rigour in the Delphi technique research. Technol Forecast Soc Change 2011, 78(9):1695-704. doi:10.1016/j.techfore.2011.04.005.
[51] Harder T, Takla A, Rehfuess E, Sanchez-Vivar A, Matysiak-Klose D, Eckmanns T, Krause G, de Carvalho Gomes H, Jansen A, Ellis S *et al*. Evidence-based decision-making in infectious diseases epidemiology, prevention and control: Matching research questions to study designs and quality appraisal tools. BMC Med Res Methodol 2014, 14(69):1-16. doi:10.1186/1471-2288-14-69.
[52] Fàbregues S, Molina-Azorín JF. Addressing quality in mixed methods research: A review and recommendations for a future agenda. Qual Quant 2017, 51(6):2847-63. doi:10.1007/s11135-016-0449-4.
[53] Critical Appraisal Skills Programme. CASP checklists, http://www.casp-uk.net/casp-tools-checklists; 2017 [accessed December 1, 2017].
[54] Joanna Briggs Institute. Critical appraisal tools, http://joannabriggs.org/research/critical-appraisal-tools.html; 2017 [accessed October 24, 2017].
[55] National Institute for Health and Clinical Excellence. The guidelines manual. London, UK: National Institute for Health and Clinical Excellence (NICE); 2009.
[56] Baker J, Lovell K, Harris N. How expert are the experts? An exploration of the concept of 'expert' within Delphi panel techniques. Nurse Res 2006, 14(1):59-70. doi:10.7748/nr2006.10.14.1.59.c6010.
[57] von der Gracht HA. Consensus measurement in Delphi studies: Review and implications for future quality assurance. Technol Forecast Soc Change 2012, 79(8):1525-36. doi:10.1016/j.techfore.2012.04.013.
[58] Blaxter M. Criteria for the evaluation of qualitative research papers. Med Soc News 1996, 22(1):68-71.
[59] Boeije HR, van Wesel F, Alisic E. Making a difference: Towards a method for weighing the evidence in a qualitative synthesis. J Eval Clin Pract 2011, 17(4):657-63. doi:10.1111/j.1365-2753.2011.01674.x.
[60] Critical Appraisal Skills Programme. 10 questions to help you make sense of qualitative research, http://docs.wixstatic.com/ugd/dded87_25658615020e427da194a325e7773d42.pdf; 2017 [accessed November 3, 2017].

[61] Sandelowski M, Barroso J. Reading qualitative studies. Int J Qual Methods 2002, 1(1):74-108. doi:10.1177/160940690200100107.
[62] Spencer L, Ritchie J, Lewis J, Dillon L. Quality in qualitative evaluation: A framework for assessing research evidence. London: Government Chief Social Researcher's Office; 2003.
[63] Vermeire E, Van Royen P, Griffiths F, Coenen S, Peremans L, Hendrickx K. The critical appraisal of focus group research articles. Eur J Gen Pract 2002, 8(3):104-8. doi:10.3109/13814780209160850.
[64] Reis S, Hermoni D, Van-Raalte R, Dahan R, Borkan JM. Aggregation of qualitative studies—From theory to practice: Patient priorities and family medicine/general practice evaluations. Patient Educ Couns 2007, 65(2):214-22. doi:10.1016/j.pec.2006.07.011.
[65] Joanna Briggs Institute. Critical appraisal checklist for qualitative research, http://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Qualitative_Research2017.pdf; 2017 [accessed October 24, 2017].
[66] Munn Z, Moola S, Riitano D, Lisy K. The development of a critical appraisal tool for use in systematic reviews: Addressing questions of prevalence. Int J Health Manag 2014, 3(3):123-8. doi:10.15171/ijhpm.2014.71.
[67] Al-Jader L, Newcombe R, Hayes S, Murray A, Layzell J, Harper P. Developing a quality scoring system for epidemiological surveys of genetic disorders. Clin Genet 2002, 62(3):230-4.
[68] Bishop FL, Prescott P, Chan YK, Saville J, von Elm E, Lewith GT. Prevalence of complementary medicine use in pediatric cancer: A systematic review. Pediatrics 2010, 125(4):768-76. doi:10.1542/peds.2009-1775.
[69] Giannakopoulos NN, Rammelsberg P, Eberhard L, Schmitter M. A new instrument for assessing the quality of studies on prevalence. Clin Oral Investig 2012, 16(3):781-8. doi:10.1007/s00784-011-0557-4.
[70] Hoy D, Brooks P, Woolf A, Blyth F, March L, Bain C, Baker P, Smith E, Buchbinder R. Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. J Clin Epidemiol 2012, 65(9):934-9. doi:10.1016/j.jclinepi.2011.11.014.
[71] Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, Janes G, Maglione M, Moher D, Nasser M *et al*. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: Pilot study of new checklists. J Clin Epidemiol 2011, 64(6):637-57. doi:10.1016/j.jclinepi.2010.08.006.
[72] Joanna Briggs Institute. Critical appraisal checklist for prevalence studies, http://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Prevalence_Studies2017.pdf; 2017 [accessed October 24, 2017].
[73] Sirriyeh R, Lawton R, Gardner P, Armitage G. Reviewing studies with diverse designs: The development and evaluation of a new tool. J Eval Clin Pract 2012, 18(4):746-52. doi:10.1111/j.1365-2753.2011.01662.x.
[74] Carroll C, Booth A. Quality assessment of qualitative evidence for systematic review and synthesis: Is it meaningful, and if so, how should it be performed? Res Synth, Methods 2015, 6(2):149-54. doi:10.1002/jrsm.1128.

[75] Federal Committee on Statistical Methodology. Measuring and reporting sources of error in surveys. Washington DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget; 2001.
[76] Groves RM, Peytcheva E. The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opin Q 2008, 72(2):167-89. doi:10.1093/poq/nfn011.
[77] Fàbregues S, Paré M-H, Meneses J. Operationalizing and conceptualizing quality in mixed methods research: A multiple case study of the disciplines of education, nursing, psychology, and sociology. J Mix Methods Res 2018, Advance online publication. doi:10.1177/1558689817751774.
[78] Pluye P, Garcia Bengoechea E, Granikov V, Kaur N, Tang DL. A world of possibilities in mixed methods: Review of the combinations of strategies used to integrate the phases, results, and qualitative and quantitative data. Int J Mult Res Approaches 2018, 10(1):41-56. doi:10.29034/ijmra.v10n1a3.
[79] Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs - Principles and practices. Health Serv Res 2013, 48(6pt2):2134-56. doi:10.1111/1475-6773.12117.
[80] Whiting P, Wolff R, Mallett S, Simera I, Savović J. A proposed framework for developing quality assessment tools. Syst Rev 2017, 6(204):1-9. doi:10.1186/s13643-017-0604-6.
[81] Davern M. Nonresponse rates are a problematic indicator of nonresponse bias in survey research. Health Serv Res 2013, 48(3):905-12. doi:10.1111/1475-6773.12070.
[82] Dillman DA, Phelps G, Tortora R, Swift K, Kohrell J, Berck J, Messer BL. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. Soc Sci Res 2009, 38(1):1-18. doi:10.1016/j.ssresearch.2008.03.007.
[83] Bryman A. Integrating quantitative and qualitative research: How is it done? Qual Res 2006, 6(1):97-113. doi:10.1177/1468794106058877.
[84] Carroll C, Booth A, Lloyd-Jones M. Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. Qual Health Res 2012, C22(10):1425-34. doi:10.1177/1049732312452937.
[85] Pluye P. Critical appraisal tools for assessing the methodological quality of qualitative, quantitative and mixed methods studies included in systematic mixed studies reviews. J Eval Clin Pract 2013, 19(4):722. doi:10.1111/jep.12017.
[86] Hawker S, Payne S, Kerr C, Hardey M, Powell J. Appraising the evidence: Reviewing disparate data systematically. Qual Health Res 2002, 12(9):1284-99. doi:10.1177/1049732302238251.
[87] Classen S, Winter S, Awadzi KD, Garvan CW, Lopez ED, Sundaram S. Psychometric testing of SPIDER: Data capture tool for systematic literature reviews. Am J Occup Ther 2008, 62(3):335-48.
[88] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JAC. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. Br Med J 2011, 343(d5928):1-9. doi:10.1136/bmj.d5928.
[89] Lynn MR. Determination and quantification of content validity. Nurs Res 1986, 35(6):382-6.

[90] Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007, 30(4):459-67. doi:10.1002/nur.20199.
[91] Garland R. The mid-point on a rating scale: Is it desirable. Mark Bull 1991, 2(1):66-70.
[92] Jamieson S. Likert scales: How to (ab)use them. Med Educ 2004, 38(12):1217-8. doi:10.1111/j.1365-2929.2004.02012.x.

Table 1. Number of Experts in Each Round of the Modified e-Delphi Study

Panel	Invitation	Round-one	Round-two
Qualitative	72	26	21
Survey	66	21	15
Mixed methods	58	26	20
Total	196	73	56

Table 2. Delphi Results with Experts in Qualitative Research (n=21)

Criteria	Agreement index
1. A qualitative approach is appropriate to answer the research question.	**1.00**
2. The methods were adapted to fit the context of the study.	0.71
3. The role(s) of researcher(s) are discussed in terms of their assumptions and position as insider/outsider relative to the phenomenon, participants, and/or setting.	0.67
4. The researcher's involvement in the data collection and analysis is appropriate for the method used.	0.57
5. The sampling strategy is appropriately justified.	0.71
6. The sample size is appropriate for the research design.	0.43
7. The sample represents the diversity of the people for whom the study is relevant.	0.24
8. The characteristics of the sample relevant to the interpretation of the findings are appropriately described.	0.71
9. The sites of recruitment are appropriate for addressing the purpose of the study.	0.38
10. The sources of qualitative data (such as archives, documents, participant observation, etc.) are appropriate to address the research question.	**0.86**
11. The qualitative data collection methods are most appropriate to address the research question.	0.76
12. The qualitative data analysis methods are appropriately addressed.	0.67
13. Appropriate explanation is given for how findings (such as themes, concepts, categories, etc.) were derived from the data.	**0.81**
14. There is coherence between qualitative data sources, collection, analysis and interpretation.	**0.81**
15. Strategies (such as prolonged engagement, peer review, etc.) are used to strengthen the findings.	0.71
16. Appropriate consideration is given to how findings relate to the context (such as the setting where the data were collected, etc.).	0.71
17. The influence of the researcher(s) on the data collection and analysis, results and interpretation is appropriately considered.	0.76
18. The interpretation of results is plausible.	**0.86**
19. The interpretation of results is sufficiently substantiated with data.	**0.90**
20. Any relevant epistemological or theoretical framework used is appropriately explained and justified.	0.62
21. The contextual relations between the researcher(s) and the participants (and/or materials) of research are appropriately addressed.	0.43

*Criteria in bold had an agreement index ≥0.80.

Table 3. Delphi Results with Experts in Survey Research (n=15)

Criteria*	Agreement index
1. The target population is clearly defined.	**1.00**
2. The study participants and the setting are described in detail.	**1.00**
3. The list from which the sample is drawn is appropriate for answering the research question.	**1.00**
4. The sampling strategy is relevant to address the research question.	**0.87**
5. The sample is representative of the target population for the main relevant variables.	**0.87**
6. The sample size is appropriate considering the population under study (such as population size, expected response rate, etc.).	0.53
7. The sample size is based on pre-study considerations of statistical power.	0.40
8. The same methods of data collection are used for all participants.	0.13
9. Standard instruments are used for the measurement of the variables.	0.33
10. The choice of variables is based on their content validity.	0.73
11. The survey instrument was pretested.	0.60
12. The survey instrument is reliable.	0.66
13. The survey instrument is valid.	0.66
14. The statistical analysis is appropriate to answer the research question.	**1.00**
15. The sampling bias is adequately addressed in the analysis.	**0.87**
16. Confounding factors are identified and accounted for in the analysis.	**0.80**
17. The response rate is acceptable (60% or above).	0.47
18. There is no significant difference in relevant sociodemographic characteristics between the respondents and the non-respondents.	0.40
19. Weighting for nonresponse is carried out.	0.60
20. A clear justification for using survey method is provided.	0.46

*Criteria in bold had an agreement index ≥0.80.

Table 4. Delphi Results with Experts in Mixed Methods Research (n=20)

Criteria*	Agreement index
1. A mixed methods research question (or purpose statement) is formulated.	0.60
2. A clear rationale is provided for using a mixed methods design to address the research problem and questions.	**0.95**
3. Key literature on mixed methods is reviewed in support of the mixed methods approach chosen by the authors.	0.20
4. The mixed methods design is consistent with the epistemological assumptions of the study.	0.30
5. Methods were selected to minimize shared bias.	0.25
6. Quantitative and qualitative components of the study are effectively integrated.	**0.85**
7. The type of integration of the quantitative and qualitative components matches the mixed methods design	0.70
8. The epistemological, ontological and teleological stances of the researcher that underlie the quantitative and quantitative approaches are successfully combined	0.10
9. Strategies for integrating phases, results and/or data are adequately performed.	**0.90**
10. Methods are implemented in a way that remains true to the mixed methods design.	0.70
11. The qualitative and quantitative components are linked in a cohesive and logical manner.	**0.85**
12. Divergences and inconsistencies between quantitative and qualitative results are adequately addressed.	**0.90**
13. Inferences derived from the quantitative and qualitative results are adequately incorporated in the meta-inferences regarding the entire study.	**0.90**
14. Meta-inferences regarding the entire study are consistent with the rationale given for using a mixed methods design.	0.50
15. The study contributes to advancing the field of mixed methods research.	0.10
16. The added value gained from using a mixed methods design in this study is described.	0.50
17. The strengths and weaknesses of methods optimize the breadth and depth of the study.	0.30
18. Threats to the trustworthiness of quantitative, qualitative and mixed methods are identified and adequately addressed.	**0.80**
19. Rigorous procedures for data collection and analysis are used in quantitative and qualitative components.	0.75
20. The study purposefully seeks out diverse perspectives (interpretive comprehension).	0.35
21. The mixed methods study generated findings and insights that would not have been possible with a mono-method study.	0.55

*Criteria in bold had an agreement index ≥0.80.

Table 5. Modifications of Three of the Five Categories of Studies of the Mixed Methods Appraisal Tool (MMAT)

Study category*	Criteria from the MMAT (version 2011)	Modifications
Qualitative research	1.1 Are the sources of qualitative data (archives, documents, informants, observations) relevant to address the research question (objective)?	Reformulate: Are the sources of qualitative data sources (such as archives, documents, participant observation, interviews, etc.) adequate to address the research question?
	1.2 Is the process for analyzing qualitative data relevant to address the research question (objective)?	Reformulate: Was adequate explanation is given for how findings (such as themes, concepts, categories, etc.) were derived from the data?
	1.3 Is appropriate consideration given to how findings relate to the context, e.g., the setting, in which the data were collected?	Remove
	1.4 Is appropriate consideration given to how findings relate to researchers' influence, e.g., through their interactions with participants?	Remove
		New: Is the qualitative approach appropriate to answer the research question?
		New: Is there coherence between qualitative data sources, collection, analysis and interpretation?
		New: Is the interpretation of results sufficiently substantiated with data?
Survey research	4.1 Is the sampling strategy relevant to address the quantitative research question (quantitative aspect of the mixed methods question)?	No change
	4.2 Is the sample representative of the population understudy?	No change
	4.3 Are measurements appropriate (clear origin, or validity known, or standard instrument)?	No change
	4.4 Is there an acceptable response rate (60% or above)?	Replacement: Is the risk of nonresponse bias low?
		New: Is the statistical analysis appropriate to answer the research question?

Mixed methods research	5.1 Is the mixed methods research design relevant to address the qualitative and quantitative research questions (or objectives), or the qualitative and quantitative aspects of the mixed methods question (or objective)?	Replacement: Is a clear rationale provided for using a mixed methods design to address the research problem and questions?
	5.2 Is the integration of qualitative and quantitative data (or results*) relevant to address the research question (objective)?	Replacement: Are the different components of the study effectively integrated to answer the research question?
	5.3 Is appropriate consideration given to the limitations associated with this integration, e.g., the divergence of qualitative and quantitative data (or results*) in a triangulation design?	Replacement: Are the divergences and inconsistencies between quantitative and qualitative results adequately addressed?
		New: Are the outputs of the integration of qualitative and quantitative components adequately interpreted?
		New: Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

*The two other categories of studies of the MMAT are randomized controlled trial and non-randomized studies.

CHAPTER 7. DISCUSSION

This chapter discusses the main findings of this project and presents a revised version of the MMAT. Then, the MMAT is compared with other existing tools. The last part of this chapter addresses the limitations, strengths and contributions of this project.

7.1 Discussion of Main Results

Two phases were performed in this project: phase 1 to identify areas for improvement needed in the MMAT and phase 2 to identify criteria that need to be modified, removed or added in the MMAT. The results of each phase will be discussed separately in the following.

7.1.1 Results of phase 1

Phase 1 consisted in a qualitative descriptive study in which 20 semi-structured interviews were conducted with MMAT users. The results of this phase were helpful to identify changes to be made in the MMAT as well as things to keep unchanged. Among the 13 themes identified in phase 1 (Table 2, Paper #2, Chapter 5), eight were judged problematic by the team and required improvement in the MMAT. The five other themes (comprehensive, easy to use, short and quick, educational tool, and accessible online) were aspects that were appreciated by the MMAT users and should be maintained. These results were discussed among the team members during a two-hour meeting.

One first point of discussion was related to the criteria on qualitative and mixed methods studies that were considered by the MMAT users more difficult to judge, more subject to interpretation than the quantitative criteria, and less precise (theme 9.1) as well as concerns about completeness of the tool (theme 3). Some suggestions were made by the team members to modify the criteria. One was to focus on the concept of ‘coherence’ between the different components of the qualitative approach as seen in the JBI tool (Joanna Briggs Institute, 2017c). The JBI tool looks at if there is congruity between the analytic method, the method of data collection, and the theoretical perspective within the qualitative approach that is applied. Another suggestion was to add the concept of ‘rigor’. Rigor would involve a clear presentation of all of the themes and give an explanation of the full picture (rather than focussing on the most frequent

themes). Also, another criterion that was suggested is how close the interpretation is to the data. The team members suggested investigating the qualitative and mixed methods criteria with a group of experts.

The team members suggested some changes in the tutorial (theme 7). First, there is a need to provide an explanation of why there is only one section for qualitative studies as opposed to three for quantitative studies. Also, more information is needed to specify how to use the MMAT. For example, if only qualitative studies or qualitative part of mixed methods research is retained, only the criteria of the qualitative component of the MMAT need to be appraised. Second, some examples provided in the criteria of the MMAT can be confusing. It was suggested to replace the notion of examples by 'hint' as seen in the CASP tools (Critical Appraisal Skills Programme, 2017a). Some members mentioned that care should be put to avoid giving too many explanations and examples or too definite guidance since users will stick with the examples and not look beyond that. Also, it is not possible to provide all types of examples of inappropriateness. Thus, it would be better to remove the examples from the criteria and add more explanation (hint) in the tutorial. Third, a suggestion was made to tailor the tutorial to the users' needs. For example, there could be two versions of the tutorial, one basic and a 'MMAT Plus' that will provide more explanations and refer to reference books and other resources for further information. The idea is to keep the tool short as appreciated by MMAT users (theme 8); not increasing the volume of information in the manual if the users do not need it.

Following the idea of developing a 'MMAT Plus', it was suggested to add specific criteria for common qualitative research approaches. For example, one specific feature in phenomenology is that researchers maintain a close connection/relationship between what was said and what was understood; in grounded theory, saturation and link with the literature are important for the development of a theory; in case study, patterns should be identified; and in ethnography, an in-depth collection/analysis of observation and participation are typical features. In the eventuality of developing a 'MMAT Plus', the specific features to each approach will need to be further explored based on the literature and experts consensus. This could address theme 4 on the need to adapt the tool to the topic of the review.

Comments were also provided by the team members regarding the criteria on mixed methods studies in the MMAT. It was recommended to add an explanation of what is mixed

methods because there exist different definitions and there is a need for common understanding (Johnson, Onwuegbuzie, & Turner, 2007). A preliminary question could be: Is it truly a mixed methods research? Another recommendation was to add the integration of phases in criterion 5.2 of the MMAT to be in line with recent evidence (Pluye et al., 2018). This criterion could be reformulated as: Is the integration of qualitative and quantitative phases, data, and results relevant to address the research questions?

A second point of discussion concerned the criteria that were often rated 'can't tell' (theme 9.3). The members mentioned that there can be two different meanings of 'can't tell' that needs to be clarified in the tutorial: it could mean that the topic is discussed but still not clear (no full understanding) vs. the authors did not talk about it (missing information). Currently, for criteria rated 'can't tell', it is advised to search for companion papers and contact the authors for more information. However, this latter recommendation is not often performed. The discussion about missing information also led to question whether the MMAT should be limited to criteria that are usually reported in papers. Opinion diverged among the team members. On the one hand, criteria should reflect what people should report and not what people are reporting. Thus, several did not agree to remove criteria that are not reported. On the other hand, in terms of content validity, criteria have to be relevant and representative (Haynes et al., 1995). The rule of thumb is that an item that is never checked might not be relevant and thus should be removed because it is useless.

Considering the team discussion, it was decided not to remove criteria on the basis that they are not well reported. Instead, it was suggested to provide more hints (indicators) to help the reviewers appraise the criteria. For example, one of the criteria that the MMAT users often mentioned rating 'can't tell' is criterion 1.4 on reflexivity. The team members proposed to reword this criterion (e.g., add the term 'reflexivity'). Because most papers do not report it, the best that can be done is to pick up really problematic papers showing an obvious lack of evidence of reflexivity that affect the research credibility and provide them as examples. Lack of reflexivity might be easier to identify. The rating of this criterion could be 0 when there is a problem with reflexivity and 1 when there is no problem with reflexivity. Another way of dealing with this could be working with the concept of 'coherence' (ensuring that the question, overall framework and method are all lined up). If there is no coherence, it may be a sign of no

reflexivity.

A third point discussed was related to the five categories of designs of the MMAT. Some MMAT users mentioned they have difficulty choosing a category of criteria to appraise, especially between NRS and quantitative descriptive studies (theme 10). It was suggested to add definitions of the five categories of designs included in the MMAT because people tend to label their studies in a certain way that does not always match the description. The important thing is not to find keywords that the authors have provided, but appraise the design and evaluate what type of design it seems to be based on the description provided. For example, a same design (e.g., cross-sectional) can refer to both analytical and descriptive studies. Also, the word 'survey' can be added in the quantitative descriptive category of the tutorial. Another suggestion was to change the labelling of the categories in the MMAT and orient it toward the research questions such as studies on the effectiveness of interventions, and studies describing a phenomenon. Moreover, the team discussed the structure of the categories of the MMAT. Instead of having RCT, NRS, and quantitative descriptive studies, it was proposed to restructure the categories of the MMAT as follows: 1-Qualitative studies, 2-Intervention studies, 3-Observational studies, 4-Descriptive studies, 5-Mixed methods studies. Intervention studies can be different from observational studies because there is an intervention to judge. Also, intervention studies can apply different designs including RCT and NRS. Several typologies of designs have made the distinction between observational and intervention studies (or experimental studies) (Carini et al., 2009; Grimes & Schulz, 2002b; National Institute for Health Care Excellence, 2012).

The MMAT was initially developed to cover the most common study designs included in SMSRs. Yet, users mentioned that some studies did not fit well with the categories of designs proposed in the MMAT (theme 2). The team discussed two options: referring to other existing tools or developing our own criteria with a team of experts. Currently, several tools on specific designs can be identified in the literature such as tools for economic studies (Adarkwah, van Gils, Hiligsmann, & Evers, 2016), diagnostic accuracy (Whiting et al., 2011), and measurement studies (Mokkink et al., 2010). Several of these tools followed a sound development process and were tested for validity and reliability. Questions were raised on the relevance of adding new criteria and which categories of designs should be added in the MMAT. For example, the team questioned whether specific criteria should be added for new trial designs such as pragmatic

trials and step-wedged design. Because of the large variety of study designs, there is a need to further investigate which categories of studies are missing in the MMAT based on their frequency of use. Thus, for now, no additional of new categories of designs is planned in the MMAT. It was suggested to add a note in the tutorial of the MMAT about this limit as well as references to other existing tools as needed.

A last point of discussion concerned the two screening questions added at the beginning of the tool for excluding papers that are not empirical studies. Since the MMAT focuses on methodological quality, papers that are not empirical studies cannot be appraised with this tool. Some MMAT users suggested removing the screening questions or integrating them with the criteria. There is a need to provide more explanation on the rationale of these two screening questions to make it clear from the beginning that they are used to rule out non-empirical papers. In the tutorial, a note can be added on the fact that these screening questions can be used during the selection step. It was suggested to leave these two questions as screening questions. Also, suggestions were made to reword the screening questions. The first screening question could be simplified: Are there clear research questions or objectives? In the second screening question, the example could be removed.

In summary, several themes identified in phase 1 of the project could be addressed by clarifying the explanations in the tutorial, reformulating some criteria, removing the examples in the criteria, providing hints for each item, and adding more information on how to use the MMAT. However, themes related to the criteria, such as criteria difficult to judge and completeness, need further investigations.

7.1.2 Results of phase 2

The phase 2 of the project focused specifically on identifying relevant criteria that could be included in the revision of the MMAT. The MMAT uses a threats-to-validity approach, i.e., focuses on specific features of a study design to judge the trustworthiness of a study. Threats to validity, also named bias, invalidity, exact bias and deviations from the truth, are systematic errors that can lead to distortion of study findings in either positive or negative direction (Delgado-Rodríguez & Llorca, 2004; Suzuki, Tsuda, Mitsuhashi, Mansournia, & Yamamoto, 2016). In the literature, one important point of debate is about what type of validity should be

appraised (Wells & Littell, 2009). Bryant and Wortman (1984) proposed to appraise four types of validity: internal, external, statistical conclusion and construct validity. They suggest using the construct and external validity to decide whether a study is relevant to the review question and the statistical conclusion and internal validity to judge the acceptability of a study. More than 20 tools appraising the external validity of studies can be found in the literature (Dyrvig, Kidholm, Gerke, & Vondeling, 2014). Yet, in more recent CATs, the focus is mainly put on the internal validity to assess how well a study was conducted to minimize bias (Bai et al., 2012). This has led to the development of a series of risk of bias tools such as the Cochrane RoB (Higgins et al., 2016), Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I) (Sterne et al., 2016), Risk Of Bias In Non-randomized Studies - of Exposures (ROBINS-E) (Morgan et al., 2017), Risk of Bias Assessment tool for Non-randomized Studies (RoBANS) (Kim et al., 2013), and Risk of Bias in N-of-1 Trials (RoBINT) (Tate et al., 2013). In line with recent development, the MMAT put emphasis on methodological quality, and more specifically, on the threats to validity of studies (biases). However, since qualitative and mixed methods studies are included in the MMAT, the term 'methodological quality' was preferred in this dissertation. Internal validity (bias) is a term mainly associated with quantitative research. In qualitative research, a similar concept would be 'credibility' (see Table 1, Paper #1, Chapter 2).

To identify the criteria that can influence the treats to validity of studies, two methods were used depending on the categories of designs: a mapping of criteria of CATs on RCT and NRS, and a modified e-Delphi technique for criteria on qualitative, survey, and mixed methods studies. The findings obtained from these methods are discussed in the following.

7.1.2.1 Criteria for randomized controlled trials and non-randomized studies

Several biases can be identified in the literature on RCT and NRS. For example, Sackett (1979) catalogued 35 biases, and Delgado-Rodríguez and Llorca (2004) listed 74 biases. Based on the literature of CATs on RCT and NRS, it is possible to identify seven core categories of biases that are usually considered during quality appraisal (Armijo-Olivo, Fuentes, Ospina, Saltaji, & Hartling, 2013; Higgins et al., 2016; Jarde, Losilla, Vives, & Rodrigo, 2013; Morgan et al., 2017; Sterne et al., 2016; Viswanathan & Berkman, 2012; Zaza, Wright-De Agüero, Briss, Truman, & Hopkins, 2000). They are related to the selection of participants, intervention and

exposure, measurement of exposures and outcomes, attrition, confounding, analysis of data, and reporting (Table 6).

Table 6. Definitions of Threats to Validity (Biases) in Randomized Controlled Trials and Non-Randomized Studies

Bias linked with	Definition	Terms used
Selection	Bias due to systematic differences between the groups that are compared. Bias due to individual being more likely to be selected than others.	• Selection bias • Bias in selection of participants into the study • Bias arising from the randomization process • Sampling bias
Intervention/ Exposure	Bias due to systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest.	• Performance bias • Bias due to deviations from intended interventions • Bias due to departure from intended exposures • Bias introduced by failure to maintain integrity of the intervention
Measurement	Bias due to systematic differences caused by variables being inaccurately measured or classified.	• Measurement bias • Detection bias • Information bias • Recall bias • Observation bias • Misclassification bias • Bias in measurement classification of interventions • Bias in classification of exposures • Bias in measurement of outcomes
Attrition	Bias due to systematic differences between groups in nonresponses, withdrawals and exclusions of participants.	• Attrition bias • Bias due to missing data
Confounding*	Bias in the distortion in the interpretation of findings due to one or more prognostic variables (factors that predict the outcome of interest) that also predicts the intervention received/exposure at baseline.	• Confounding bias • Bias due to confounding
Analysis	Bias due to errors in the analytical procedures used.	• Analytic bias • Statistical bias

Reporting	Bias due to systematic differences between reported and unreported findings.	• Reporting bias • Bias in the selection of the reported result

* In some reference books, confounding bias is considered as a selection bias (Viswanathan & Berkman, 2012).

To inform the revision of the two sets of the MMAT criteria not addressed in the e-Delphi study (i.e., RCT and NRS), a mapping of the criteria included in CATs was performed. Among the 52 CATs identified in the literature review (Chapter 3), 33 were on RCTs, cohort studies, case-control studies, intervention studies, and observational studies. All the criteria listed in these CATs were extracted and entered in an Excel spreadsheet. Then, the criteria were grouped into the seven categories of biases identified in the literature (Table 6): (a) sample and sampling methods/allocation, (b) intervention/exposure, (c) measurement, (d) attrition/follow-up, (e) analysis, (f) confounding, and (g) reporting. In addition, all the criteria on blinding were grouped together. Criteria related to other categories such as ethics (e.g., informed consent obtained), conflict of interest (e.g., source of funding stated), objectives (e.g., clear aim and hypothesis), design (e.g., appropriate study design), results (e.g., findings clearly described), and conclusion (e.g., findings support the conclusions) were not included in the mapping because there are not related to specific threats to validity.

To identify the most relevant criteria, the number of CATs that used each item was counted and their frequency was calculated, i.e., the number of CATs that included an item divided by the total number of CATs. The criteria with the highest frequency were used to inform the revision of the MMAT. For RCT, the results of the mapping were compared to those of a review on CATs used in general health research and physical therapy (Armijo-Olivo et al., 2013). They analyzed 26 CATs on RCT, of which 17 were included in this review. Table 7 presents the criteria with the highest frequencies in each category of bias. Detailed results of the mapping for RCT and NRS are presented in Appendix 9. In Appendix 9, the CATs specific to only RCT or only NRS were grouped together as well as those that included criteria for both RCT and NRS. This was done to check if the criteria were specific to RCT or NRS. For example, in RCT, a criterion on confounding was found in seven CATs but only one CAT was specific to RCT (the six others include criteria for both RCT and NRS). This might indicate that confounding bias is less important in RCT since randomization can minimize this bias.

Table 7. Most Frequent Criteria Used in Critical Appraisal Tools of Randomized Controlled Trials and Non-Randomized Studies

Bias	Randomized controlled trials (n=23)	Non-randomized studies (n=19)
Selection bias	▪ Random allocation (19, 83%) ▪ Baseline equivalence (15, 65%) ▪ Clear selection criteria (15, 65%) ▪ Allocation concealment (11, 48%)	▪ Clear selection criteria (9, 47%) ▪ Subjects representative of the target population (8, 42%) ▪ Random allocation (8, 42%) ▪ Baseline equivalence (7, 37%)
Performance bias	▪ Blinding of subjects (17, 74%) ▪ Intervention well described (12, 52%) ▪ Blinding carer (11, 48%) ▪ Adherence with intervention (11, 48%) ▪ Cointervention avoided or similar across groups (10, 43%)	▪ Blinding of subjects (9, 47%) ▪ Intervention administered as designed (5, 26%) ▪ Adherence with intervention (5, 26%)
Measurement bias (misclassification bias)		▪ Clear definition or description of intervention/exposure (10, 53%) ▪ Intervention/exposure assessed using valid and reliable measures (6, 32%)
Measurement bias (measurement of outcome bias)	▪ Blinding of assessor/outcome measure not influenced by knowledge of intervention received (19, 83%) ▪ Valid/reliable/standard measures (12, 52%) ▪ Clear description/justification of outcomes measured (7, 30%) ▪ Relevant outcomes (7, 30%)	▪ Blinding of assessor/outcome measure not influenced by knowledge of intervention/exposure received (14, 74%) ▪ Valid/reliable/standard measures (12, 63%) ▪ Clear description of the outcome measured (8, 42%)
Attrition bias	▪ Number and reasons of subjects lost to follow-up (13, 57%) ▪ Complete data/low lost to follow-up (10, 43%)	▪ Complete data/low lost to follow-up (11, 58%) ▪ Number and reasons of subjects lost to follow-up (8, 42%) ▪ Dropout rates/reasons similar between groups (7, 37%)
Confounding bias		▪ Confounders accounted for in analysis (13, 68%) ▪ Confounders accounted for in design (8, 42%)

Analytical bias	▪ Appropriate statistical analysis (14, 61%) ▪ Intention-to-treat (13, 57%) ▪ Sample size justification (11, 48%) ▪ Report point measures and measures of variability (10, 43%)	▪ Appropriate statistical analysis (11, 58%) ▪ Sample size justification (8, 42%) ▪ Report point measures and measures of variability (6, 32%)
Reporting bias	▪ Selective reporting or complete reporting of findings (3, 13%)	▪ Selective reporting or complete reporting of findings (6, 32%)

The results of the mapping were used to inform the revision of MMAT for RCT and NRS. The suggestion made during the team meeting of restructuring the categories of the MMAT to 'Intervention studies' and 'Observational studies' instead of RCT and NRS was not retained after analyzing the literature on CATs. Three main reasons can explain that the current structure of the MMAT remained unchanged. First, there are two categories of bias (misclassification and confounding biases) in NRS that are not found in RCT because randomization minimizes these biases. Second, in RCT, more importance is put on the proper execution of the sampling technique (i.e., randomization). For example, in the new version of the Cochrane RoB (Higgins et al., 2016), the selection bias category was renamed 'bias arising from the randomization process' (Higgins et al., 2016). Third, the biases between intervention studies that are not RCT and observational studies are comparable. For example, the Cochrane team developed a risk of bias tool for non-randomized intervention studies (ROBINS-I) (Sterne et al., 2016) and for non-randomized exposure studies (ROBINS-E) (Morgan et al., 2017). These two tools have been developed grounded on the same biases and they share similar criteria. The main difference is in the terminology used (e.g., the term 'intervention' is replaced with 'exposure').

Among the seven categories of biases identified in Table 6, two were not considered for RCT and NRS in the MMAT. The first category is reporting bias because the MMAT focuses on methodological quality and few CATs included a criterion on complete reporting of findings. Also, previous studies on the Cochrane RoB Tool showed that the criteria on this bias are often rated as 'unclear' or 'low' with a low reliability among reviewers (Hartling et al., 2013; Sterne, 2013; Vale, Tierney, & Burdett, 2013). Moreover, the developers wrote: "it makes little sense to classify all findings from a study as biased on the basis that it failed to report one or more

particular results (e.g., relating to an adverse effect)" (Sterne, 2013, p. 2). A recent review compared the reporting content between protocols/registrations and full reports of primary biomedical research and found that inconsistencies are frequent, common and suboptimal (Li et al., 2018). Also, criteria limited with the quality of reporting (e.g., criteria starting with 'clear description of') were not considered in this project.

Another bias not included in the RCT and NRS categories of the MMAT is analytic bias. Although several criteria on this bias can be found in CATs, analytic bias often overlaps with the other biases. In several CATs, some criteria on analytical bias are integrated with other biases. For example, the criterion "Did the authors use an appropriate analysis method that controlled for all the important confounding domains" is considered as a confounding bias although it is on statistical analysis (Sterne et al., 2016). Also, in RCT, intention-to-treat analysis is used to minimize attrition and selection biases (Armijo-Olivo et al., 2013; Higgins et al., 2016). Thus, no criterion specific to analytic bias (e.g., appropriate statistical analysis) was added in the RCT and NRS categories of the new version of the MMAT.

Some criteria of the mapping are linked with random errors such as adequate sample size and power calculation (Armijo-Olivo et al., 2013). These criteria are related to the threats to precision that occur due to unpredictable fluctuations that impact the precision of the estimates (Vetter & Mascha, 2017; Viswanathan & Berkman, 2012). As mentioned earlier, the MMAT adheres to a threats-to-validity approach and focus on systematic errors that can occur in a study. Thus, these criteria were not considered in the MMAT.

Table 8 presents modifications for the criteria on RCT in the MMAT. The MMAT (version 2011) has two criteria on selection bias, and two on attrition bias. On the basis of the results of the mapping and interviews with MMAT users, it is suggested to remove one criterion on attrition, reformulate one on selection, replace one on selection, and add two new criteria (one on measurement bias and one on performance bias) (Table 8). Criterion 2.1 on randomization was kept because found in most CATs (19/23). However, it was reformulated: the first part on description was removed to focus on methodological quality and the formulation was simplified. Criterion 2.2 on concealment was replaced with another criterion on selection bias on baseline equivalence (15/23). Concealment is partly addressed in the new criterion 2.1. Indeed, to judge if randomization is well performed, appropriate sequence generation and concealment of

assignment are necessary. Some MMAT users mentioned that the distinction between criteria 2.3 and 2.4 is not clear (theme 9.4). Since both are related to the same category of bias, it is suggested to remove criterion 2.4. In criterion 2.3, the cut-off value was removed because it is not absolute and there is no standard. In the literature, complete data value ranged from 80% (Thomas, Ciliska, Dobbins, & Micucci, 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder, Furlan, Bombardier, Bouter, & Editorial Board of the Cochrane Collaboration Back Review Group, 2003) and 30% for follow-up of more than one year (Viswanathan & Berkman, 2012). Two new criteria were added to cover the different categories of bias found in RCT (i.e., measurement and performance bias). The most frequent criterion on measurement bias identified in the mapping is on the blinding of outcome assessment. It was partly included in the criterion 2.2 in the previous version of the MMAT. Since criterion 2.2 was replaced, a criterion on blinding of assessor was added. Another criterion added was related to performance bias. The most frequent performance bias criteria found in the mapping were related to blinding of subjects and carers. However, blinding of subjects and carers is not always possible, especially in nonpharmacological trials such as in rehabilitation and public health (Armijo-Olivo et al., 2014; Victora, Habicht, & Bryce, 2004). Thus, another criterion that could be applicable to most RCT was chosen instead and concerns the adherence to intervention.

Table 8. Modifications to the Randomized Controlled Trials Criteria of the MMAT

MMAT (version 2011)	**Modifications suggested for MMAT (version 2018)**
Selection bias 2.1. Is there a clear description of the randomization (or an appropriate sequence generation)?	Reformulate Is randomization appropriately performed?
Selection, Measurement and Performance biases 2.2. Is there a clear description of the allocation concealment (or blinding when applicable)?	Replace Were the groups comparable at baseline?

Attrition bias 2.3. Are there complete outcome data (80% or above)?	Reformulate Are there complete outcome data?
Attrition bias 2.4. Is there low withdrawal/drop-out (below 20%)?	Remove
	New on measurement bias Are outcome assessors blinded to the intervention provided?
	New on performance bias Did the participants adhere to the assigned intervention?

Regarding NRS, the MMAT (version 2011) has two criteria on selection bias, one on measurement and one on attrition. Based on the mapping, it is suggested to remove one criterion on selection, reformulate three (one on measurement, one on selection and one on attrition), and add two new criteria (Table 9). Since there are two criteria on selection bias, it is suggested to remove criterion 3.3. Criterion 3.1 was reformulated to focus on the representativeness of subjects. Criterion 3.2 was reformulated to include both the measurement of exposure and outcome. For the same reason explained in the previous paragraph, the cut-off value for complete outcome data was removed from criterion 3.4. A note on the acceptable rates found in the literature was added in the tutorial. Two new criteria were added to address confounding and performance biases. Regarding confounding bias, the two most frequent criteria mentioned in the CATs were confounders taken into account in the design and/or the analysis. These criteria were added in the MMAT. Concerning performance bias, the most frequent criterion found in the mapping was blinding of subjects. For same reasons mentioned for RCT, blinding of subjects was not added in the MMAT. Also, this criterion was mainly mentioned in the CATs that were developed to assess both RCT and NRS; only two CATs specific to NRS mentioned blinding. Instead, a criterion on the intervention/exposure integrity was added.

Table 9. Modifications to the Non-Randomized Studies Criteria of the MMAT

MMAT (version 2011)	Modifications suggested for MMAT (version 2018)
Selection bias 3.1. Are participants (organizations) recruited in a way that minimizes selection bias?	Reformulate Are the participants representative of the target population?
Measurement bias 3.2. Are measurements appropriate (clear origin, or validity known, or standard instrument; and absence of contamination between groups when appropriate) regarding the exposure/intervention and outcomes?	Reformulate Are measurements appropriate regarding both the outcome and intervention (or exposure)?
Selection bias 3.3. In the groups being compared (exposed vs. non-exposed; with intervention vs. without; cases vs. controls), are the participants comparable, or do researchers take into account (control for) the difference between these groups?	Remove
Attrition bias 3.4. Are there complete outcome data (80% or above), and, when applicable, an acceptable response rate (60% or above), or an acceptable follow-up rate for cohort studies (depending on the duration of follow-up)?	Reformulate Are there complete outcome data?
	New on confounding bias Are the confounders accounted for in the design and analysis?
	New on performance bias During the study period, is the intervention administered (or exposure occurred) as intended?

7.1.2.2 Criteria for qualitative, mixed methods and quantitative descriptive studies

To address the three other sets of criteria in the MMAT, a modified e-Delphi technique was used with methodological experts (Chapter 6). The results of this study identified six qualitative criteria, seven mixed methods criteria, and eight survey criteria.

Qualitative research

In the literature, there is still a lack of consensus on how critical appraisal should be performed, especially for qualitative research. Some consider that the application of scales and checklists as seen for quantitative studies is too reductionist, idiosyncratic, unreliable and prescriptive for qualitative studies (Barbour, 2001; Leeman, Voils, & Sandelowski, 2015). Moreover, some aspects of qualitative research, such as the quality of insight and interpretation, are considered difficult to appraise and subjective (Dixon-Woods, Shaw, Agarwal, & Smith, 2004). Dixon-Woods et al. (2007) found that the use of checklists sensitizes reviewers to specific aspects of research practice or reporting and that they do not produce higher levels of agreement between or within reviewers compared with unprompted judgment. Yet, CATs are considered helpful to ensure that the key methodological limitations are examined in a systematic and transparent manner (Munthe-Kaas et al., 2018). Also, it has been advocated "to identify a minimum set of 'core domains' for assessing methodological limitations" of qualitative studies (Lewin et al., 2015, p. 13). In a Delphi study with 18 qualitative experts on the development of reporting guidance, the experts were against having fixed criteria; they instead agreed on the importance of having criteria that are open, adaptable and flexible to better respond to methodological changes and research questions (Hannes, Heyvaert, Slegers, Vandenbrande, & Van Nuland, 2015).

From the literature review performed on CATs (Chapter 3), only nine CATs on qualitative studies were retained. This number can be considered small compared to other reviews of this topic. For example, a recent review analyzed 58 assessment guidelines for qualitative studies with 56 experts and users in health sciences (Santiago-Delefosse et al., 2016). They found that consensus can be reached for 12 general criteria that are not necessarily specific to a qualitative method and that follow the logical course of a research plan: theoretical framework, research question, goals/objectives, literature review, methodology/method/design, sampling, data, analysis, reflexivity, credibility, transferability, and ethics (Santiago-Delefosse et al., 2016). Several of these criteria were also found in other CATs but were excluded during the e-Delphi study because they are not specific to qualitative studies (e.g., research ethics respected) or related to reporting quality (e.g., explicit theoretical framework).

Based on the results of the e-Delphi study (Chapter 6), three new qualitative criteria were added in the MMAT relating to the qualitative approach used, the coherence between qualitative data sources, collection, analysis, and the interpretation of data. A new criterion was added in the MMAT on the appropriateness of the approach used to answer the research question. This criterion can be found in four other tools (Critical Appraisal Skills Programme, 2017a; Kmet et al., 2004; National Institute for Health Care Excellence, 2012; Vermeire et al., 2002). This criterion was judged important to add in the MMAT since there is only one category of criteria for qualitative studies. Compared with quantitative studies where different tools exist for different study designs, many tools for qualitative studies encompass a wide range of study designs (Lewin et al., 2015). Among the CATs for qualitative studies identified during the literature review (Chapter 3), all tools could be applied to any qualitative approaches except for one tool for focus groups (Vermeire et al., 2002),

The second qualitative criterion added in the MMAT is on coherence. This criterion was suggested during the team meeting and is also found in four other tools (Joanna Briggs Institute, 2017b; Reis, Hermoni, Van-Raalte, Dahan, & Borkan, 2007; Sandelowski & Barroso, 2002; Spencer, Ritchie, Lewis, & Dillon, 2003). Some of these tools referred to other constructs, such as congruity and logical consistency, which were considered similar to coherence. This criterion is considered important to judge the credibility of a qualitative study, i.e., the confidence in the truthfulness of the findings (Lincoln & Guba, 1985). Santiago-Delefosse et al. (2016, p. 149) identified credibility as one important criterion and defined it as follows: “a researcher's credibility is based on the logical consistency that exists between the theoretical reference, research question, collection techniques and data analysis.”

A third new qualitative criterion concerned the interpretation of data. This criterion is found in two other tools (Boeije, van Wesel, & Alisic, 2011; Sandelowski & Barroso, 2002). The original item in the Delphi study was ‘The interpretation of results is plausible and sufficiently substantiated with data’. The experts suggested separating this item for Round-two of the e-Delphi because it addresses two concepts. Both reached consensus among the experts. The one on plausibility was not retained because the index of agreement was slightly lower than the other one and some experts mentioned that it might be harder to judge.

Mixed methods research

The literature on the quality of mixed methods studies is more recent but also subject to similar debates to qualitative research such as the use of a standard/fixed vs. flexible approach to quality appraisal (Burrows, 2013; Fàbregues et al., 2018). Also, since qualitative and quantitative methods are combined, new constructs have been introduced for mixed methods studies. For example, it was suggested to address the construct of 'legitimation' instead of validity (Onwuegbuzie & Johnson, 2006). Other constructs specific to mixed methods studies are found such as 'meta-inferences', which is defined as the inferences derived from integrating qualitative and quantitative findings (Teddlie & Tashakkori, 2009), and 'shared bias', which refers to the fact that both qualitative and quantitative methods are subject to the same biases (Curry & Nunez-Smith, 2014). The literature on the quality of mixed methods studies has seen an increase in the number of papers published. Two reviews on this topic can be identified in the literature. A first one searched the literature until December 2009 and identified 18 papers on 13 checklists developed to evaluate the methodological quality of primary mixed studies research (Heyvaert et al., 2013a). They identified specific and generic criteria for mixed methods research; those pertaining to the design, rationale, integration and interpretation were among the most popular. A more recent review identified 64 papers published until February 2016 and found 46 criteria cited in at least two papers (Fàbregues & Molina-Azorín, 2017). The results of this latter review were used in the e-Delphi study.

Although several papers on the quality of mixed methods studies can be found, few tools have been validated. Besides from the MMAT, only two other CATs specifically designed for mixed methods studies were found in the literature review (Chapter 3): the QATSDD (Sirriyeh et al., 2012) and the Evaluative Tool for Mixed Methods Studies (Long et al., 2002). These tools include generic criteria and specific qualitative and quantitative criteria, but do not have any specific mixed methods criteria such as the integration of both components, which is a core characteristic of mixed methods studies (Creswell & Plano Clark, 2018). To our knowledge, the MMAT is the only available CAT that also includes criteria for mixed methods studies. Some recent tools have been developed after the period covered in the literature review performed for this dissertation (e.g., MIXED framework (Eckhardt & DeVon, 2017)), but have not been tested for validity and reliability.

Based on the results of the e-Delphi study, three criteria were replaced and two new were added. Still few empirical studies have been conducted on the quality of mixed methods studies; the existing papers on the topic are mainly opinion or theoretical papers. Besides from the e-Delphi study performed in this dissertation, two other studies on this topic can be found. First, a multi-phase mixed methods study was conducted with 12 methodologists to develop a reporting checklist and to pilot test it with five reviewers: the Burrows Rubric for Evaluating Mixed Methods (BREMM) (Burrows, 2013). The BREMM includes 15 criteria rated on a 6-point scale (from strongly disagree to strongly agree). Several criteria in this tool are not found in the MMAT because they focus on reporting quality and are generic criteria on the research process such as clear statements on the purpose, research questions, literature review, philosophical assumption, and results.

Second, a recent qualitative multiple case study was performed to describe and compare how researchers from four different disciplines operationalize and conceptualize the quality of mixed methods research (Fàbregues et al., 2018). They interviewed 44 researchers in mixed methods studies and conducted within-case and cross-case analysis for pattern seeking. In this study, they identified 14 criteria that were mentioned more than five times by the participants, among which the most popular were: the quality of quantitative and qualitative components, the provision of a rationale for using a mixed methods research design, the effective integration of the quantitative and qualitative components of the study, and a clear and accurate description of the mixed methods design implemented. Three of these four criteria were included in the MMAT. No criteria on design were added in the MMAT because they did not reach a high level of consensus in the e-Delphi study. Also, criteria pertaining to design often focus on reporting quality (e.g., clear description of the mixed methods research design). Besides, two other criteria were added in the MMAT on inferences and divergences. These criteria were also found in Fàbregues et al. (2018) but with less prominence.

Quantitative descriptive studies

Regarding quantitative descriptive studies, the change made in this category was based on the results of the e-Delphi with experts in survey research. The team decided to focus on survey for the e-Delphi because it has not been done before, survey is one of the most used quantitative descriptive studies, and it is an umbrella term for other descriptive designs. For

example, cross-sectional study is also named prevalence study and frequency survey (Grimes & Schulz, 2002b). Two changes were made to criteria of the previous version of the MMAT based on the results of the e-Delphi study: one criterion on response rate was replaced and one new criterion on statistical analysis was added. The cut-off value for response rate (60%) was removed because it was considered arbitrary by the experts of the e-Delphi. Also, the appropriate response rate can vary a lot depending on several factors such as the topic, participants characteristics, data collection mode (e.g., phone, internet, mail), and survey design (e.g., display, text appearance, length) (Hoonakker & Carayon, 2009). Instead, the experts suggested that what is more of a concern is nonresponse bias. Having a good response rate does not necessarily minimize nonresponse bias (Groves & Peytcheva, 2008); response rates can be low, but still be unbiased whilst response rates could be high but biased (i.e., non-respondents may be significantly different from respondents). The criterion on statistical analysis was judged highly relevant by the experts, and was added in the MMAT. This criterion can be found in two other CATs on descriptive studies (Joanna Briggs Institute, 2017a; Munn, Moola, Riitano, & Lisy, 2014).

In epidemiology, several descriptive designs can be found: prevalence studies, case series, case reports, surveillance data, descriptive analyses of routinely collected data (such as registries and mortality data), correlational studies, and trend studies (Grimes & Schulz, 2002a; Hennekens & Buring, 1987; Pai & Filion, 2014). Interviews with MMAT users and discussion with team members suggested adding the term 'survey' in the list of quantitative descriptive design. Besides from the aim (i.e., to provide a description), one common characteristic of these descriptive studies is that they include one individual or one group of participants; stated otherwise, there is no comparison group in these studies (Grimes & Schulz, 2002b; Hartling et al., 2010). In papers on classification of study designs, descriptive studies are also named 'non-comparative studies' (Hartling et al., 2010; Seo et al., 2016; West et al., 2002).

Compared to RCT and NRS, few validated CATs on quantitative descriptive studies were identified in the literature review. Several existing tools are developed for prevalence studies (Al-Jader et al., 2002; Giannakopoulos, Rammelsberg, Eberhard, & Schmitter, 2012; Hoy et al., 2012; Joanna Briggs Institute, 2017a; Munn et al., 2014; Shamliyan et al., 2011). Two CATs on intervention case series can be found (Guo, Moga, Harstall, & Schopflocher, 2016; Yang et al.,

2009). Concerning survey, no validated tool was identified. Only a reporting guideline for survey research that was developed from a systematic review of the literature can be found (Bennett et al., 2011). Also, a recent CAT not included in the literature review was published: Appraisal tool for Cross-Sectional Studies (AXIS tool) (Downes et al., 2016). This tool was developed from a Delphi study with 18 experts and includes 20 criteria. The five criteria in the MMAT on sampling, coverage, measurement, nonresponse and analysis are found in this tool. The other criteria in the AXIS tool are mainly on reporting quality and focus on the objective, results, discussion, funding sources, and ethics.

7.2 MMAT Version 2018: A Revised Version of the MMAT Version 2011

From the results of phases 1 and 2, a revised version of the MMAT was produced (see Appendix 10): MMAT version 2018 checklist and tutorial. Three main changes were made to the previous version of the MMAT (version 2011):

1. Modify criteria: Modifications were made to the MMAT to address two problematic themes identified in phase 1 of this project: concerns about completeness (theme 3) and items not clear or difficult to judge (theme 9). The findings of phase 2 (e-Delphi and mapping) informed the changes needed in the criteria of the MMAT: four criteria were removed, seven were reformulated, five were replaced, and ten new were added. During the development of the revised version of the checklist and tutorial, care was taken to maintain as much as possible the characteristics that were appreciated (e.g., easy to use, short and quick). The revised version of the MMAT includes five criteria for each category of studies, which is one additional criterion for qualitative and quantitative studies and two for mixed methods studies compared with the previous version. The mixed methods studies category now has the same number of criteria as the other categories. A criterion on the quality of each method was added. This was addressed in the previous version of the MMAT by asking reviewers to use three sets of criteria (qualitative set, one of the quantitative sets, and mixed methods set) when appraising a mixed methods study. Adding this criterion makes it more explicit in the new version of the MMAT. In the other categories, new criteria were added to cover different categories of biases. Also, the formulation of some criteria was simplified by removing details and examples (these were put in the tutorial instead).

2. Remove the overall scoring: MMAT users also mentioned that there was a need to clarify how to compute the overall score in the tutorial (theme 7). This led to changes in the overall scoring of the MMAT. There is still much debate about the use of summative score in critical appraisal (Glenny, 2005). The use of a summative numerical score is a simple way of providing an overall idea of the quality of a study. However, a single number does not provide information on what aspects of studies are problematic and can even hide serious defects (Crowe & Sheppard, 2011). Also, it is unclear whether criteria should be weighted or not (Colle, Rannou, Revel, Fermanian, & Poiraudeau, 2002; Higgins & Green, 2008). Currently, it is discouraged to calculate an overall score from the ratings of each criterion (Herbison, Hay-Smith, & Gillespie, 2006; Higgins & Green, 2008; Viswanathan et al., 2012). On this basis, it was decided to remove the summative numerical score from the MMAT. Instead, it is advised to provide a more detailed presentation of the ratings of each criterion to better inform the quality of the included studies, and encourage sensitivity analysis.
3. Modification in the tutorial: The format of the tutorial was maintained since the MMAT users appreciated the lists of study designs. Explanations were provided to help the reviewers judge the criteria in the MMAT (theme 7). Besides, two other problematic themes led to minor changes in the tutorial of the MMAT. Concerning the study designs that cannot be assessed using the MMAT (theme 2), it was decided among the team members not to add new categories of studies for now. The MMAT suggests general criteria that could be applied for qualitative, quantitative, and mixed methods studies. However, there are some types of studies that may require more specific criteria such as economic studies. A note on this limit was added in the tutorial. In future development of the MMAT, new categories of studies could be added if needs are expressed by more MMAT users. Regarding the screening questions, no change was made but a note was added in the tutorial to explain why there are suggested (i.e., to exclude non-empirical studies from the appraisal). Finally, an algorithm was added in the tutorial to help MMAT users choose the set(s) of criteria to use for their review (theme 10). Algorithm is a visual step-by-step process that allows, through decision rules, to classify study designs (Hartling et al., 2010). It was developed based on several existing algorithms of quantitative study designs (Hartling et al., 2010; Hartling, Bond, Santaguida, Viswanathan, & Dryden, 2011; National Institute for Health Care Excellence, 2012; Scottish Intercollegiate Guidelines Network; Seo et al., 2016; West et al., 2002; Zaza

et al., 2000). These algorithms were simplified for the purpose of the MMAT and study designs of qualitative and mixed methods studies were added. Only the main study designs are presented in the algorithm; the list is not exhaustive.

7.3 Comparison of the MMAT With Other Existing Critical Appraisal Tools

Currently, there exist more than 500 CATs. Based on the literature review of CATs, four general categories of CATs were identified: (a) generic, (b) generic and specific criteria, (c) specific criteria for categories of studies, and (d) specific criteria for study designs. The most common categories are the specific tools, i.e., tools including specific criteria for a category of studies (e.g., qualitative studies, epidemiological studies) or a study design (e.g., RCT, cohort studies). When using these categories of tools in SMSRs, it implies that different sets of criteria are used for different study designs. The generic tools can be advantageous for SMSRs since a same set of criteria can be used for several study designs. However, they often focus on reporting quality criteria since what is usually common throughout the different designs are general information that must be included when reporting of a study (e.g., purposes/objectives, methods, ethics, results, discussion, and conclusion).

Even though the MMAT can be used for different study designs, it is considered as a specific tool since it includes core criteria for categories of studies (i.e., qualitative, NRS, descriptive, and mixed methods studies) and for one specific study design (i.e., RCT). The MMAT can be compared to tools developed by JBI (Joanna Briggs Institute, 2017c), NICE (National Institute for Health Care Excellence, 2012), SIGN (Scottish Intercollegiate Guidelines Network, 2017b), and CASP (Critical Appraisal Skills Programme, 2017b). They offer a variety of tools (up to 13) for specific study designs or categories of studies. Table 10 presents the study designs that are covered in these tools.

Table 10. Comparison of the MMAT with Critical Appraisal Tools Developed by CASP, JBI, NICE, and SIGN

Category	Tools				
	MMAT	**CASP**	**JBI**	**NICE**	**SIGN**
QUAL	Yes	Yes	Yes	Yes	No
RCT	Yes	Yes	Yes	Yes (1 tool: quantitative intervention studies)	Yes
NRS	Yes	Yes (2 tools: case-control studies and cohort studies)	Yes (4 tools: case-control studies, quasi-experimental studies, analytical cross-sectional studies, cohort studies)	Yes (1 tool: quantitative studies reporting correlations and associations)	Yes (2 tools: cohort studies and case-control studies)
Descriptive	Yes	No	Yes (3 tools: case reports, case series, and prevalence studies)	No	No
Mixed methods	Yes	No	No	No	No
Other	No	Yes (4 tools: economic studies, diagnostic studies, systematic reviews, and clinical prediction rules)	Yes (4 tools: economic studies, diagnostic test accuracy studies, systematic reviews, and text and opinion)	Yes (1 tool: economic studies)	Yes (3 tools: economic studies, diagnostic studies, and systematic reviews and meta-analyses)
Total number of criteria	25	89	126	80	69

Five main differences can be found. First, the CASP, JBI, NICE and SIGN tools include 6 to 27 criteria per tool for a specific design. The MMAT has fewer criteria since it focuses on the core ones. Second, compared to JBI, SIGN and CASP, the MMAT has only one set of

criteria for NRS. In the MMAT, the criteria that are common to case-control, cohort and quasi-experimental studies were identified and these designs were grouped into one broad category (i.e., NRS). Third, besides from the MMAT, only JBI has a tool for descriptive studies. Still very few validated tools have been developed for descriptive studies. Fourth, compared to the others, the MMAT does not have specific criteria for economic, diagnostic studies, and systematic reviews. The MMAT includes general categories of studies that are usually found in SMSRs. In the interviews with the MMAT users, some mentioned having difficulties appraising certain types of studies (theme 2). This could be further investigated to identify if there is a need to include other categories of studies in the MMAT. Also, there are no criteria for systematic reviews because the MMAT was developed for use in SMSRs that include empirical studies. One last difference is that the MMAT includes specific criteria for mixed methods studies, which are not found in the other tools.

The MMAT can also be comparable to risk of bias tools (such as the Cochrane RoB) since the focus is on methodological quality and does not follow the hierarchy of evidence approach. However, the Cochrane Collaboration Handbook makes a clear distinction between quality and bias. In their view, the term 'assessment of methodological quality' suggests that the appraisal is on the extent to which a study followed the highest possible standards (Higgins & Green, 2008). They argue that a study may respect high standards, yet can still have important biases. Thus, they coined the term 'assessment of risk of bias' and focus their appraisal on the extent to which results of a study can be trustworthy (Higgins & Green, 2008). They reserved the term 'quality of evidence' to describe the "extent to which one can be confident that an estimate of effect is near the true value for an outcome, across studies" (Higgins & Green, 2008, p. 190). This distinction was not made in this dissertation since the MMAT includes criteria for different study designs. The construct of 'bias' is mainly used in quantitative studies, less in qualitative and mixed methods studies.

The field of CATs is constantly growing. The literature search was carried out in 2015 (chapter 3) and several new CATs not included in this dissertation can already be identified. Recent tool developments focused mainly on specific fields. For example, recent tools have been developed for RCT and NRS of research on birth place (Vedam, Rossiter, Homer, Stoll, & Scarf, 2017), and for RCT, observational and systematic reviews on drug adverse events (Faillie et al.,

2017). This might explain why there exist so many CATs. The MMAT was developed to be applied to any field of human subjects research. During the interviews with MMAT users, some mentioned they would appreciate a more flexible tool allowing to add other criteria they judge important to appraise for their topic under review (theme 4). This can lead to a new way of conceptualizing CATs. For example, in addition to core criteria, a bank of validated criteria could be made optional to allow the tool to be tailored as appropriate.

7.4 Limitations of This Project

Several challenges and limitations were encountered during this project. Regarding the literature review, the lists of SMSRs and CATs identified are not exhaustive. For the search of SMSRs, no other sources than bibliographic databases were searched. Since the field of SMSR is still new, there is no specific controlled vocabulary and only keyword terms were used in the search strategy. General controlled terms were tested such as 'review' or 'literature review as topic', which generated too much noise. Also, since it was a methodological review, backward or forward citation tracking of SMSRs would not have provided additional references. For the search for CATs, the majority of the retained reviews of CATs focused on the quality of quantitative studies. Only five of the 17 reviews addressed other types of study designs. More recent reviews of CATs for qualitative studies can be found in the literature (Munthe-Kaas, Lewin, & Glenton, 2017; Santiago-Delefosse et al., 2016).

Another limitation during the literature review concerns a potential selection and interpretation bias. The selection of CATs and interpretation of findings on measurement properties were performed by one reviewer. One challenge encountered was the lack of information on critical appraisal in the included SMSRs. For example, several reviews did not provide a clear description of how many tools were used. Several mentioned using the tools from CASP or JBI, without detailing which one. Yet, the CASP has eight different CATs and the JBI has 13 CATs. To simplify the analysis, the CATs from a same organization were counted as one tool in Appendix 3. Also, the analysis of the literature was limited to the textual content available. The authors were not contacted for missing or unclear information. Also, several SMSRs used CATs that were developed and used in other systematic reviews. These tools were not extracted from the original sources. Similarly, 152 SMSRs mentioned developing their own

criteria. It could be interesting to go back to the origin of each tool and to further explore the reasons leading to choose/modify an existing CAT or to develop a new one.

No quality appraisal of the retained SMSRs was performed during the literature review. The literature review focused mainly on the review process (i.e., how the review was conducted) than on the findings of the SMSRs. Also, currently, the existing tools for appraising the reporting and methodological quality of systematic reviews are for quantitative reviews (Moher et al., 2009; Shea et al., 2009; Whiting et al., 2016), and are not adapted for SMSRs (Bouchard, Dubuisson, Simard, & Dorval, 2011).

Regarding phase 1 of the project, the potential participants were identified through published systematic reviews and contacts with the developer. This might have biased the types of participants that were largely doctoral students and postdoctoral fellows. Other potential MMAT users were not contacted, such as HTA professionals and master students. Also, the interviews of the qualitative descriptive study were performed by one person that is familiar with the MMAT. Prior to data collection, she participated in four reviews as second reviewer for the appraisal of studies using the MMAT. While working on these reviews, she has identified some areas that need improvement (e.g., criteria that need further clarification, missing criteria). Her preconceptions might had bias the interviews and analyses. Care was taken to minimize the impact of this bias and make sure that the data collected and analyzed represent the experience of the MMAT users. A semi-structured interview guide was developed prior to starting data collection so that similar questions were asked to the participants. Also, a second coder that has not used the MMAT in a systematic review was involved in the data analysis of the interviews. No major difficulties were encountered between the coders since the level of interpretation of data was low. Compared with other qualitative designs, qualitative descriptive study usually entails a low-inference interpretation level since the aim is to provide a comprehensive summary of events using the terms of the events (Sandelowski, 2000, 2010). Moreover, the interpretation of the identified themes involved a third coder.

Regarding phase 2, the mapping of criteria of the RCT and NRS was performed by one person. An interpretation bias can be present especially for criteria that are not clearly described in the papers. Concerning the modified e-Delphi study, no consensus was reached for the criteria on measurement and nonresponse bias for surveys. Due to the difficulty recruiting these experts

and the results obtained, it was decided not to further investigate these criteria by conducting supplementary Delphi rounds. The recruitment of experts was challenging since it aimed at mobilizing several researchers during a same period of time. Several reminders were sent and a longer turnaround time was given for panel members (one month). The start date of this phase was chosen to avoid recruitment during a busy period (e.g., during grants submission deadlines or at the beginning of the school semester) and also considering that it was preferable that the project ends in June at the latest before the start of summer vacation.

7.5 Strengths of This Project

A structured stepwise project using a sequential exploratory mixed methods design was performed. The integration of the qualitative and quantitative component occurred between the phases and after the results of both phases to inform the revision of the MMAT (Pluye et al., 2018). This design is often used for tool development (Creswell & Plano Clark, 2018). The sequential exploratory design has the advantage of being straightforward to implement and report (Creswell & Plano Clark, 2018). However, each phase depends on the success of the previous one. Regular monitoring of the progress of the project was performed to ensure it was completed within the timeframe planned.

Although the literature review was not exhaustive, a large sample of SMSRs was found (n=459). To our knowledge, only one other review on SMSRs was conducted back in 2006 and had identified 17 SMSRs (Pluye et al., 2009). Moreover, 17 reviews on CATs were identified and accounted for more than 500 CATs. The literature review provided a good overview of the existing tools and current state of knowledge of SMSRs.

The numbers of participants recruited in this project could be considered acceptable. In phase 1, data saturation was achieved with 20 MMAT users. In phase 2, the number targeted in each group was at least 15 experts after two rounds. This number was estimated from the literature on the Delphi technique and content validity. In the literature on Delphi technique, a size of 10 to 15 participants is usually deemed sufficient for homogeneous samples of experts (Keeney et al., 2011). In the literature on content validation, the numbers of experts ranging from 5 to 12 were suggested (Lynn, 1986; Polit, Beck, & Owen, 2007). In this project, respectively 21, 15, and 20 experts participated in Round-two for qualitative, survey, and mixed methods studies.

The participants came from different countries: eight countries in phase 1 and 11 countries in phase 2. The large number of countries involved in this project demonstrates that the field of critical appraisal and systematic reviews is of international interest. It can also provide some evidence on the ecological validity of the MMAT. Ecological validity is a subset to external validity and refers to the transferability of findings from an experimental context to the real-world environment (Khorsan & Crawford, 2014; Schmuckler, 2001). This project allowed identifying areas for improvement of the MMAT with 'real-world' users (i.e., users who were not directly involved in the initial development of the MMAT).

7.6 Contribution to Knowledge

This project provided new implications for conducting SMSRs and to the advancement of knowledge in the fields of SMSRs and critical appraisal. In the following, six main contributions will be presented into three categories: conceptual, methodological, and practical.

7.6.1 Conceptual contributions

First, this project contributed to clarifying the construct of SMSR (Figure 1 in Paper #1, section 2.2). SMSR is defined as a type of systematic review in which qualitative, quantitative, and mixed methods studies are combined. In SMSRs, the integration can be seen at the level of studies (i.e., combining different study designs) and the level of synthesis (i.e., combining quantitative and qualitative synthesis methods) (Heyvaert et al., 2013b; Hong et al., 2017). This conceptualization of SMSR was helpful for distinguishing the different terms currently being used to designate this type of review (section 2.1). Several categories of terminology used in reviews can be identified. Under the category on the types of studies (Table 2), it can be possible to distinguish SMSRs from quantitative reviews (i.e., reviews of quantitative studies only) and qualitative reviews (i.e., reviews of qualitative studies only); the term 'mixed' in SMSR means combining qualitative, quantitative and mixed methods studies. The confusion in the terminology is mainly seen between SMSRs and mixed methods reviews. In our view, SMSR is more global and encompasses mixed methods review. SMSR is about combining studies of different designs and can use one or several synthesis methods. Whereas mixed methods review is about combining different synthesis methods (such as using meta-analysis and thematic synthesis).

Second, this project provided a conceptual framework in which the different dimensions of quality and purposes for performing critical appraisal in SMSRs are presented (Figure 2 and Table 2 in Paper #1, section 2.2). Three dimensions of quality were identified: methodological, conceptual, and reporting. The methodological quality provides information on trustworthiness and is the most frequently mentioned, studied and debated dimension. Most of the CATs developed, especially the most recent ones, have been used to appraise this dimension (Bai et al., 2012). The conceptual quality is mainly described in interpretive synthesis to judge if a study provides rich insight into a concept. Still few papers have addressed this dimension of quality. Reporting quality has become a requirement of several journals that asked to complete a reporting checklist before the submission of a manuscript to ensure that all the important information on a specific study design is included. Although not advised as a proxy for methodological quality (Dreier, Borutta, Stahmeyer, Krauth, & Walter, 2010; Higgins & Green, 2008; Wells & Littell, 2009), reporting quality is often appraised in systematic reviews. Researchers do not always distinguish between reporting and methodological quality criteria. Yet, this distinction can be important since the results, recommendations, and conclusion provided from the reviews will be influenced by the quality dimension assessed. A review that used a reporting tool can provide information on the accuracy, transparency, and completeness of a paper but cannot necessarily infer on the trustworthiness of the studies.

7.6.2 Methodological contributions

Third, this project makes a methodological contribution to critical appraisal. The literature review identified more than 500 CATs that can be classified into four general categories: generic tools, generic tools with specific criteria, specific tools for categories of studies, and specific tools for study designs. These categories were helpful to differentiate the CATs and situate the MMAT. The MMAT has often been considered as a generic tool since it can be applied to different study designs. However, based on the categories identified in the literature review, generic CATs use a same set of criteria for all studies. Thus, the MMAT, does not fit in this category because it includes methodological criteria that are specific to categories of studies (qualitative, NRS, quantitative descriptive, and mixed methods studies) and one study design (RCT). Also, this project contributed to revising the MMAT, which was specifically designed to be used in SMSRs. This tool is unique as it includes criteria for different categories

of studies including mixed methods studies. The findings of the modified e-Delphi study with methodological experts (Chapter 6) and the mapping of the criteria of the CATs (Appendix 9) contributed to strengthen the content validity of the MMAT. Content validity is an important type of validity to consider in tool development to ensure that it includes criteria that are representative and relevant for the appraisal of methodological quality of studies (Haynes et al., 1995).

Fourth, in addition to a methodological contribution to critical appraisal, the review of the 459 SMSRs also contributed to providing an overview on the current state of knowledge in SMSRs and on the synthesis approaches used. The results of this analysis are presented in Appendix 2. This review showed the exponential growth of SMSRs over the past decade. The first identified SMSR was published in 1998. Since 2013, greater interest in this type of review can be found with more than 100 SMSRs published yearly. Also, this review proposed a typology of synthesis method designs including two main designs (convergent and sequential) and three levels of integration were identified (level of data, results of synthesis, and interpretation) (Appendix 2). Several other typologies of synthesis methods have been developed (Frantzen & Fetters, 2015; Heyvaert et al., 2013b; Sandelowski et al., 2012) but remain theoretical. In this review, the typology was developed from the analysis of 459 SMSRs. This typology can help reviewers develop their protocol, and better understand how to perform the synthesis and integration of qualitative, quantitative, and mixed methods studies.

7.6.3 Practical contributions

Fifth, at a practical level, the interviews conducted with MMAT users (Chapter 5) helped to identify areas for improvement of the usefulness (usability and utility) of the MMAT. Addressing usefulness is important to facilitate the adoption of the MMAT. The user manual was revised and more explanation on each criterion was provided (Appendix 10). Guidance is important to describe how to assess each criterion included in the tool (Whiting et al., 2017). References to the explanations were also added in the MMAT user manual so that users can refer to them if needed. Moreover, an algorithm was added to help users select the proper set of criteria to use. This can address problems often seen in scientific papers such as the mismatch between the label and description of the study design used, and inconsistent terminology

(Hartling et al., 2010).

Another usability issue will need to be addressed in a near future: update of the MMAT website. Up to now, dissemination strategies used for this project have been publications in scientific journals and presentations at international conferences. Using a website is another useful strategy, which can allow for wider distribution. For example, as of January 2018, the MMAT website has been consulted more than 30,000 times since March 2013 (this represents around 6,000 visits per year). Currently, the MMAT website is presented as a wiki using the PBworks platform (http://mixedmethodsappraisaltoolpublic.pbworks.com), which is a collaborative website where a visitor can become an active participant by getting involved in creating and editing the content (Boulos, Maramba, & Wheeler, 2006). However, such a website is less relevant in the case of the MMAT as its use is more informative than collaborative. Changes in the MMAT website might be warranted. Some MMAT users mentioned that the website is not user-friendly (theme 12). Once the MMAT (version 2018) will be pilot tested, the website will be updated to improve its navigation features including a simple menu bar, easier access to download the tool, and a list of additional resources such as a list of complementary tools (see Figure 7). In addition to the tool and user guide, other information can be included on the website such as information on training, detail on contributors and funding, and translations of the tool (Whiting et al., 2017). The MMAT (version 2011) was translated into French and Brazilian Portuguese (Robert, 2015; Souto et al., submitted). These translations will be updated with the latest version. The team will also collaborate with other research teams interested in translating the MMAT in another language. Finally, a rating sheet will be made available to facilitate the compilation of the ratings of the studies (theme 13). This rating sheet template will be developed in Microsoft Excel.

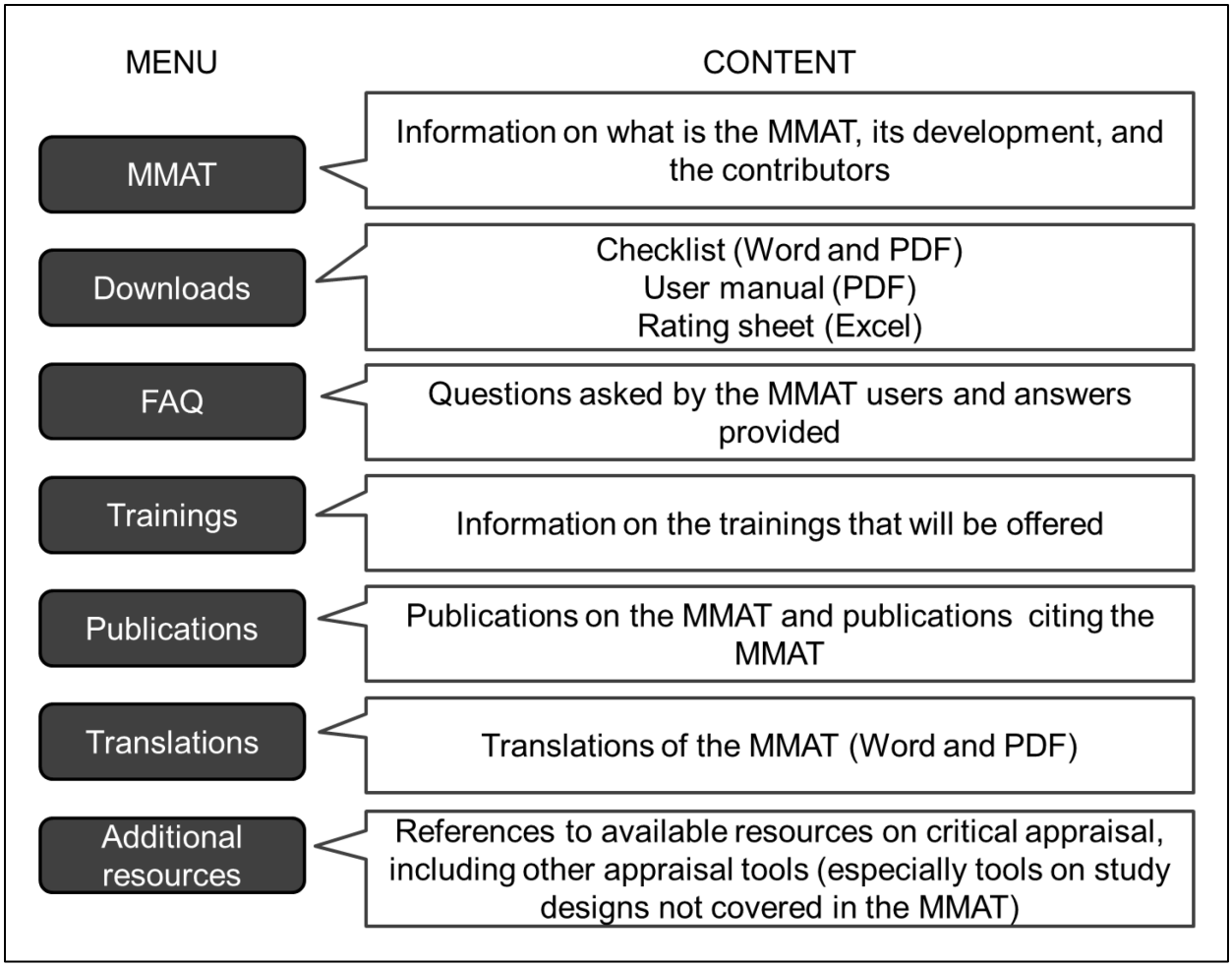


Figure 7. Plan for the Website Update of the MMAT

Finally, since several usefulness issues were addressed in this dissertation, it could be expected that the MMAT (version 2018) will be more useful compared to the MMAT (version 2011) to facilitate the critical appraisal of the methodological quality of several types of studies included in SMSRs, specifically the appraisal of qualitative, quantitative, and mixed methods studies. It can be compelling compared to its alternative, which consists of selecting and using separate CATs for each type of studies. This latter option can be complex as well as time and resources consuming since it requires that reviewers learn how to use several tools. Also, by limiting to core criteria, the MMAT can provide a more time efficient appraisal. In studies on the previous version of the MMAT (version 2011), the average time to complete the MMAT for one study ranged from 11 to 14 minutes (Pace et al., 2012; Souto et al., 2015). The efficiency of the

new MMAT (version 2018) will need to be tested. A similar duration might be expected since only one criterion is added for each category of studies (except for mixed methods studies that have two additional criteria). Moreover, the MMAT is not limited to one specific field; the reviewers can come from different fields such as education, health services and policy, and social sciences. The MMAT can be used by several reviewers involved in SMSRs such as researchers, graduate students, decision- and policy-makers, librarians, and service planners. As SMSRs gains in popularity, it can be anticipated that this tool be useful to a growing number of reviewers involved in this type of review.

CHAPTER 8. CONCLUSION

SMSRs are systematic reviews combining qualitative, quantitative, and mixed methods studies. They are increasingly popular in several fields due to their potential to address complex questions, interventions and phenomena as well as to provide more meaningful, practical and useful recommendations to clinicians, decision- and policy-makers, patients and researchers. Yet, several challenges need to be surmounted due the heterogeneity of studies included in these reviews. One of them is related to the critical appraisal of studies of diverse designs. This project addressed this challenge by revising a critical appraisal tool for use in SMSRs, the MMAT.

This project used a sequential exploratory mixed methods design; the results of a qualitative descriptive study (phase 1) informed the data collection of a modified e-Delphi study (phase 2). The integration occurred between and after the phases. In phase 1, interviews with researchers who have used the MMAT were conducted to understand their views and experiences. Then, the research team met to discuss the results of the analysis of interviews and potential areas for improving the MMAT. One important point concerned the clarity, relevance and completeness of the MMAT criteria. In phase 2, item-related issues were further addressed with a group of methodological experts. A modified e-Delphi study was performed and helped to identify relevant criteria to appraise the quality of studies. Also, a mapping of criteria from 33 existing critical appraisal tools was performed to identify the core criteria for RCT and NRS. This led to modify seven criteria, remove four, replace five, and add ten new in the MMAT, and allowed developing a revised MMAT (version 2018). Finally, to facilitate its usability, the MMAT user manual was revised. Further development in a near future will include modifying the website and adding a rating sheet. To conclude this dissertation, some final remarks regarding critical appraisal in SMSRs will be made in light of the knowledge and experience gained throughout this doctoral process as well as directions for future research.

8.1 Final Remarks Regarding Critical Appraisal in SMSRs

First, developing clear and consensual criteria is a challenging endeavour. Quantitative, qualitative, and mixed methods research use different terminologies and are associated with different dominant worldviews (Heyvaert, Hannes, & Onghena, 2016; Lincoln & Guba, 1985). Although there is no consensus on the best way to perform critical appraisal, the approach used

in the MMAT is to acknowledge these differences by suggesting different sets of criteria using the language of each type of studies. Ideally, SMSR teams should include members with expertise in qualitative, quantitative, and mixed methods research (depending on the type of studies included in the review). Moreover, the use of the MMAT might be more challenging for first-time users, especially for graduate students and novice researchers not familiar with the potential diversity of the types of designs across included studies in SMSRs. Graduate students should ask colleagues with expertise in qualitative, quantitative, and mixed methods research as needed to help for appraising studies with designs with which they are less comfortable or were never trained.

Second, the critical appraisal process is about judgment making. Many appraisal tools, such as the MMAT, are developed to assist and structure this process by suggesting criteria that need to be taken into consideration for appraising the methodological quality of included studies (in contrast to other tools designed for assessing the reporting or conceptual quality). However, the MMAT should not be used as a cookbook that needs to be strictly followed. Instead, it is intended to provide cues on what to look for in a paper and make the process more transparent and systematic. The interpretation of the criteria can vary depending on the fields and knowledge of the reviewers. Because the process of critical appraisal is subjective, it is usually recommended to involve at least two reviewers that will independently appraise the studies and discuss any discrepancies (and a third-party for making a final decision when disagreements between the two reviewers cannot be easily solved). Also, the MMAT does not replace the experts' judgment. For example, some reviewers might judge that other criteria are important to add depending on the topic and review question. It is usually recommended to agree on the criteria and their interpretation at the beginning of the review appraisal process, and apply them uniformly throughout the included studies.

Third, the impact of the critical appraisal process on the review findings should be made more explicit. In the literature review conducted on SMSRs (Chapter 3), some SMSRs described the critical appraisal process used but did not report any findings of this process. Also, other SMSRs presented an overall mean score of the quality of the included papers without detailing and discussing their impact on the interpretation of findings and recommendations. This led to question the reviewers' rationale for performing the appraisal process, and its consequences on

the review findings. Critical appraisal is a mandatory step in systematic review and is a very demanding process, especially if the criteria are subject to interpretation. Beyond meeting the requirement for conducting systematic reviews, there is a need to better justify why it is necessary to perform and what would be the impact on the findings of the review.

Fourth, in Chapter 2, an analogy was made between the critical appraisal and courtroom trial process: a jury (review) needs to judge the evidence (studies) conveyed by lawyers (research papers). How the evidence is reported can greatly influence the judgment made. A distinction between studies and research papers was made because the dimensions of quality are different. A lot of work has been done on reporting quality with the EQUATOR Network and several reporting guidelines have been developed since 2008 (Altman, Simera, Hoey, Moher, & Schulz, 2008). It can be expected that the reporting quality of papers will improve throughout the years. For methodological quality, there exist hundreds of tools but they are harder to identify due to the lack of a centralized library such as the EQUATOR Network. The National Collaborating Center for Methods and Tools (NCCMT) provides a registry of some available CATs in public health (Peirson, Catallo, & Chera, 2013). Such a registry could be extended to include a more exhaustive list of CATs, which will greatly facilitate the selection of appropriate tools. Also, several tools include a mix of reporting and methodological quality criteria. Both dimensions are interdependent but the appraisal of methodological quality requires a further step of judging the trustworthiness of the findings of a study from the reported information. Further research is needed to explore the interdependence of these dimensions and how it can translate into improving the appraisal of studies. For example, although the MMAT was developed for appraising the methodological quality of studies included in SMSRs, an indirect application could be to influence how researchers will report their future papers. Also, still few tools have addressed conceptual quality of studies.

Fifth, the courtroom trial analogy pinpoints the importance of providing good evidence to demonstrate the trustworthiness of the studies. In health sciences, some recommendations made can be a question of life and death, and multiple treatment recommendations balance health improvement vs. potential adverse events. Therefore, reviewers want to make sure that their review relies on strong evidence before recommending a treatment, especially if evidence is contradictory or finds it to be harmful. Similarly, a jury would not want to wrongly convict an

innocent person. In addition, from this analogy, a question arises as to whether the level of complexity in the process could be based on the importance of the case and its consequence. For example, in a courtroom trial process, a murder trial can be much more complex than a financial trial. Also, the consequences of the verdict made (e.g., fine, community sentence, or jail sentence) could influence the level of complexity of the process. Could different levels of appraisal be suggested depending on the review questions and consequences of the recommendations made in a review? For example, a systematic review recommending the use of a cardiac drug to reduce the mortality rate can have vital consequence (life vs. death) compared to a review interested in understanding a phenomenon (e.g., understand why people decide to exercise) where the immediate consequence is less critical. In the former, the appraisal process could provide a more complete account of all the validity criteria so that the results of the appraisal clearly inform which papers should be excluded or considered in sensitivity analysis. Conversely, in the latter, the appraisal could be used for description or synthesis purpose (less for exclusion purpose) and could focus on the most important criteria related to the phenomenon of interest. Thus, there is a need to better understand when critical appraisal is needed and how it could be adapted.

8.2 Directions for Future Research on the MMAT

This project focused on the content validity and usefulness of the MMAT. A revised version of the MMAT was developed and further testing is needed. There is a need to pilot test the tool with a group of experts and users to collect comments and feedback on the clarity of the wording and relevance of the criteria. This can lead to further modifying some criteria and improving the content validity of the MMAT. This step is important to make sure the criteria are clear and the MMAT adequately covers the construct under assessment, i.e., the methodological quality of qualitative, quantitative, and mixed methods studies.

Then, concurrent validity testing could be performed, i.e., to test the extent to which the tool correlates with measures of the same construct administered at the same time (De Vet, Terwee, Mokkink, & Knol, 2011). Since the MMAT covers different designs, different validated CATs will need to be compared such as the Cochrane RoB for RCT (Higgins et al., 2011) and the EPHPP for NRS (Thomas et al., 2004). Methodological experts' judgment is another measure

that can be used since the construct of interest concerns methodological quality of studies. High correlations with other tools and experts' judgment will indicate that the MMAT properly serves its intended purpose of assessing the methodological quality of different study designs.

The construct validity (including convergent and discriminant validity) is another type of validity that should be considered in tool development. Convergent validity studies on other CATs have used different related measures such as journal impact factor, citation rates, effect size, conflict of interest, years of publication, and funding sources (Kim et al., 2013; Moncrieff, Churchill, Drummond, & McGuire, 2001; Reed et al., 2007). These measures and others that can be relevant to the MMAT will need to be explored to properly test its construct validity. Also, during the interviews with the MMAT users (Chapter 5) concerns were expressed about the discriminant validity of the tool. This can be tested by choosing a sample of articles with equal distribution of good, moderate, and bad studies and check if the ratings of the MMAT differ among these studies.

In addition to validity, reliability testing of the MMAT is needed to ensure that it can be used in different circumstances and by different raters. The interrater reliability could be tested by asking several raters with different backgrounds to use the MMAT for appraising a preselected sample of research articles published in peer-reviewed journals. The same raters could also be asked to rerate the same articles after a predetermined time interval to ensure that same ratings are obtained on repeated use of the MMAT (test-retest reliability). The time interval will need to be long enough to avoid a memory bias. One advantage of working with studies is that the condition evaluated does not change over time as seen in patients' outcomes.

Continuous development of the MMAT is required. All measurement instruments need to be revised since there is never a 'final' version (Nunnally & Bernstein, 1994). The quality of the reporting of studies can improve in the years to come and might influence the methodological quality criteria that can be assessed. Also, more and more researchers are interested in meta-research studies and new evidence on critical appraisal is being created. For example, there has been a call to develop more empirical evidence on the association between the methodological quality criteria and treatment effects (Armijo-Olivo et al., 2013). As evidence develops, modifications might be necessary in the MMAT to keep it up to date with the latest developments. Moreover, there is a need to explore whether criteria on other study designs

should be added in the MMAT. For instance, several tools recently developed focus on specific study designs such as for the appraisal of prediction modelling studies (Wolff et al., 2017) and the appraisal of moderator and predictor analysis (van Hoorn et al., 2017). Also, currently, there is one set of criteria for all qualitative studies in the MMAT. Further studies could explore the need to new add criteria regarding specific qualitative approaches (e.g., qualitative description, grounded theory, phenomenology, and ethnography). This could lead to the development of a 'MMAT Plus' as suggested during the team meeting of the MMAT developers.

To keep up to date with the evidence on CATs, a continuing surveillance of the literature trends is needed. A monitoring system, such as eSRAP, could be used. eSRAP uses crowdsourcing to filter information and identify emerging peer-reviewed papers, and offers a structured and continuously updated knowledge repository (Granikov, Tang, Bouthillier, & Pluye, 2016). New tools could be added on the MMAT website. Also, continuing data collection on the use of the MMAT could contribute to informing the required modifications and improvement. One possible option to address this issue would be to provide MMAT users with a log in access to a web platform where they could enter their ratings and comments about the tool. This will contribute to continuously collect data on its use and update the MMAT as necessary, making it a 'living' tool.

REFERENCES

Abbott, A. (1998). The causal devolution. *Sociological Methods & Research, 27*(2), 148-181.

Adarkwah, C. C., van Gils, P. F., Hiligsmann, M., & Evers, S. M. (2016). Risk of bias in model-based economic evaluations: The ECOBIAS checklist. *Expert Review of Pharmacoeconomics & Outcomes Research, 16*(4), 513-523.

Al-Jader, L., Newcombe, R., Hayes, S., Murray, A., Layzell, J., & Harper, P. (2002). Developing a quality scoring system for epidemiological surveys of genetic disorders. *Clinical Genetics, 62*(3), 230-234.

Altman, D. G., Simera, I., Hoey, J., Moher, D., & Schulz, K. (2008). EQUATOR: Reporting guidelines for health research. *Open Medicine, 2*(2), e49.

Armijo-Olivo, S., Cummings, G. G., Fuentes, J., Saltaji, H., Ha, C., Chisholm, A., et al. (2014). Identifying items to assess methodological quality in physical therapy trials: A factor analysis. *Physical Therapy, 94*(9), 1272-1284.

Armijo-Olivo, S., Fuentes, J., Ospina, M., Saltaji, H., & Hartling, L. (2013). Inconsistency in the items included in tools used in general health research and physical therapy to evaluate the methodological quality of randomized controlled trials: A descriptive analysis. *BMC Medical Research Methodology, 13*(116), 1-19.

Armola, R. R., Bourgault, A. M., Halm, M. A., Board, R. M., Bucher, L., Harrington, L., et al. (2009). AACN levels of evidence: What's new? *Critical Care Nurse, 29*(4), 70-73.

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., et al. (2004). Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches The GRADE Working Group. *BMC Health Services Research, 4*(38), 1-7.

Bai, A., Shukla, V. K., Bak, G., & Wells, G. (2012). *Quality assessment tools project report*. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health.

Banta, D. (2003). The development of health technology assessment. *Health Policy, 63*(2), 121-132.

Barbour, R. S. (2001). Checklists for improving rigour in qualitative research: A case of the tail wagging the dog? *British Medical Journal, 322*(7294), 1115-1117.

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine, 7*(9), e1000326.

Bennett, C., Khangura, S., Brehaut, J. C., Graham, I. D., Moher, D., Potter, B. K., et al. (2011). Reporting guidelines for survey research: An analysis of published guidance and reporting practices. *PLoS Medicine, 8*(8), e1001069.

Bjork, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research: An International Electronic Journal, 14*(1), 1-14.

Boeije, H. R., van Wesel, F., & Alisic, E. (2011). Making a difference: Towards a method for weighing the evidence in a qualitative synthesis. *Journal of Evaluation in Clinical Practice, 17*(4), 657-663.

Booth, A., Papaioannou, D., & Sutton, A. (2012). *Systematic approaches to a successful literature review*. London: SAGE Publications.

Bouchard, K., Dubuisson, W., Simard, J., & Dorval, M. (2011). Systematic mixed-methods reviews are not ready to be assessed with the available tools. *Journal of Clinical Epidemiology, 64*(8), 926-928.

Boulos, M. N. K., Maramba, I., & Wheeler, S. (2006). Wikis, blogs and podcasts: A new generation of Web-based tools for virtual collaborative clinical practice and education. *BMC Medical Education, 6*(41), 1-8.

Bradshaw, C., Atkinson, S., & Doody, O. (2017). Employing a qualitative description approach in health care research. *Global Qualitative Nursing Research, 4*, 1-8.

Brouwers, M. C., Kho, M. E., Browman, G. P., Burgers, J. S., Cluzeau, F., Feder, G., et al. (2010). AGREE II: Advancing guideline development, reporting and evaluation in health care. *Canadian Medical Association Journal, 182*(18), E839-E842.

Bryant, F. B., & Wortman, P. M. (1984). Methodological issues in the meta-analysis of quasi-experiments. *New Directions for Program Evaluation, 1984*(24), 5-24.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97-113.

Bryman, A., Becker, S., & Sempik, J. (2008). Quality criteria for quantitative, qualitative and mixed methods research: A view from social policy. *International Journal of Social Research Methodology, 11*(4), 261-276.

Bunn, F., Trivedi, D., Alderson, P., Hamilton, L., Martin, A., Pinkney, E., et al. (2015). The impact of Cochrane Reviews: A mixed-methods evaluation of outputs from Cochrane Review Groups supported by the National Institute for Health Research. *Health Technology Assessment, 19*(28), 1-100.

Burls, A. (2009). *What is critical appraisal?* (2nd ed.). Newmarket, UK: Hayward Medical Communications.

Burrows, T. (2013). *A preliminary rubric design to evaluate mixed methods research* (Doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA. Retrieved from https://vtechworks.lib.vt.edu/bitstream/handle/10919/19324/Burrows_T_D_2013.pdf

Busse, R., Orvain, J., Velasco, M., Perleth, M., Drummond, M., Jørgensen, T., et al. (2002). Best practice in undertaking and reporting health technology assessments. *International Journal of Technology Assessment in Health Care, 18*(02), 361-422.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin Company.

Carini, S., Pollock, B. H., Lehmann, H. P., Bakken, S., Barbour, E. M., Gabriel, D., et al. (2009). *Development and evaluation of a study design typology for human research.* Paper presented at the AMIA Annual Symposium Proceedings.

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

Carroll, C., Booth, A., Leaviss, J., & Rick, J. (2013). "Best fit" framework synthesis: Refining the method. *BMC Medical Research Methodology, 13*(37), 1-16.

Chalmers, I. (1993). The Cochrane Collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences, 703*(1), 156-165.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*(1), 12-37.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services*. London: Nuffield Provincial Hospitals Trust.

Colle, F., Rannou, F., Revel, M., Fermanian, J., & Poiraudeau, S. (2002). Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. *Archives of Physical Medicine and Rehabilitation, 83*(12), 1745-1752.

Colorafi, K. J., & Evans, B. (2016). Qualitative descriptive methods in health science research. *Health Environments Research & Design Journal, 9*(4), 16-25.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(1), 104-126.

Cottrell, E., Whitlock, E., Kato, E., Uhl, S., Belinson, S., Chang, C., et al. (2014). *Defining the benefits of stakeholder engagement in systematic reviews. Research white paper*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ).

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Creswell, J. W., & Plano Clark, V. (2018). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Critical Appraisal Skills Programme. (2017a). *10 questions to help you make sense of qualitative research*. Retrieved November 3, 2017, from http://docs.wixstatic.com/ugd/dded87_25658615020e427da194a325e7773d42.pdf

Critical Appraisal Skills Programme. (2017b). *CASP checklists*. Retrieved December 1, 2017, from http://www.casp-uk.net/casp-tools-checklists

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

Curry, L., & Nunez-Smith, M. (2014). *Mixed methods in health sciences research: A practical primer* (Vol. 1). Thousand Oaks, CA: SAGE Publications.

Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M., et al. (2007). A hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology, 60*(1), 43-49.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47*(2), 108-121.

de Vet, H. C., de Bie, R. A., van der Heijden, G. J., Verhagen, A. P., Sijpkes, P., & Knipschild, P. G. (1997). Systematic reviews on the basis of methodological criteria. *Physiotherapy, 83*(6), 284-289.

De Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27), i-186.

Delgado-Rodríguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health, 58*(8), 635-641.

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., et al. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology, 6*(35), 1-13.

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.

Dixon-Woods, M., Sutton, A., Shaw, R., Miller, T., Smith, J., Young, B., et al. (2007). Appraising qualitative research for inclusion in systematic reviews: A quantitative and qualitative comparison of three methods. *Journal of Health Services Research & Policy, 12*(1), 42-47.

Downes, M. J., Brennan, M. L., Williams, H. C., & Dean, R. S. (2016). Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *BMJ Open, 6*(12), e011458.

Dreier, M., Borutta, B., Stahmeyer, J., Krauth, C., & Walter, U. (2010). Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. *GMS Health Technology Assessment, 6*, Doc07 (20100614).

Droitcour, J., Silberman, G., & Chelimsky, E. (1993). Cross-design synthesis: A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. *International Journal of Technology Assessment in Health Care, 9*(03), 440-449.

Dyrvig, A.-K., Kidholm, K., Gerke, O., & Vondeling, H. (2014). Checklists for external validity: A systematic review. *Journal of Evaluation in Clinical Practice, 20*(6), 857-864.

Eckhardt, A. L., & DeVon, H. A. (2017). The MIXED framework: A novel approach to evaluating mixed-methods rigor. *Nursing Inquiry, 24*(4), e12189.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing, 12*(1), 77-84.

Fàbregues, S., & Molina-Azorín, J. F. (2017). Addressing quality in mixed methods research: A review and recommendations for a future agenda. *Quality & Quantity, 51*(6), 2847-2863.

Fàbregues, S., Paré, M.-H., & Meneses, J. (2018). Operationalizing and conceptualizing quality in mixed methods research: A multiple case study of the disciplines of education, nursing, psychology, and sociology. *Journal of Mixed Methods Research, Advance online publication*, https://doi.org/10.1177/1558689817751774.

Faillie, J.-L., Ferrer, P., Gouverneur, A., Driot, D., Berkemeyer, S., Vidal, X., et al. (2017). A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. *Journal of Clinical Epidemiology, 86*(6), 168-175.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs - Principles and practices. *Health Services Research, 48*(6pt2), 2134-2156.

Fetters, M. D., & Freshwater, D. (2015). The 1+ 1= 3 integration challenge. *Journal of Mixed Methods Research, 9*(2), 115-117.

Fitzpatrick-Lewis, D., Ganann, R., Krishnaratne, S., Ciliska, D., Kouyoumdjian, F., & Hwang, S. W. (2011). Effectiveness of interventions to improve the health and housing status of homeless people: A rapid systematic review. *BMC Public Health, 11*(638), 638.

Frantzen, K. K., & Fetters, M. D. (2015). Meta-integration for synthesizing data in a systematic mixed studies review: Insights from research on autism spectrum disorder. *Quality & Quantity, 50*(5), 2251-2277.

Giannakopoulos, N. N., Rammelsberg, P., Eberhard, L., & Schmitter, M. (2012). A new instrument for assessing the quality of studies on prevalence. *Clinical Oral Investigations, 16*(3), 781-788.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Glasziou, P., Vandenbroucke, J., & Chalmers, I. (2004). Assessing the quality of research. *British Medical Journal, 328*(7430), 39-41.

Glenny, A.-M. (2005). No "gold standard" critical appraisal tool for allied health research. *Evidence-Based Dentistry, 6*(4), 100-101.

Goldsmith, M. R., Bankhead, C. R., & Austoker, J. (2007). Synthesising quantitative and qualitative research in evidence-based patient information. *Journal of Epidemiology & Community Health, 61*(3), 262-270.

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews, 1*(28), 1-9.

Grad, R. M., Pluye, P., Mercer, J., Marlow, B., Beauchamp, M.-E., Shulha, M., et al. (2008). Impact of research-based synopses delivered as daily e-mail: A prospective observational study. *Journal of the American Medical Informatics Association, 15*(2), 240-245.

Granikov, V., Tang, D. L., Bouthillier, F., & Pluye, P. (2016). *eSRAP: A system for collaborative monitoring of latest trends in patient oriented research.* Paper presented at the Joint meeting of the Medical Library Association (MLA), the Canadian Health Libraries Association (CHLA), and the International Clinical Librarian Conference (ICLC), Toronto, ON.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Librairies Journal, 26*(2), 91-108.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science & Medicine, 61*(2), 417-430.

Grimes, D. A., & Schulz, K. F. (2002a). Descriptive studies: What they can and cannot do. *The Lancet, 359*(9301), 145-149.

Grimes, D. A., & Schulz, K. F. (2002b). An overview of clinical research: The lay of the land. *The Lancet, 359*(9300), 57-61.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly, 72*(2), 167-189.

Gu, X., & Blackmore, K. L. (2016). Recent trends in academic journal growth. *Scientometrics, 108*(2), 693-716.

Guetterman, T. C., Fetters, M. D., & Creswell, J. W. (2015). Integrating quantitative and qualitative results in health science mixed methods research through joint displays. *The Annals of Family Medicine, 13*(6), 554-561.

Gulmezoglu, A., Chandler, J., Shepperd, S., & Pantoja, T. (2013). Reviews of qualitative evidence: A new milestone for Cochrane. *Cochrane Database of Systematic Reviews, 11*, ED000073.

Guo, B., Moga, C., Harstall, C., & Schopflocher, D. (2016). A principal component analysis is conducted for a case series quality appraisal checklist. *Journal of Clinical Epidemiology, 69*(1), 199-207.

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., et al. (2011a). GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology, 64*(4), 383-394.

Guyatt, G. H., Oxman, A. D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., et al. (2011b). GRADE guidelines 6. Rating the quality of evidence - imprecision. *Journal of Clinical Epidemiology, 64*(12), 1283-1293.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., et al. (2011c). GRADE guidelines: 8. Rating the quality of evidence - indirectness. *Journal of Clinical Epidemiology, 64*(12), 1303-1310.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., et al. (2011d). GRADE guidelines: 7. Rating the quality of evidence - inconsistency. *Journal of Clinical Epidemiology, 64*(12), 1294-1302.

Hannes, K., Heyvaert, M., Slegers, K., Vandenbrande, S., & Van Nuland, M. (2015). Exploring the potential for a consolidated standard for reporting guidelines for qualitative research: An argument Delphi approach. *International Journal of Qualitative Methods, 14*(4), 1-16.

Harden, A., & Gough, D. (2012). Quality and relevance appraisal. In D. Gough, S. Oliver & J. Thomas (Eds.), *An introduction to systematic reviews* (pp. 153-178). London: SAGE Publications.

Hartikainen, S., Lönnroos, E., & Louhivuori, K. (2007). Medication as a risk factor for falls: Critical systematic review. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 62*(10), 1172-1181.

Hartling, L., Bond, K., Harvey, K., Santaguida, P. L., Viswanathan, M., & Dryden, D. M. (2010). *Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ).

Hartling, L., Bond, K., Santaguida, P. L., Viswanathan, M., & Dryden, D. M. (2011). Testing a tool for the classification of study designs in systematic reviews of interventions and exposures showed moderate reliability and low accuracy. *Journal of Clinical Epidemiology, 64*(8), 861-871.

Hartling, L., Hamm, M. P., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., et al. (2013). Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology, 66*(9), 973-981.

Hasson, F., & Keeney, S. (2011). Enhancing rigour in the Delphi technique research. *Technological Forecasting and Social Change, 78*(9), 1695-1704.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing, 32*(4), 1008-1015.

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.

Hayden, J. A., Côté, P., & Bombardier, C. (2006). Evaluation of the quality of prognosis studies in systematic reviews. *Annals of Internal Medicine, 144*(6), 427-437.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Hennekens, C. H., & Buring, J. E. (1987). *Epidemiology in medicine*. Philadelphia: Lippincott Williams and Wilkins.

Herbison, P., Hay-Smith, J., & Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*(12), 1249-1256.

Heyvaert, M., Hannes, K., Maes, B., & Onghena, P. (2013a). Critical appraisal of mixed methods studies. *Journal of Mixed Methods Research, 7*(4), 302-327.

Heyvaert, M., Hannes, K., & Onghena, P. (2016). *Using mixed methods research synthesis for literature reviews: The mixed methods research synthesis approach*. Thousand Oaks, CA: SAGE Publications.

Heyvaert, M., Maes, B., & Onghena, P. (2013b). Mixed methods research synthesis: Definition, framework, and potential. *Quality & Quantity, 47*(2), 659-676.

Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley Online Library.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal, 343*(d5928), 1-9.

Higgins, J. P. T., Sterne, J. A. C., Savović, J., Page, M. J., Hróbjartsson, A., Boutron, I., et al. (2016). A revised tool for assessing risk of bias in randomized trials. In J. Chandler, J. McKenzie, I. Boutron & V. Welch (Eds.), *Cochrane Methods. Cochrane Database of Systematic Reviews*: Issue 10 (Suppl 1).

Hong, Q. N., Gonzalez-Reyes, A., & Pluye, P. (2018). Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). *Journal of Evaluation in Clinical Practice, 24*(3), 459-467.

Hong, Q. N., & Pluye, P. (2018). A conceptual framework for critical appraisal in systematic mixed studies reviews. *Journal of Mixed Methods Research, Advance online publication*, https://doi.org/10.1177/1558689818770058.

Hong, Q. N., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews, 6*(61), 1-14.

Hoonakker, P., & Carayon, P. (2009). Questionnaire survey nonresponse: A comparison of postal mail and internet surveys. *International Journal of Human–Computer Interaction, 25*(5), 348-373.

Hoy, D., Brooks, P., Woolf, A., Blyth, F., March, L., Bain, C., et al. (2012). Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. *Journal of Clinical Epidemiology, 65*(9), 934-939.

Hunt, H., Pollock, A., Campbell, P., Estcourt, L., & Brunton, G. (2018). An introduction to overviews of reviews: Planning a relevant research question and objective for an overview. *Systematic Reviews, 7*(1), 39.

Hunter, A., & Brewer, J. D. (2015). Designing multimethod research. In S. Hesse-Biber & B. Johnson (Eds.), *The Oxford handbook of multimethod and mixed methods research inquiry*. Oxford, UK: Oxford University Press.

Ioannidis, J. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Quarterly, 94*(3), 485-514.

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology, 13*(10), e1002264.

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLoS Biology, 16*(3), e2005468.

Jarde, A., Losilla, J.-M., & Vives, J. (2012). Methodological quality assessment tools of non-experimental studies: A systematic review. *Anales de Psicología, 28*(2), 617-628.

Jarde, A., Losilla, J.-M., Vives, J., & Rodrigo, M. F. (2013). Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology, 13*(2), 138-146.

Jinha, A. E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258-263.

Joanna Briggs Institute. (2017a). *Critical appraisal checklist for prevalence studies*. Retrieved October 24, 2017, from http://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Prevalence_Studies2017.pdf

Joanna Briggs Institute. (2017b). *Critical appraisal checklist for qualitative research*. Retrieved October 24, 2017, from http://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Qualitative_Research2017.pdf

Joanna Briggs Institute. (2017c). *Critical appraisal tools*. Retrieved October 24, 2017, from http://joannabriggs.org/research/critical-appraisal-tools.html

Joanna Briggs Institute. (2017d). *JBI levels of evidence*. Retrieved October 31, 2017, from http://joannabriggs.org/jbi-approach.html#tabbed-nav=Levels-of-Evidence

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

Jones, D., Dixon-Woods, M., Abrams, K., & Fitzpatrick, R. (1999). *Meta-analysis of qualitative and quantitative evidence*. Leicester, UK: University of Leicester.

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

Keeney, S., Hasson, F., & McKenna, H. (2011). *The Delphi technique in nursing and health research*. Chichester, UK: Wiley Online Library.

Khorsan, R., & Crawford, C. (2014). How to assess the external validity and model validity of therapeutic trials: A conceptual approach to systematic review methodology. *Evidence-Based Complementary & Alternative Medicine, 2014*(ID 694804), 1-12.

Kim, H., Sefcik, J. S., & Bradway, C. (2017). Characteristics of qualitative descriptive studies: A systematic review. *Research in Nursing & Health, 40*(1), 23-42.

Kim, S. Y., Park, J. E., Lee, Y. J., Seo, H.-J., Sheen, S.-S., Hahn, S., et al. (2013). Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology, 66*(4), 408-414.

Kmet, L., Lee, R., & Cook, L. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Edmonton, AB: Alberta Heritage Foundation for Medical Research.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lawrence, M., & Kinn, S. (2012). Defining and measuring patient-centred care: An example from a mixed-methods systematic review of the stroke literature. *Health Expectations, 15*(3), 295-326.

Leeman, J., Voils, C., & Sandelowski, M. (2015). Conducting mixed methods literature reviews: Synthesizing the evidence needed to develop and implement complex social and health interventions. In S. Hesse-Biber & B. Johnson (Eds.), *The Oxford handbook of multimethod and mixed methods research inquiry* (pp. 167-184). Oxford; New York: Oxford University Press.

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gülmezoglu, M., et al. (2015). Using qualitative evidence in decision making for health and social interventions: An approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Medicine, 12*(10), e1001895.

Li, G., Abbade, L. P. F., Nwosu, I., Jin, Y., Leenus, A., Maaz, M., et al. (2018). A systematic review of comparisons between protocols or registrations and full reports in primary biomedical research. *BMC Medical Research Methodology, 18*(9), 1-20.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*: Harvard University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review, 41*(4), 429-471.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75). Newbury Park, CA: SAGE Publications.

Lind, J. (1757). *A treatise on the scurvy. In three parts. Containing an inquiry into the nature, causes and cure of that disease. Together with a critical and chronological view of what has been published on the subject*. London: A. Millar in the Strand.

Lomas, J., Culyer, T., McCutcheon, C., McAuley, L., & Law, S. (2005). *Conceptualizing and combining evidence for health system guidance* Ottawa, ON: Canadian Health Services Research Foundation.

Long, A. F., Godfrey, M., Randall, T., Brettle, A. J., & Grant, M. J. (2002). *HCPRDU evaluative tool for mixed methods studies*. Leeds, UK: University of Leeds, Nuffield Institute for Health.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-386.

MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., & Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment, 4*(34), 1-154.

Martin, V., Renaud, J., & Dagenais, P. (2013). *Les normes de production des revues systématiques: Guide méthodologique*. Montréal, QC: Institut national d'excellence en santé et en services sociaux (INESSS).

McKenzie, J. E., & Brennan, S. E. (2017). Overviews of systematic reviews: great promise, greater challenge. *Systematic Reviews, 6*(1), 185.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., et al. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology, 63*(8), e1-e37.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539-549.

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

Morgan, R., Sterne, J., Higgins, J., Thayer, K., Schunemann, H., Rooney, A., et al. (2017). *A new instrument to assess Risk of Bias in Non-randomised Studies of Exposures (ROBINS-E): Application to studies of environmental exposure*. Abstracts of the Global Evidence Summit, Cape Town, South Africa. Cochrane Database of Systematic Reviews 2017, Issue 9 (Suppl 1). https://doi.org/10.1002/14651858.CD201702.

Morley, R. F., Norman, G., Golder, S., & Griffith, P. (2016). A systematic scoping review of the evidence for consumer involvement in organisations undertaking systematic reviews: Focus on Cochrane. *Research Involvement and Engagement, 2*(1), 36.

Moynihan, R. (2004). *Evaluating health services: A reporter covers the science of research synthesis*. New York: Milbank Memorial Fund.

Munn, Z., Moola, S., Riitano, D., & Lisy, K. (2014). The development of a critical appraisal tool for use in systematic reviews: Addressing questions of prevalence. *International Journal of Health Policy & Management, 3*(3), 123-128.

Munn, Z., Stern, C., Aromataris, E., Lockwood, C., & Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology, 18*(5), 1-9.

Munthe-Kaas, H., Bohren, M. A., Glenton, C., Lewin, S., Noyes, J., Tunçalp, Ö., et al. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings - Paper 3: How to assess methodological limitations. *Implementation Science, 13*(9), 1-8.

Munthe-Kaas, H., Lewin, S., & Glenton, C. (2017). *Assessing the methodological strengths and limitations of qualitative evidence: What are the key criteria?* Abstracts of the Global Evidence Summit, Cape Town, South Africa. Cochrane Database of Systematic Reviews 2017, Issue 9 (Suppl 1). https://doi.org/10.1002/14651858.CD201702.

National Health and Medical Research Council. (1998). *A guide to the development, implementation and evaluation of clinical practice guidelines*. Canberra, AU: National Health and Medical Research Council (NHMRC).

National Health and Medical Research Council. (2000). *How to use the evidence: Assessment and application of scientific evidence*. Canberra, AU: National Health and Medical Research Council (NHMRC).

National Institute for Health Care Excellence. (2012). *Methods for the development of NICE public health guidance*. London: National Institute for Health and Care Excellence (NICE).

Neergaard, M. A., Olesen, F., Andersen, R. S., & Sondergaard, J. (2009). Qualitative description – The poor cousin of health research? *BMC Medical Research Methodology, 9*(52), 1-5.

Noblit, G. W., & Hare, R. D. (1988). *Meta-ethnography: Synthesizing qualitative studies* (Vol. 11). Thousand Oaks, CA: SAGE Publications.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw.

O'Cathain, A., Murphy, E., & Nicholl, J. (2008). The quality of mixed methods studies in health services research. *Journal of Health Services Research & Policy, 13*(2), 92-98.

O'Cathain, A., Thomas, K. J., Drabble, S. J., Rudolph, A., & Hewison, J. (2013). What can qualitative research do for randomised controlled trials? A systematic mapping review. *BMJ Open, 3*(6), e002889.

Office of Health Technology Assessment. (1982). *MEDLARS and health information policy: A technical memorandum*. Washington, DC: US Government Printing Office.

Olsson, L. E., Jakobsson Ung, E., Swedberg, K., & Ekman, I. (2013). Efficacy of person-centred care as an intervention in controlled trials – A systematic review. *Journal of Clinical Nursing, 22*(3-4), 456-465.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13*(1), 48-63.

Ostlund, U., Kidd, L., Wengstrom, Y., & Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: A methodological review. *International Journal of Nursing Studies, 48*(3), 369-383.

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., et al. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies, 49*(1), 47-53.

Pai, M., & Filion, K. (2014). *Teach epidemiology: Classification of study designs (version 8)*. Retrieved December 15, 2017, from http://www.teachepi.org/documents/courses/Classification%20Design.pdf

Pai, M., McCulloch, M., Gorman, J. D., Pai, N., Enanoria, W., Kennedy, G., et al. (2003). Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *The National Medical Journal of India, 17*(2), 86-95.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management, 52*(2), 183-199.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review - A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy, 10* (Suppl 1), 21-34.

Pearson, A., White, H., Bath-Hextall, F., Apostolo, J., Salmond, S., & Kirkpatrick, P. (2014). *Methodology for JBI mixed methods systematic reviews The Joanna Briggs Institute Reviewers Manual* (pp. 5-34). Adelaide, AU: Joanna Briggs Institute.

Peirson, L., Catallo, C., & Chera, S. (2013). The Registry of Knowledge Translation Methods and Tools: A resource to support evidence-informed public health. *International Journal of Public Health, 58*(4), 493-500.

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P., Mayhew, A., Pantoja, T., et al. (2013). Synthesizing evidence on complex interventions: How meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of Clinical Epidemiology, 66*(11), 1230-1243.

Petticrew, M., & Roberts, H. (2006). How to appraise the studies: An introduction to assessing study quality. In M. Petticrew & H. Roberts (Eds.), *Systematic reviews in the social sciences: A practical guide* (pp. 125-163). Padstow, UK: Wiley-Blackwell.

Pill, J. (1971). The Delphi method: Substance, context, a critique and an annotated bibliography. *Socio-Economic Planning Sciences, 5*(1), 57-71.

Pincus, T., Miles, C., Froud, R., Underwood, M., Carnes, D., & Taylor, S. J. (2011). Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: A consensus study. *BMC Medical Research Methodology, 11*(14), 1-14.

Plano Clark, V. L., & Ivankova, N. V. (2015). *Mixed methods research: A guide to the field*. Thousand Oaks, CA: SAGE Publications.

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.

Pluye, P., Garcia Bengoechea, E., Granikov, V., Kaur, N., & Tang, D. L. (2018). A world of possibilities in mixed methods: Review of the combinations of strategies used to integrate the phases, results, and qualitative and quantitative data. *International Journal of Multiple Research Approaches, 10*(1), 41-56.

Pluye, P., Grad, R., Granikov, V., Theriault, G., Fremont, P., Burnand, B., et al. (2012). Feasibility of a knowledge translation CME program: Courriels Cochrane. *Journal of Continuing Education in the Health Professions, 32*(2), 134-141.

Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health, 35*, 29-45.

Pluye, P., Hong, Q. N., Bush, P. L., & Vedel, I. (2016). Opening-up the definition of systematic literature review: The plurality of worldviews, methodologies and methods for reviews and syntheses. *Journal of Clinical Epidemiology, 73*(5), 2-5.

Pluye, P., Robert, E., Cargo, M., Bartlett, G., O'Cathain, A., Griffiths, F., et al. (2011). *Proposal: A Mixed Methods Appraisal Tool for systematic mixed studies reviews*. Retrieved November 15, 2013, from http://mixedmethodsappraisaltoolpublic.pbworks.com

Polit, D. F., & Beck, C. T. (2014). Supplment for Chapter 14 - Qualitative descriptive studies. In D. F. Polit & C. T. Beck (Eds.), *Essentials of nursing research: Appraising evidence for nursing practice* (8th ed.). Philadelphia: Lippincott Williams & Wilkins.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459-467.

Pope, C., Mays, N., & Popay, J. (2007). Mixed approaches to evidence synthesis. In C. Pope, N. Mays & J. Popay (Eds.), *Synthesizing qualitative and quantitative health evidence : A guide to methods* (pp. 95-114). New York: Open University Press, McGraw Hill Education.

Powell, A., Rushmer, R., & Davies, H. (2009). *A systematic narrative review of quality improvement models in health care*: NHS Quality Improvement Scotland.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *Journal of the American Medical Association, 298*(9), 1002-1009.

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.

Robert, E. (2015). *Mixed Methods Appraisal Tool pour l'évaluation de la qualité des études de la revue réaliste*. Unpublished document. Available at http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/108871951/MMAT%20traduction%20FR_Emilie-Robert_2015.pdf.

Robinson, K. A., Chou, R., Berkman, N. D., Newberry, S. J., Fu, R., Hartling, L., et al. (2016). Twelve recommendations for integrating existing systematic reviews into new reviews: EPC guidance. *Journal of Clinical Epidemiology, 70*(2), 38-44.

Rowe, G., Wright, G., & Bolger, F. (1991). Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change, 39*(3), 235-251.
Sackett, D. (1979). Bias in analytic research. *Journal of Chronic Diseases, 32*(1-2), 51-63.
Sackett, D. L., Rosenberg, W. M., Gray, J., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*(7023), 71-72.
Sale, J. E. M., & Brazil, K. (2004). A strategy to identify critical appraisal criteria for primary mixed-method studies. *Quality & Quantity, 38*(4), 351-365.
Sandelowski, M. (2000). Focus on research methods - Whatever happened to qualitative description? *Research in Nursing & Health, 23*(4), 334-340.
Sandelowski, M. (2010). What's in a name? Qualitative description revisited. *Research in Nursing & Health, 33*(1), 77-84.
Sandelowski, M., & Barroso, J. (2002). Reading qualitative studies. *International Journal of Qualitative Methods, 1*(1), 74-108.
Sandelowski, M., Docherty, S., & Emden, C. (1997). Focus on qualitative methods - Qualitative metasynthesis: Issues and techniques. *Research in Nursing and Health, 20*, 365-372.
Sandelowski, M., Voils, C. I., & Barroso, J. (2006). Defining and designing mixed research synthesis studies. *Research in the Schools, 13*(1), 29-40.
Sandelowski, M., Voils, C. I., Leeman, J., & Crandell, J. L. (2012). Mapping the mixed methods–mixed research synthesis terrain. *Journal of Mixed Methods Research, 6*(4), 317-331.
Santiago-Delefosse, M., Gavin, A., Bruchez, C., Roux, P., & Stephen, S. (2016). Quality of qualitative research in the health sciences: Analysis of the common criteria present in 58 assessment guidelines by expert users. *Social Science & Medicine, 148*(1), 142-151.
Saunders, L. D., Soomro, G. M., Buckingham, J., Jamtvedt, G., & Raina, P. (2003). Assessing the methodological quality of nonrandomized intervention studies. *Western Journal of Nursing Research, 25*(2), 223-237.
Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy, 2*(4), 419-436.
Schryen, G., Wagner, G., & Benlian, A. (2015). *Theory of knowledge for literature reviews: An epistemological model, taxonomy and empirical analysis of IS literature*. Paper presented at the Thirty Sixth International Conference on Information Systems, Forth Worth, TX.
Schuerman, J., Soydan, H., Macdonald, G., Forslund, M., de Moya, D., & Boruch, R. (2002). The Campbell collaboration. *Research on Social Work Practice, 12*(2), 309-317.
Scottish Intercollegiate Guidelines Network. (2017a). *Algorithm for classifying study design for questions of effectiveness*. Retrieved December 1, 2017, from http://www.sign.ac.uk/assets/study_design.pdf
Scottish Intercollegiate Guidelines Network. (2017b). *Critical appraisal notes and checklists*. Retrieved December 1, 2017, from http://www.sign.ac.uk/checklists-and-notes.html
Seo, H.-J., Kim, S. Y., Lee, Y. J., Jang, B.-H., Park, J.-E., Sheen, S.-S., et al. (2016). A newly developed tool for classifying study designs in systematic reviews of interventions and exposures showed substantial reliability and validity. *Journal of Clinical Epidemiology, 70*(2), 200-205.
Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology, 63*(10), 1061-1070.

Shamliyan, T. A., Kane, R. L., Ansari, M. T., Raman, G., Berkman, N. D., Grant, M., et al. (2011). Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: Pilot study of new checklists. *Journal of Clinical Epidemiology, 64*(6), 637-657.

Shaw, R. L., Larkin, M., & Flowers, P. (2014). Expanding the evidence within evidence-based healthcare: Thinking about the context, acceptability and feasibility of interventions. *Evidence-Based Medicine, 19*(6), 201-203.

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., et al. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 62*(10), 1013-1020.

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.

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice, 18*(4), 746-752.

Slim, K., Nini, E., Forestier, D., Kwiatkowski, F., Panis, Y., & Chipponi, J. (2003). Methodological index for non-randomized studies (MINORS): Development and validation of a new instrument. *ANZ Journal of Surgery, 73*(9), 712-716.

Souto, R. Q., Khanassov, V., Hong, Q. N., Bush, P. L., Vedel, I., & Pluye, P. (2015). Systematic mixed studies reviews: Updating results on the reliability and efficiency of the Mixed Methods Appraisal Tool. *International Journal of Nursing Studies, 52*(1), 500-501.

Souto, R. Q., Lima, K. S. d. A., Pluye, P., Hong, Q. N., Barbosa, K., & Djogovic, T. (submitted). Tradução e adaptação transcultural do instrumento Mixed Methods Appraisal Tool ao contexto brasileiro. *Revista da Escola de Enfermagem da USP/Journal of School of Nursing - University of São Paulo*.

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*. London: Government Chief Social Researcher's Office.

Sterne, J. A. (2013). Why the Cochrane risk of bias tool should not include funding source as a standard item. *Cochrane Database of Systematic Reviews, 12*, ED000076.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal, 355*(i4919).

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions, 25*(1), 76-97.

Sudsawad, P. (2007). *Knowledge translation: Introduction to models, strategies, and measures*. Austin, TX: Southwest Educational Development Laboratory, National Center for the Dissemination of Disability Research.

Sutton, A. J., Jones, D. R., Abrams, K. R., Sheldon, T. A., & Song, F. (1999). Systematic reviews and meta-analysis: A structured review of the methodological literature. *Journal of Health Services Research & Policy, 4*(1), 49-55.

Suzuki, E., Tsuda, T., Mitsuhashi, T., Mansournia, M. A., & Yamamoto, E. (2016). Errors in causal inference: An organizational schema for systematic error and random error. *Annals of Epidemiology, 26*(11), 788-793.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., et al. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*(5), 619-638.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: SAGE Publications.

The Cochrane Collaboration. (2017). *About us*. Retrieved June 18, 2017, from http://www.cochrane.org/ca/about-us

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176-184.

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods, 2*(1), 1-14.

Thorne, S. (2008). *Interpretive description* (Vol. 2). Walnut Creek, CA: Left Coast Press.

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.

Tricco, A. C., Cardoso, R., Thomas, S. M., Motiwala, S., Sullivan, S., Kealey, M. R., et al. (2015). Barriers and facilitators to uptake of systematic reviews by policy makers and health care managers: A scoping review. *Implementation Science, 11*(4), 1-20.

Tricco, A. C., Langlois, E. V., & Straus, S. E. (2017). *Rapid reviews to strengthen health policy and systems: A practical guide*. Geneva: World Health Organization.

Tricco, A. C., Tetzlaff, J., & Moher, D. (2011). The art and science of knowledge synthesis. *Journal of Clinical Epidemiology, 64*(1), 11-20.

U.S. National Library of Medicine. (2013, February 20). *Fact sheet - MEDLINE®*. Retrieved April 5, 2014, from http://www.nlm.nih.gov/pubs/factsheets/medline.html

Vale, C. L., Tierney, J. F., & Burdett, S. (2013). Can trial quality be reliably assessed from published reports of cancer trials: Evaluation of risk of bias assessments in systematic reviews. *British Medical Journal, 346*(f1798), 1-10.

van Hoorn, R., Tummers, M., Booth, A., Gerhardus, A., Rehfuess, E., Hind, D., et al. (2017). The development of CHAMP: A checklist for the appraisal of moderators and predictors. *BMC Medical Research Methodology, 17*(173), 1-9.

Van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., & Editorial Board of the Cochrane Collaboration Back Review Group. (2003). Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine, 28*(12), 1290-1299.

Vedam, S., Rossiter, C., Homer, C. S., Stoll, K., & Scarf, V. L. (2017). The ResQu Index: A new instrument to appraise the quality of research on birth place. *PLoS One, 12*(8), e0182991.

Verhagen, A. P., de Vet, H. C., de Bie, R. A., Kessels, A. G., Boers, M., Bouter, L. M., et al. (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.

Vermeire, E., Van Royen, P., Griffiths, F., Coenen, S., Peremans, L., & Hendrickx, K. (2002). The critical appraisal of focus group research articles. *European Journal of General Practice, 8*(3), 104-108.

Vetter, T. R., & Mascha, E. J. (2017). Bias, confounding, and interaction: Lions and tigers, and bears, oh my! *Anesthesia & Analgesia, 125*(3), 1042-1048.

Victora, C. G., Habicht, J.-P., & Bryce, J. (2004). Evidence-based public health: Moving beyond randomized trials. *American Journal of Public Health, 94*(3), 400-405.

Viswanathan, M., Ansari, M. T., Berkman, N. D., Chang, S., Hartling, L., McPheeters, M., et al. (2012). *Assessing the risk of bias of individual studies in systematic reviews of health care interventions*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ) Methods Guide for Comparative Effectiveness Reviews.

Viswanathan, M., & Berkman, N. D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology, 65*(2), 163-178.

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting and Social Change, 79*(8), 1525-1536.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., Vandenbroucke, J. P., et al. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology, 61*(4), 344-349.

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*. The Netherlands: STM: International Association of Scientific, Technical and Medical Publishers.

Wells, K., & Littell, J. H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice, 19*(1), 52-62.

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*(2), 235-268.

West, S. L., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., et al. (2002). *Systems to rate the strength of scientific evidence*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ).

Whiting, P., Savovic, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., et al. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology, 69*(1), 225-234.

Whiting, P., Wolff, R., Mallett, S., Simera, I., & Savović, J. (2017). A proposed framework for developing quality assessment tools. *Systematic Reviews, 6*(204), 1-9.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., et al. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*(8), 529-536.

Whittemore, R., Chao, A., Jang, M., Minges, K. E., & Park, C. (2014). Methods for knowledge synthesis: An overview. *Heart & Lung, 43*(5), 453-461.

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing, 52*(5), 546-553.

Wilson, P. M., Petticrew, M., Calnan, M. W., & Nazareth, I. (2010). Disseminating research findings: What should researchers do? A systematic scoping review of conceptual frameworks. *Implementation Science, 5*(91), 1-16.

Wolff, R., Moons, K., Riley, R., Whiting, P., Westwood, M., Collins, G., et al. (2017). *PROBAST – A risk-of-bias tool for prediction-modelling studies*. Abstracts of the Global Evidence Summit, Cape Town, South Africa. Cochrane Database of Systematic Reviews 2017, Issue 9 (Suppl 1). https://doi.org/10.1002/14651858.CD201702.

Wortman, P. M. (1994). Judging research quality. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 97-109). New York: Russel Sage Foundation.

Yang, A. W., Li, C. G., Da Costa, C., Allan, G., Reece, J., & Xue, C. C. (2009). Assessing quality of case series studies: Development and validation of an instrument by herbal medicine CAM researchers. *Journal of Alternative and Complementary Medicine, 15*(5), 513-522.

Yates, S. L., Morley, S., Eccleston, C., & Williams, A. C. d. C. (2005). A scale for rating the quality of psychological trials for pain. *Pain, 117*(3), 314-325.

Yin, R. K., & Heald, K. A. (1975). Using the case survey method to analyze policy studies. *Administrative Science Quarterly, 20*(3), 371-381.

Zaza, S., Wright-De Agüero, L. K., Briss, P. A., Truman, B. I., & Hopkins, D. P. (2000). Data collection instrument and procedure for systematic reviews in the guide to community preventive services. *American Journal of Preventive Medicine, 18*(Suppl 1), 44-74.

Zhao, S. (1991). Metatheory, metamethod, meta-data-analysis: What, why, and how? *Sociological Perspectives, 34*(3), 377-390.

APPENDIX 1. MIXED METHODS APPRAISAL TOOL (MMAT), VERSION 2011

Mixed Methods Appraisal Tool (MMAT) – Version 2011

For dissemination, application, and feedback: Please contact pierre.pluye@mcgill.ca, Department of Family Medicine, McGill University, Canada.

The MMAT is comprised of two parts (see below): criteria (Part I) and tutorial (Part II). While the content validity and the reliability of the pilot version of the MMAT have been examined, this critical appraisal tool is still in development. Thus, the MMAT must be used with caution, and users' feedback is appreciated. Cite the present version as follows.

> Pluye, P., Robert, E., Cargo, M., Bartlett, G., O'Cathain, A., Griffiths, F., Boardman, F., Gagnon, M.P., & Rousseau, M.C. (2011). *Proposal: A mixed methods appraisal tool for systematic mixed studies reviews.* Retrieved on [date] from http://mixedmethodsappraisaltoolpublic.pbworks.com. Archived by WebCite® at http://www.webcitation.org/5tTRTc9yJ.

Purpose: The MMAT has been designed for the appraisal stage of complex systematic literature reviews that include qualitative, quantitative and mixed methods studies (mixed studies reviews). The MMAT permits to concomitantly appraise and describe the methodological quality for three methodological domains: mixed, qualitative and quantitative (subdivided into three sub-domains: randomized controlled, non-randomized, and descriptive). Therefore, using the MMAT requires experience or training in these domains. E.g., MMAT users may be helped by a colleague with specific expertise when needed. The MMAT allows the appraisal of most common types of study methodology and design. For appraising a qualitative study, use section 1 of the MMAT. For a quantitative study, use section 2 or 3 or 4, for randomized controlled, non-randomized, and descriptive studies, respectively. For a mixed methods study, use section 1 for appraising the qualitative component, the appropriate section for the quantitative component (2 or 3 or 4), and section 5 for the mixed methods component. For each relevant study selected for a systematic mixed studies review, the methodological quality can then be described using the corresponding criteria. This may lead to exclude studies with lowest quality from the synthesis, or to consider the quality of studies for contrasting their results (e.g., low quality vs. high).

Scoring metrics: For each retained study, an overall quality score may not be informative (in comparison to a descriptive summary using MMAT criteria), but might be calculated using the MMAT. Since there are only a few criteria for each domain, the score can be presented using descriptors such as *, **, ***, and ****. For qualitative and quantitative studies, this score can be the number of criteria met divided by four (scores varying from 25% (*) -one criterion met- to 100% (****) -all criteria met-). For mixed methods research studies, the premise is that the overall quality of a combination cannot exceed the quality of its weakest component. Thus, the overall quality score is the lowest score of the study components. The score is 25% (*) when *QUAL=1 or QUAN=1 or MM=0*; it is

50% (**) when *QUAL=2 or QUAN=2 or MM=1*; it is 75% (***) when *QUAL=3 or QUAN=3 or MM=2*; and it is 100% (****) when *QUAL=4 and QUAN=4 and MM=3* (QUAL being the score of the qualitative component; QUAN the score of the quantitative component; and MM the score of the mixed methods component).

Rationale: There are general criteria for planning, designing and reporting mixed methods research (Creswell and Plano Clark, 2010), but there is no consensus on key specific criteria for appraising the methodological quality of mixed methods studies (O'Cathain, Murphy and Nicholl, 2008). Based on a critical examination of 17 health-related systematic mixed studies reviews, an initial 15-criteria version of MMAT was proposed (Pluye, Gagnon, Griffiths and Johnson-Lafleur, 2009). This was pilot tested in 2009. Two raters assessed 29 studies using the pilot MMAT criteria and tutorial (Pace, Pluye, Bartlett, Macaulay et al., 2010). Based on this pilot exercise, it is anticipated that applying MMAT may take on average 15 minutes per study (hence efficient), and that the Intra-Class Correlation might be around 0.8 (hence reliable). The present 2011 revision is based on feedback from four workshops, and a comprehensive framework for assessing the quality of mixed methods research (O'Cathain, 2010).

Conclusion: The MMAT has been designed to appraise the *methodological quality* of the studies retained for a systematic mixed studies review, not the quality of their *reporting* (writing). This distinction is important, as good research may not be 'well' reported. If reviewers want to genuinely assess the former, companion papers and research reports should be collected when some criteria are not met, and authors of the corresponding publications should be contacted for additional information. Collecting additional data is usually necessary to appraise *qualitative research and mixed methods studies*, as there are no uniform standards for reporting study characteristics in these domains (www.equator-network.org), in contrast, e.g., to the CONSORT statement for reporting randomized controlled trials (www.consort-statement.org).

Authors and contributors: Pierre Pluye[1], Marie-Pierre Gagnon[2], Frances Griffiths[3] and Janique Johnson-Lafleur[1] proposed an initial version of MMAT criteria (Pluye et al., 2009). Romina Pace[1] and Pierre Pluye[1] led the pilot test. Gillian Bartlett[1], Belinda Nicolau[4], Robbyn Seller[1], Justin Jagosh[1], Jon Salsberg[1] and Ann Macaulay[1] contributed to the pilot work (Pace et al., 2010). Pierre Pluye[1], Émilie Robert[5], Margaret Cargo[6], Alicia O'Cathain[7], Frances Griffiths[3], Felicity Boardman[3], Marie-Pierre Gagnon[2], Gillian Bartlett[1], and Marie-Claude Rousseau[8] contributed to the present 2011 version.

Affiliations: 1. Department of Family Medicine, McGill University, Canada; 2. Faculté des sciences infirmières, Université Laval, Canada; 3. Warwick Medical School, University of Warwick, UK; 4. Faculty of Dentistry, McGill University, Canada; 5. Centre de recherche du CHUM, Université de Montréal, Canada; 6. School of Health Sciences, University of South Australia, Australia; 7. Medical Care Research Unit, ScHARR, University of Sheffield, UK; 8. INRS-Institut Armand Frappier, Laval, Canada.

PART I. MMAT criteria & one-page template (to be included in appraisal forms)

Types of mixed methods study components or primary studies	Methodological quality criteria (see tutorial for definitions and examples)	Responses			
		Yes	No	Can't tell	Comments
Screening questions (for all types)	• Are there clear qualitative and quantitative research questions (or objectives*), or a clear mixed methods question (or objective*)?				
	• Do the collected data allow address the research question (objective)? E.g., consider whether the follow-up period is long enough for the outcome to occur (for longitudinal studies or study components).				
	Further appraisal may not be feasible or appropriate when the answer is 'No' or 'Can't tell' to one or both screening questions.				
1. Qualitative	1.1. Are the sources of qualitative data (archives, documents, informants, observations) relevant to address the research question (objective)?				
	1.2. Is the process for analyzing qualitative data relevant to address the research question (objective)?				
	1.3. Is appropriate consideration given to how findings relate to the context, e.g., the setting, in which the data were collected?				
	1.4. Is appropriate consideration given to how findings relate to researchers' influence, e.g., through their interactions with participants?				
2. Quantitative randomized controlled (trials)	2.1. Is there a clear description of the randomization (or an appropriate sequence generation)?				
	2.2. Is there a clear description of the allocation concealment (or blinding when applicable)?				
	2.3. Are there complete outcome data (80% or above)?				
	2.4. Is there low withdrawal/drop-out (below 20%)?				
3. Quantitative non-randomized	3.1. Are participants (organizations) recruited in a way that minimizes selection bias?				
	3.2. Are measurements appropriate (clear origin, or validity known, or standard instrument; and absence of contamination between groups when appropriate) regarding the exposure/intervention and outcomes?				

	3.3. In the groups being compared (exposed vs. non-exposed; with intervention vs. without; cases vs. controls), are the participants comparable, or do researchers take into account (control for) the difference between these groups?				
	3.4. Are there complete outcome data (80% or above), and, when applicable, an acceptable response rate (60% or above), or an acceptable follow-up rate for cohort studies (depending on the duration of follow-up)?				
4. Quantitative descriptive	4.1. Is the sampling strategy relevant to address the quantitative research question (quantitative aspect of the mixed methods question)?				
	4.2. Is the sample representative of the population understudy?				
	4.3. Are measurements appropriate (clear origin, or validity known, or standard instrument)?				
	4.4. Is there an acceptable response rate (60% or above)?				
5. Mixed methods	5.1. Is the mixed methods research design relevant to address the qualitative and quantitative research questions (or objectives), or the qualitative and quantitative aspects of the mixed methods question (or objective)?				
	5.2. Is the integration of qualitative and quantitative data (or results*) relevant to address the research question (objective)?				
	5.3. Is appropriate consideration given to the limitations associated with this integration, e.g., the divergence of qualitative and quantitative data (or results*) in a triangulation design?				
	Criteria for the qualitative component (1.1 to 1.4), and appropriate criteria for the quantitative component (2.1 to 2.4, or 3.1 to 3.4, or 4.1 to 4.4), must also be applied.				

*These two items are not considered as double-barreled items since in mixed methods research, (1) there may be research questions (quantitative research) or research objectives (qualitative research), and (2) data may be integrated, and/or qualitative findings and quantitative results can be integrated.

PART II. MMAT tutorial

Types of mixed methods study components or primary studies	Methodological quality criteria
1. Qualitative Common types of qualitative research methodology include: A. Ethnography The aim of the study is to describe and interpret the shared cultural behaviour of a group of individuals. B. Phenomenology The study focuses on the subjective experiences and interpretations of a phenomenon encountered by individuals. C. Narrative The study analyzes life experiences of an individual or a group. D. Grounded theory Generation of theory from data in the process of conducting research (data collection occurs first). E. Case study In-depth exploration and/or explanation of issues intrinsic to a particular case. A case can be anything from a decision-making process, to a person, an organization, or a country.	1.1. Are the sources of qualitative data (archives, documents, informants, observations) relevant to address the research question (objective)? E.g., consider whether (a) the selection of the participants is clear, and appropriate to collect relevant and rich data; and (b) reasons why certain potential participants chose not to participate are explained.
	1.2. Is the process for analyzing qualitative data relevant to address the research question (objective)? E.g., consider whether (a) the method of data collection is clear (in depth interviews and/or group interviews, and/or observations and/or documentary sources); (b) the form of the data is clear (tape recording, video material, and/or field notes for instance); (c) changes are explained when methods are altered during the study; and (d) the qualitative data analysis addresses the question.
	1.3. Is appropriate consideration given to how findings relate to the context, e.g., the setting, in which the data were collected? E.g., consider whether the study context and how findings relate to the context or characteristics of the context are explained (how findings are influenced by or influence the context). "For example, a researcher wishing to observe care in an acute hospital around the clock may not be able to study more than one hospital. (…) Here, it is essential to take care to describe the context and particulars of the case [the hospital] and to flag up for the reader the similarities and differences between the case and other settings of the same type" (Mays & Pope, 1995). The notion of context may be conceived in different ways depending on the approach (methodology) tradition.

F. Qualitative description There is no specific methodology, but a qualitative data collection and analysis, e.g., in-depth interviews or focus groups, and hybrid thematic analysis (inductive and deductive). Key references: Creswell, 1998; Schwandt, 2001; Sandelowski, 2010.	1.4. Is appropriate consideration given to how findings relate to researchers' influence, e.g., through their interactions with participants? E.g., consider whether (a) researchers critically explain how findings relate to their perspective, role, and interactions with participants (how the research process is influenced by or influences the researcher); (b) researcher's role is influential at all stages (formulation of a research question, data collection, data analysis and interpretation of findings); and (c) researchers explain their reaction to critical events that occurred during the study. The notion of reflexivity may be conceived in different ways depending on the approach (methodology) tradition. E.g., "at a minimum, researchers employing a generic approach [qualitative description] must explicitly identify their disciplinary affiliation, what brought them to the question, and the assumptions they make about the topic of interest" (Caelli, Ray & Mill, 2003, p. 5).

Types of mixed methods study components or primary studies	Methodological quality criteria
2. Quantitative randomized controlled (trials) Randomized controlled clinical trial: A clinical study in which individual participants are allocated to intervention or control groups by randomization (intervention assigned by researchers). Key references: Higgins & Green, 2008; Porta, 2008; Oxford Center for Evidence based medicine, 2009.	2.1. Is there a clear description of the randomization (or an appropriate sequence generation)? In a randomized controlled trial, the allocation of a participant (or a data collection unit, e.g., a school) into the intervention or control group is based solely on chance, and researchers describe how the randomization schedule is generated. "A simple statement such as 'we randomly allocated' or 'using a randomized design' is insufficient". *Simple randomization:* Allocation of participants to groups by chance by following a predetermined plan/sequence. "Usually it is achieved by referring to a published list of random numbers, or to a list of random assignments generated by a computer". *Sequence generation:* "The rule for allocating interventions to participants must be specified, based on some chance (random) process". Researchers provide sufficient detail to allow a readers' appraisal of whether it produces comparable groups. E.g., blocked randomization (to ensure particular allocation ratios to the intervention groups), or stratified randomization (randomization performed separately within strata), or minimization (to make small groups closely similar with respect to several characteristics).
	2.2. Is there a clear description of the allocation concealment (or blinding when applicable)? *The allocation concealment protects assignment sequence until allocation.* E.g., researchers and participants are unaware of the assignment sequence up to the point of allocation. E.g., group assignment is concealed in opaque envelops until allocation. *The blinding protects assignment sequence after allocation.* E.g., researchers and/or participants are unaware of the group a participant is allocated to during the course of the study.
	2.3. Are there complete outcome data (80% or above)? E.g., almost all the participants contributed to almost all measures.
	2.4. Is there low withdrawal/drop-out (below 20%)? E.g., almost all the participants completed the study.

Types of mixed methods study components or primary studies	Methodological quality criteria
3. Quantitative non-randomized Common types of design include (A) non-randomized controlled trials, and (B-C-D) observational analytic study or component where the intervention/exposure is defined/assessed, but not assigned by researchers. A. Non-randomized controlled trials The intervention is assigned by researchers, but there is no randomization, e.g., a pseudo-randomization. A non-random method of allocation is not reliable in producing alone similar groups. B. Cohort study Subsets of a defined population are assessed as exposed, not exposed, or exposed at different degrees to factors of interest. Participants are followed over time to determine if an outcome occurs (prospective longitudinal). C. Case-control study Cases, e.g., patients, associated with a certain outcome are selected, alongside a corresponding group of controls. Data is collected on whether cases and controls were exposed to the factor under study (retrospective). D. Cross-sectional analytic study At one particular time, the relationship between health-related characteristics (outcome) and other factors (intervention/exposure) is examined. E.g., the frequency of outcomes is compared in different population sub-groups according to the presence/absence (or level) of the	3.1. Are participants (organizations) recruited in a way that minimizes selection bias? At recruitment stage: For cohort studies, e.g., consider whether the exposed (or with intervention) and non-exposed (or without intervention) groups are recruited from the same population. For case-control studies, e.g., consider whether same inclusion and exclusion criteria were applied to cases and controls, and whether recruitment was done independently of the intervention or exposure status. For cross-sectional analytic studies, e.g., consider whether the sample is representative of the population. 3.2. Are measurements appropriate (clear origin, or validity known, or standard instrument; and absence of contamination between groups when appropriate) regarding the exposure/intervention and outcomes? At data collection stage: E.g., consider whether (a) the variables are clearly defined and accurately measured; (b) the measurements are justified and appropriate for answering the research question; and (c) the measurements reflect what they are supposed to measure. For non-randomized controlled trials, the intervention is assigned by researchers, and so consider whether there was absence/presence of a contamination. E.g., the control group may be indirectly exposed to the intervention through family or community relationships.

intervention/exposure. Key references for observational analytic studies: Higgins & Green, 2008; Wells, Shea, O'Connell, Peterson, et al., 2009.	3.3. In the groups being compared (exposed vs. non-exposed; with intervention vs. without; cases vs. controls), are the participants comparable, or do researchers take into account (control for) the difference between these groups? At data analysis stage: For cohort, case-control and cross-sectional, e.g., consider whether (a) the most important factors are taken into account in the analysis; (b) a table lists key demographic information comparing both groups, and there are no obvious dissimilarities between groups that may account for any differences in outcomes, or dissimilarities are taken into account in the analysis.
	3.4. Are there complete outcome data (80% or above), and, when applicable, an acceptable response rate (60% or above), or an acceptable follow-up rate for cohort studies (depending on the duration of follow-up)?

Types of mixed methods study components or primary studies	Methodological quality criteria
4. Quantitative descriptive studies Common types of design include single-group studies: A. Incidence or prevalence study without comparison group In a defined population at one particular time, what is happening in a population, e.g., frequencies of factors (importance of problems), is described (portrayed). B. Case series A collection of individuals with similar characteristics are used to describe an outcome. C. Case report An individual or a group with a unique/unusual outcome is described in details. Key references: Critical Appraisal Skills Programme, 2009; Draugalis, Coons & Plaza, 2008.	4.1. Is the sampling strategy relevant to address the quantitative research question (quantitative aspect of the mixed methods question)? E.g., consider whether (a) the source of sample is relevant to the population under study; (b) when appropriate, there is a standard procedure for sampling, and the sample size is justified (using power calculation for instance).
	4.2. Is the sample representative of the population understudy? E.g., consider whether (a) inclusion and exclusion criteria are explained; and (b) reasons why certain eligible individuals chose not to participate are explained.
	4.3. Are measurements appropriate (clear origin, or validity known, or standard instrument)? E.g., consider whether (a) the variables are clearly defined and accurately measured; (b) measurements are justified and appropriate for answering the research question; and (c) the measurements reflect what they are supposed to measure.
	4.4. Is there an acceptable response rate (60% or above)? The response rate is not pertinent for case series and case report. E.g., there is no expectation that a case series would include all patients in a similar situation.

Types of mixed methods study components or primary studies	Methodological quality criteria
5. Mixed methods Common types of design include: A. Sequential explanatory design The quantitative component is followed by the qualitative. The purpose is to explain quantitative results using qualitative findings. E.g., the quantitative results guide the selection of qualitative data sources and data collection, and the qualitative findings contribute to the interpretation of quantitative results. B. Sequential exploratory design The qualitative component is followed by the quantitative. The purpose is to explore, develop and test an instrument (or taxonomy), or a conceptual framework (or theoretical model). E.g., the qualitative findings inform the quantitative data collection, and the quantitative results allow a generalization of the qualitative findings. C. Triangulation design The qualitative and quantitative components are concomitant. The purpose is to examine the same phenomenon by interpreting qualitative and quantitative results (bringing data analysis together at the interpretation stage), or by integrating qualitative and quantitative datasets (e.g., data on same cases), or by transforming data (e.g., quantization of qualitative data). D. Embedded design The qualitative and quantitative components are concomitant. The purpose is to support a qualitative study with a quantitative sub-study (measures), or to better understand a specific issue of a quantitative study using a qualitative sub-study, e.g., the efficacy or the implementation of an intervention based on the views of participants. Key references: Creswell & Plano Clark, 2007; O'Cathain, 2010.	5.1. Is the mixed methods research design relevant to address the qualitative and quantitative research questions (or objectives), or the qualitative and quantitative aspects of the mixed methods question (or objective)? E.g., the rationale for integrating qualitative and quantitative methods to answer the research question is explained. 5.2. Is the integration of qualitative and quantitative data (or results) relevant to address the research question (objective)? E.g., there is evidence that data gathered by both research methods was brought together to form a complete picture, and answer the research question; authors explain when integration occurred (during the data collection-analysis or/and during the interpretation of qualitative and quantitative results); they explain how integration occurred and who participated in this integration. 5.3. Is appropriate consideration given to the limitations associated with this integration, e.g., the divergence of qualitative and quantitative data (or results)?

References

- Caelli, K., Ray, L., & Mill, J. (2003). 'Clear as Mud': Toward greater clarity in generic qualitative research. *International Journal of Qualitative Methods, 2(*2), 1-23.
- Creswell, J., & Plano Clark, V. (2007). *Designing and conducting mixed methods research*. London: SAGE.
- Creswell, J. (1998). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks: SAGE.
- Critical Appraisal Skills Programme (2009). CASP appraisal tools. Retrieved on August 26, 2009 from: www.phru.nhs.uk/pages/PHD/resources.htm.
- Draugalis, J.R., Coons, S.J., & Plaza, C.M. (2008). Best practices for survey research reports: A synopsis for authors and reviewers. *American Journal of Pharmaceutical Education, 72(1),* e11.
- Higgins, J.P.T. & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions* - Version 5.0.1 [updated September 2008]. The Cochrane Collaboration. Retrieved on August 26, 2009 from www.cochrane-handbook.org
- Mays, N., & Pope, C. (1995). Qualitative Research: Rigour and qualitative research. *British Medical Journal, 311(6997),* 109-112.
- O'Cathain, A., Murphy, E. & Nicholl, J. (2008). The quality of mixed methods studies in health services research. *Journal of Health Services Research and Policy, 13(2),* 92-98.
- O'Cathain, A. (2010). Assessing the quality of mixed methods research: Towards a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research (2nd edition)* (pp. 531-555). Thousand Oaks: SAGE.
- Pace, R., Pluye, P., Bartlett, G., Macaulay, A., Salsberg, J., Jagosh, J., & Seller, R. (2010). *Reliability of a tool for concomitantly appraising the methodological quality of qualitative, quantitative and mixed methods research: A pilot study*. 38th Annual Meeting of the North American Primary Care Research Group (NAPCRG), Seattle, USA.
- 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),* 529-46.
- Oxford Center for Evidence Based Medicine (2009). Levels of evidence. Retrieved on July 7, 2009 from www.cebm.net/levels_of_evidence.asp
- Porta, M. (2008). *A dictionary of epidemiology*. New York: Oxford University Press.
- Sandelowski, M. (2010). What's in a name? Qualitative description revisited. *Research in Nursing and Health, 33*(1), 77-84.
- Schwandt, T. (2001). *Dictionary of qualitive inquiry*. Thousand Oaks: SAGE.
- Wells, G.A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2009). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses*. The Cochrane Non-Randomized Studies Method Group. Retrieved on July 7, 2009 from www.ohri.ca/programs/clinical_epidemiology/oxford.htm

APPENDIX 2. PAPER PUBLISHED ON THE REVIEW OF SYSTEMATIC MIXED STUDIES REVIEWS

Hong, Q.N., Pluye, P., Bujold, M., & Wassef, M. (2017). Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Systematic Reviews, 6*(61), 1-14.

Hong et al. Systematic Reviews (2017) 6:61
DOI 10.1186/s13643-017-0454-2

Systematic Reviews

RESEARCH Open Access

Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence

Quan Nha Hong[1*], Pierre Pluye[1], Mathieu Bujold[1] and Maggy Wassef[2]

Abstract

Background: Systematic reviews of qualitative and quantitative evidence can provide a rich understanding of complex phenomena. This type of review is increasingly popular, has been used to provide a landscape of existing knowledge, and addresses the types of questions not usually covered in reviews relying solely on either quantitative or qualitative evidence. Although several typologies of synthesis designs have been developed, none have been tested on a large sample of reviews. The aim of this review of reviews was to identify and develop a typology of synthesis designs and methods that have been used and to propose strategies for synthesizing qualitative and quantitative evidence.

Methods: A review of systematic reviews combining qualitative and quantitative evidence was performed. Six databases were searched from inception to December 2014. Reviews were included if they were systematic reviews combining qualitative and quantitative evidence. The included reviews were analyzed according to three concepts of synthesis processes: (a) synthesis methods, (b) sequence of data synthesis, and (c) integration of data and synthesis results.

Results: A total of 459 reviews were included. The analysis of this literature highlighted a lack of transparency in reporting how evidence was synthesized and a lack of consistency in the terminology used. Two main types of synthesis designs were identified: convergent and sequential synthesis designs. Within the convergent synthesis design, three subtypes were found: (a) data-based convergent synthesis design, where qualitative and quantitative evidence is analyzed together using the same synthesis method, (b) results-based convergent synthesis design, where qualitative and quantitative evidence is analyzed separately using different synthesis methods and results of both syntheses are integrated during a final synthesis, and (c) parallel-results convergent synthesis design consisting of independent syntheses of qualitative and quantitative evidence and an interpretation of the results in the discussion.

Conclusions: Performing systematic reviews of qualitative and quantitative evidence is challenging because of the multiple synthesis options. The findings provide guidance on how to combine qualitative and quantitative evidence. Also, recommendations are made to improve the conducting and reporting of this type of review.

Keywords: Systematic review, Research synthesis, Mixed studies review, Mixed methods review, Integrative review, Mixed methods research

* Correspondence: quan.nha.hong@mail.mcgill.ca
[1]Department of Family Medicine, McGill University, 5858 Chemin de la Côte-des-Neiges, 3rd Floor, Montreal, QC H3S 1Z1, Canada
Full list of author information is available at the end of the article

Background

Systematic reviews have been used by policy-makers, researchers, and health service providers to inform decision-making [1]. Traditionally, systematic reviews have given preference to quantitative evidence (mainly from randomized controlled trials (RCTs) and to clinical effectiveness questions). However, a focus on quantitative evidence is insufficient in areas where research is not dominated by RCTs [2]. For example, in several fields such as public health, RCTs are not always appropriate nor sufficient to address complex and multifaceted problems [3]. Also, while reviews focusing on RCTs can help to answer the question, "What works for whom?," other important questions remain unanswered such as "Why does it work?," "How does it work?," or "What works for whom in what context?." Such questions can be addressed by reviewing qualitative evidence. Indeed, the analysis of qualitative evidence can complement those of quantitative studies by providing better understanding of the impact of contextual factors, helping to focus on outcomes that are important for patients, families, caregivers, and the population and exploring the diversity of effects across studies [4].

In recent years, there has been a growing interest in synthesizing evidence derived from studies of different designs. This new type of review has been labelled with various terms such as integrative review [5], mixed methods review [6], mixed methods research synthesis [7], mixed research synthesis [8], and mixed studies review [9, 10]. These reviews can yield a rich and highly practical understanding of complex interventions and programs [9, 10]. They can be used to provide (a) a deeper understanding of quantitative evidence, (b) a statistical generalization of findings from qualitative evidence, or (c) a corroboration of knowledge obtained from quantitative and qualitative evidence [9].

The past decade has been rich with methodological advancements of reviews of qualitative and quantitative evidence. For example, several critical appraisal tools for assessing the quality of quantitative and qualitative studies have been developed [9, 11, 12]. Also, new synthesis methods have been developed to integrate qualitative and quantitative evidence such as critical interpretive synthesis, meta-narrative synthesis, and realist synthesis [4, 13, 14]. In addition, researchers have been interested in defining and categorizing different types of synthesis designs (see Table 1). These types were inspired by the literature on mixed methods research, which is a research process integrating quantitative and qualitative methods of data collection and analysis [15]. The types of synthesis design developed are, as yet, theoretical; they have not been tested on a large sample of reviews. Therefore, it is necessary to gain a better understanding of how reviews of qualitative and quantitative evidence are carried out. The aim of this review of reviews was to identify and develop a typology of synthesis designs and methods and to propose strategies for synthesizing qualitative and quantitative evidence.

This review of reviews will contribute to a better understanding of the extent of this literature and justify its relevance. The results will also provide a comprehensive roadmap on how reviews of qualitative and quantitative evidence are carried out. It will provide guidance for conducting and reporting this type of review.

Methods

A review of systematic reviews combining qualitative and quantitative evidence (hereafter, systematic mixed studies reviews (SMSR)) was performed (Table 2). SMSR follows the typical stages of systematic review, with the particularity of including evidence from qualitative, quantitative, and/or mixed method studies [7, 10]. It uses a mixed methods approach [7, 10].

The focus of this review of reviews was on the synthesis process that is the sequence of events and activities regarding how the findings of the included studies were brought together. Thus, a "process-data conceptualization" was conducted [16] using a deductive-inductive approach, i.e., using concepts from the literature on mixed methods research as a starting point, but allowing for new concepts to emerge. Based on the literature on mixed methods research, three main questions were asked: (a) Was the evidence synthesized using qualitative and/or quantitative synthesis methods?, (b) Was there a sequence in the synthesis of the evidence?, and (c) Where did the integration of quantitative and qualitative evidence occur?

Information sources and search strategy

Reviews were searched in six databases (Medline, PsycInfo, Embase, CINAHL, AMED, and Web of Science) from their respective inception dates through December 8, 2014. A search strategy was developed by the first author with the help of two specialized librarians. It included only free text searching since the field of SMSR is still new and no controlled vocabulary exists (see Table 3 for full-search strategy in Medline). All the records were transferred to a reference manager software (EndNote X7) and duplicates were removed using the Bramer-method [17].

Eligibility criteria and selection

SMSRs were included in this review of reviews if they provided a clear description of search and selection strategies, a quality appraisal of included studies, and combined either (a) qualitative, quantitative, and/or mixed methods studies; (b) qualitative and mixed methods studies; (c) quantitative and mixed methods studies; or (d) only mixed methods studies. However,

Table 1 Typology of synthesis designs suggested in the literature

Authors	Synthesis designs[a]
Frantzen and Fetters [40]	1. Convergent meta-integration: quantitative, qualitative, and mixed methods studies are synthesized without data transformation. 2. Convergent qualitative meta-integration: quantitative data are transformed into qualitative format. 3. Convergent quantitative meta-integration: qualitative data are transformed into quantitative format. Each design can be of basic type (when a review includes quantitative and qualitative studies) or advanced type (when a review includes qualitative, quantitative, and mixed methods studies).
Heyvaert et al. [22]	An 18-design framework based on the emphasis of approaches (equal or dominant status of qualitative or quantitative approach), the temporal orientation (sequential or convergent), and the level of integration (partial or full integration).
Pluye and Hong [10]	1. Sequential exploratory: results of the qualitative synthesis inform the quantitative synthesis. 2. Sequential explanatory: results of the quantitative synthesis inform the qualitative synthesis. 3. Convergent: results of qualitative and quantitative studies are integrated using data transformation techniques.
Sandelowski et al. [8]	1. Segregated: qualitative and quantitative findings are treated separately. 2. Integrated: qualitative findings are transformed into quantitative data (quantitizing) or quantitative finding are transformed into qualitative data (qualitizing). 3. Contingent: cycle of research synthesis studies conducted to answer questions raised by previous synthesis.

[a]These synthesis designs are theoretical and not tested on a large sample of reviews

reviews that combined qualitative and mixed methods studies but only analyzed the qualitative evidence of the mixed methods studies were excluded. Likewise, reviews that included quantitative and mixed methods studies but only analyzed quantitative evidence were excluded. SMSRs limited to bibliometric analysis, as well as those that contained only a secondary analysis of studies from previous systematic reviews, were excluded. Also, reviews not published in English or French were excluded.

A three-step selection process was followed. First, all publications that were not journal papers were excluded in EndNote. Second, the remaining records were transferred to the DistillerSR software and two reviewers independently screened all the bibliographic records (titles and abstracts). When the two reviewers disagreed regarding the inclusion/exclusion of a bibliographic record, it was retained for further scrutiny at the next step. Third, two independent reviewers read the full texts of the potentially eligible reviews. Reviews for which the type of studies was not clear (e.g., no description of included studies) were excluded. Also, some reviews were excluded during the analysis because they

Table 2 Three levels of research

	Level of research		
	Primary	Secondary[a]	Tertiary
Research	Empirical study: research based directly on observation, experiment, or simulation rather than on reasoning or theory alone [26, 47].	Systematic review: collation and interpretation of existing empirical studies using systematic and explicit methods [48].	Review of reviews: collation and interpretation of existing systematic reviews [48].
Types of research	Qualitative study: research that aims at exploring and understanding phenomena in terms of the meanings people bring to them [49, 50].	Qualitative review: review combining qualitative studies.	Review of qualitative reviews: review combining qualitative reviews.
	Quantitative study: research that aims at testing theories by examining the relationship among variables [49].	Quantitative review: review combining quantitative studies.	Review of quantitative reviews: review combining quantitative reviews.
	Mixed methods study: research involving collecting and integrating both quantitative and qualitative data [49].	Mixed studies review: review combining qualitative, quantitative, and/or mixed methods studies.	Review of mixed studies reviews: review combining mixed studies reviews.
Data	Primary data collected from fieldwork or lab work.	Findings from included studies.	Findings from included reviews.
Data analysis	Analysis: a step within empirical study of investigating, making sense of, interpreting, and/or theorizing primary data using statistical and/or text analysis procedures [49, 51].	Synthesis: a step within a systematic review consisting of creating something new of findings from included studies [48].	Synthesis of findings across included reviews.

[a]Secondary research is different from secondary analysis. Secondary analysis is used to designate the reanalysis of primary data to answer new questions [52]

Table 3 Search strategy (in Medline)

Concepts	Terms searched
Mixing studies, methods, or data	1. mixed method*.mp 2. mixed stud*.mp 3. mixed research.mp 4. mixed knowledge.mp 5. multi-method*.mp 6. multimethod* 7. multiple method*.mp 8. OR/1-7
Quantitative and qualitative	9. quantitative.mp 10. trial*.mp 11. qualitative.mp 12. 9 or 10 13. 11 and 12
Reviews or syntheses	14. systemat* review*.mp 15. systemat* synthes*.mp 16. critical review*.mp 17. critical synthes*.mp 18. structured review*.mp 19. structured synthes*.mp 20. integrat* review*.mp 21. integrat* synthes*.mp 22. (literature adj3 review*).mp 23. (literature adj3 synthes*).mp 24. research review*.mp 25. research synthes*.mp 26. evidence review*.mp 27. evidence synthes*.mp 28. comprehensive review*.mp 29. comprehensive synthes*.mp 30. OR/14-29
Specific synthesis methods	31. realist review*.mp 32. realist synthes*.mp 33. meta-narrative review*.mp 34. meta-narrative synthes*.mp 35. critical interpretive review*.mp 36. critical interpretive synthes*.mp 37. 31 or 32 or 33 or 34 or 35 or 36
Combination and limits	38. 8 or 13 39. 30 and 38 40. 37 or 39 41. limit 40 to (English or French)

considered quantitative surveys as qualitative studies. Disagreements were reconciled through discussion or arbitration by a third reviewer.

Data collection and synthesis

One reviewer extracted the following data using NVivo 10: year, country, number of included studies, review title, justification for combining qualitative and quantitative evidence, and synthesis methods mentioned.

The quality of the retained reviews was not critically appraised because the aim of this review of reviews was to have a better understanding of how the synthesis is performed in SMSRs. In general, performing an appraisal is useful to check the trustworthiness of individual studies to a review and if the quality might impact the review findings [18]. This review of reviews did not focus on the findings of each review but put emphasis on the synthesis method used and how the findings were presented. Also, while some tools for appraising systematic reviews of quantitative studies exist [19, 20], to our knowledge, there is no tool for appraising the quality of SMSRs.

The data describing the synthesis processes of included reviews were analyzed using the visual mapping technique, which is commonly used for conceptualizing process data [16]. Two reviewers created visual diagrams to represent the synthesis process, i.e., the means by which the qualitative and quantitative evidence, synthesis methods, and findings were linked. These diagrams were then compared and categorized into ideal types. An ideal type is defined as the grouping of characteristics that are common to most cases of a given phenomenon [21].

The analysis focused on three concepts inspired by the literature on mixed methods research [22–24]: (a) synthesis methods, (b) sequence of data synthesis, and (c) integration of data and synthesis results.

(a) *Synthesis methods*: Synthesis consists of the stage of a review when the evidence extracted from the individual sources is brought together [13]. The synthesis method was identified from information provided in the Methods and Results sections. In line with the literature on mixed methods research, the synthesis methods were classified as quantitative or qualitative based on the process and output generated. A synthesis method was considered quantitative when the main results on specific variables across included studies were summarized or combined [25]. Quantitative output is based on numerical values of variables, which are typically produced using validated and reliable checklists and scales and are used to produce numerical data and summaries (such as frequency, mean, confidence interval, and standard error) and conduct statistical analyses [26]. Conversely, a synthesis method was considered qualitative when it summarized or interpreted data to generate outputs such as themes, concepts, frameworks, or theories (inter-related concepts). The distinction between qualitative and quantitative synthesis methods was clear in most cases. However, some synthesis methods required further discussion between the reviewers. For example, in this review of reviews, a distinction between qualitative and quantitative content analysis was made. Content analysis described in Neuendorf [27] and Krippendorff [28] was considered quantitative synthesis method because the coded categories are reliable variables and values allowing descriptive and analytical statistics. This method was developed over a century ago and is defined "as the systematic, objective, quantitative analysis of message characteristics" [27]. In contrast, qualitative content analysis produces themes and subthemes that are

qualitative in nature [29]. Also, in some SMSRs, the synthesis methods were not considered quantitative even if numbers were provided in the results. For example, some presented a table of frequencies of the number of studies for each theme identified from a thematic synthesis. The synthesis was considered qualitative since the main outputs were themes, while the numbers did not provide a combined estimate of a specific variable. Moreover, some synthesis methods are not exclusively qualitative or quantitative. For example, configurational comparative method has been considered simultaneously quantitative and qualitative by the developers [30]. In this review of reviews, this method was considered quantitative because it relies on logical inferences (Boolean algebra) and aims to reduce cases to a series of variables. Another synthesis method requiring discussion was vote counting that is considered quantitative in the literature [31]. In this review of reviews, vote counting was considered qualitative when the results were only used for descriptive purpose.

Tables 4 and 5 present a list of quantitative and qualitative synthesis methods found in the literature [13, 32–34]. When there was a discrepancy between the method described and the method used, the information from the latter was considered during the analysis. For example, some reviews described meta-analysis in the Methods section yet indicated in the Results section that the data were too heterogeneous to be combined quantitatively and a narrative analysis was, thus, used. In this case, the synthesis was considered as qualitative.

Within each review, one or several synthesis methods could be used. The synthesis process could be either qualitative (i.e., used one or several qualitative synthesis methods to analyze the included studies), quantitative (i.e., used one or several quantitative synthesis methods to analyze the included studies), or mixed (i.e., used both qualitative and quantitative synthesis methods to analyze the included studies).

(b) *Sequence*: In the literature on mixed methods research, a sequence refers to a temporal relationship between qualitative and quantitative methods of data collection and analysis [15]. In this review of reviews, the sequence of the analysis was determined based on the number of phases of synthesis and whether the results of one phase informed the synthesis of a subsequent phase. For example, a qualitative synthesis of qualitative studies is done first to identify the components of an intervention (phase 1). Then, the quantitative studies are analyzed to quantify the effect of each component (phase 2). In this case, we considered there was a sequence because the results of the qualitative synthesis informed the quantitative synthesis.

(c) *Integration*: In the literature on mixed methods research, integration is defined as the process of bringing (mixing) qualitative and quantitative approaches together and can be achieved at the level of the design (e.g., sequential and convergent designs), the methods (data collection and analysis), and the interpretation and reporting [35, 36].

In this review of reviews, we adapted these levels of integration: (1) data, i.e., all evidence analyzed using a same synthesis method, (2) results of syntheses, i.e., the results of the synthesis of qualitative and quantitative evidence are compared or combined, (3) interpretation, i.e., the discussion of the results of the synthesis of qualitative and quantitative evidence, and (4) design.

Results

Description of included reviews

The bibliographic database search yielded 7003 records of which 459 SMSRs were included in this review of reviews (Fig. 1). As seen in Fig. 2, there has been an exponential progression of the number of publications per year, especially since 2010. In over a decade, the number has passed from nearly 10 per year to more than 100. The topics of the SMSRs were mainly in health and varied widely, from health care to public health. Some were on information sciences, management, education, and research. The first authors of the SMSRs came from 28 different countries. The countries producing the most SMSRs are England ($n = 179$), Australia ($n = 71$), the USA ($n = 53$), Canada ($n = 45$), and the Netherlands ($n = 20$).

Several labels were used to name this type of review, with the most common being "systematic review" ($n = 277$), followed by "literature review" ($n = 39$), "integrative review" ($n = 35$), and "mixed methods reviews" ($n = 24$). Among those using the term systematic review, a small number specified in the title that they combined different types of evidence: "mixed systematic review" ($n = 2$), and "systematic review of quantitative and qualitative" data, evidence, literature, research, or studies ($n = 23$).

The number of studies included in the SMSRs ranged from 2 to 295 (mean = 29; SD = 33). The majority of SMSRs included qualitative and quantitative studies ($n = 249$) or qualitative, quantitative, and mixed methods studies ($n = 200$). Few included only quantitative and mixed methods studies ($n = 8$) or only qualitative and mixed methods studies ($n = 2$).

Only 24% ($n = 110$) of included reviews provided a clear rationale for combining quantitative and qualitative evidence. Authors described various reasons for performing SMSRs that fall into the following eight

Table 4 Quantitative synthesis methods

Synthesis method	Aim	Description
Bayesian synthesis [53]	To measure the likelihood of different values for parameters of interest.	Incorporates prior distributions of unknown parameter values that are then updated by deriving posterior probability distributions generated through statistical analysis of the estimates.
Case survey [54, 55]	To identify and statistically test patterns across individual case studies.	Converts qualitative cases into quantitative variables by extracting data using a same set of closed-ended questions. The answers to these questions are then aggregated to establish frequency of occurrence (that can be further statistically analyzed, as appropriate).
Configurational comparative method [56]	To build or test theories and assumptions by identifying configurations of causal conditions, i.e., combination of conditions (independent variables) that are necessary and/or sufficient for a given outcome (dependent variable).	Consists in a comparative case-oriented research approach that uses Boolean algebra to generate configurations between conditions and outcomes across cases.
Cross-design synthesis [57]	To combine results from quantitative studies with complementary designs (e.g., RCT and observational studies).	Involves an in-depth assessment of key biases of each study, an adjustment of each study's results based on the identified biases and the development of a model for combining the results within and across designs.
Meta-analysis [58]	To obtain a single summarized "effect size."	Uses statistical methods for combining results of studies into a weighted average of point estimates.
Meta-regression [59]	To relate the size of effect to one or more characteristics of the included studies (to explore sources of heterogeneity across included studies).	Uses a combination of meta-analytic and regression principles.
Meta-summary [60]	To quantitatively aggregate qualitative findings.	Consists of extraction, grouping, abstraction, and formatting of findings and the calculation of frequency and intensity effect sizes.
Quantitative content analysis [27, 28]	To transform qualitative data into few variables (numerical value) for statistical analysis.	Categorizes data and provides statistical description of the categories.
Vote counting [61]	To calculate the frequencies of categories of results across included studies.	The included studies are sorted into three categories (negative significant, positive significant, and statistically insignificant), and the number of studies for each category is calculated. The category with the most studies is the "winner."

categories: (a) nature of the literature on a topic—to adapt the review method because of the limited evidence on the topic or absence of RCTs, (b) complexity of the phenomenon—to address a complex and multifaceted phenomenon, (c) broad coverage—to provide broader perspective and cover a wide range of purposes, (d) comprehensiveness—to provide a complete picture and deduce the maximum information from the literature, (e) thorough understanding—to gain better and detailed understanding of a phenomenon, (f) complementarity—to address different review questions (e.g., why and how) and complement the strengths and limitations of quantitative and qualitative evidence, (g) corroboration—to strengthen and support the results through triangulation, and (h) practical implication—to provide more meaningful and relevant evidence for practice.

Only 39% ($n = 179$) of included reviews provided a full description of the synthesis method(s) with methodological references. The remainder provided information without reference ($n = 149$), simply mentioned (labelled) the synthesis method used ($n = 41$), or did not provide information about the synthesis ($n = 90$). A variety of synthesis methods were used in the included reviews. Among the SMSRs that provided information on the synthesis methods, the most common method mentioned was thematic synthesis ($n = 129$), followed by narrative synthesis ($n = 64$), narrative summary ($n = 30$), categorization/grouping ($n = 20$), content analysis ($n = 30$), meta-synthesis ($n = 25$), meta-analysis ($n = 27$), narrative analysis ($n = 11$), meta-ethnography ($n = 9$), textual narrative ($n = 7$), framework synthesis ($n = 7$), and realist synthesis ($n = 6$).

Synthesis of results

Based on the sequence and integration concepts, two main types of synthesis designs were identified (Fig. 3): convergent and sequential synthesis designs. Within the

Table 5 Qualitative synthesis methods

Synthesis method	Aim	Description
Critical interpretive synthesis [62]	To build a theory from the synthesis of a diverse body of evidence.	Adapted the strategies of meta-ethnography (reciprocal translational analysis, lines-of-argument synthesis, and refutational syntheses) for qualitative and quantitative evidence.
Framework synthesis [63]	To produce a new framework based on a priori and new themes.	Consists of analyzing data using an a priori framework, creating new themes by performing thematic synthesis, and producing a new framework.
Grouping and clustering [44]	To describe included studies.	Summarizes and organizes included studies into groups (categories).
Meta-ethnography [64]	To build a theory from the synthesis of qualitative studies.	Uses three main strategies: translating the concepts from studies into one another (reciprocal translational analysis), exploring and explaining contradictions between studies (refutational synthesis), and linking constructs and building a picture of the whole from studies (lines-of-argument synthesis).
Meta-narrative synthesis [65]	To make sense of complex and conflicting findings by unfolding the storyline of research traditions.	Maps research traditions and consider how they have been conceptualized, theorized, and empirically studied over time.
Meta-synthesis [66]	To understand a phenomenon of interest across qualitative studies.	Uses hermeneutic (portraying individual constructions) and dialectic (comparing and contrasting the constructions) approaches.
Narrative synthesis [44]	To summarize and explain the findings of included studies.	Adopts a textual approach to the process of synthesis and follows four elements: develop a theory of how the intervention works, why, and for whom; develop a preliminary synthesis; explore relationships within and between studies; and assess the robustness of the synthesis.
Qualitative content analysis [29]	To understand a phenomenon of interest by focusing on the manifest (patent) content or contextual meaning of text.	Uses an analytical coding process to organize content of textual data into fewer content categories.
Realist synthesis [67, 68]	To unpack how interventions work in particular contexts through theoretical explanation (middle-range theory).	Uses theory-driven context-mechanism-outcome configurations, demi-regularities, and abduction (hunches).
Textual description [44]	To describe included studies.	Provides a descriptive paragraph of each study.
Textual narrative synthesis [69]	To describe included studies.	Arranges studies into homogeneous groups and compares similarities and differences across studies.
Thematic synthesis [70]	To identify and develop themes across included studies.	Uses line-by-line coding, develops descriptive themes, and generates analytical themes. This might lead to propose a conceptual framework.

convergent synthesis design, three subtypes were found: data-based, results-based, and parallel-results convergent synthesis designs. These synthesis designs were cross tabulated with the three types of synthesis methods (qualitative, quantitative, and mixed). This led to a total of 12 possible synthesis strategies that are represented in Table 6. Reviews were found for eight of these possibilities.

I. Convergent synthesis design: In this design, the quantitative and qualitative evidence is collected and analyzed during the same phase of the research process in a parallel or a complementary manner. Three subtypes were identified based on where the integration occurred.

(a) Data-based convergent synthesis design (Fig. 3a): This design was the most common type of synthesis design (Table 6). In this design, all included studies are analyzed using the same synthesis method and results are presented together. Since only one synthesis method is used for all evidence, data transformation is involved (e.g., qualitative data transformed into numerical values or quantitative data are transformed into categories/themes). This design usually addressed one review question. Among the SMSRs in this

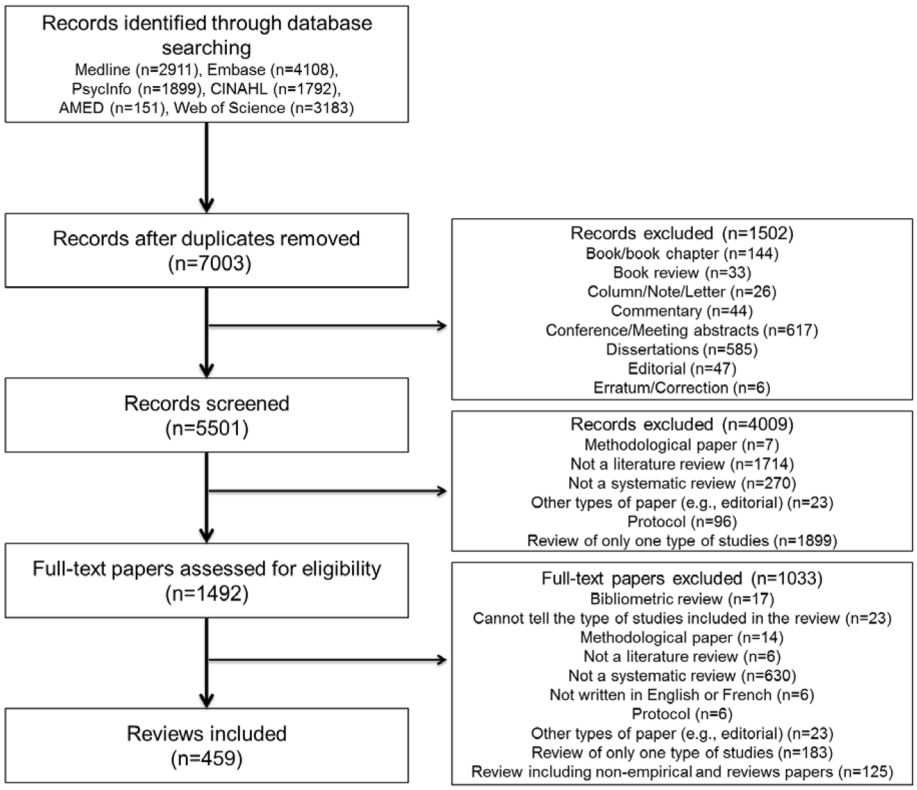


Fig. 1 Flowchart

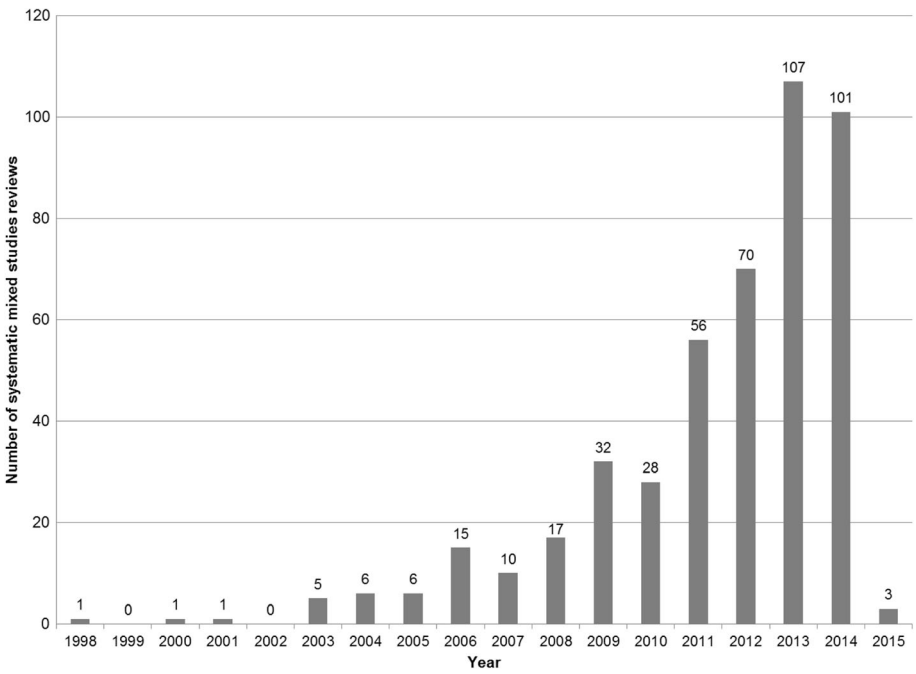


Fig. 2 Number of systematic mixed studies reviews published per year

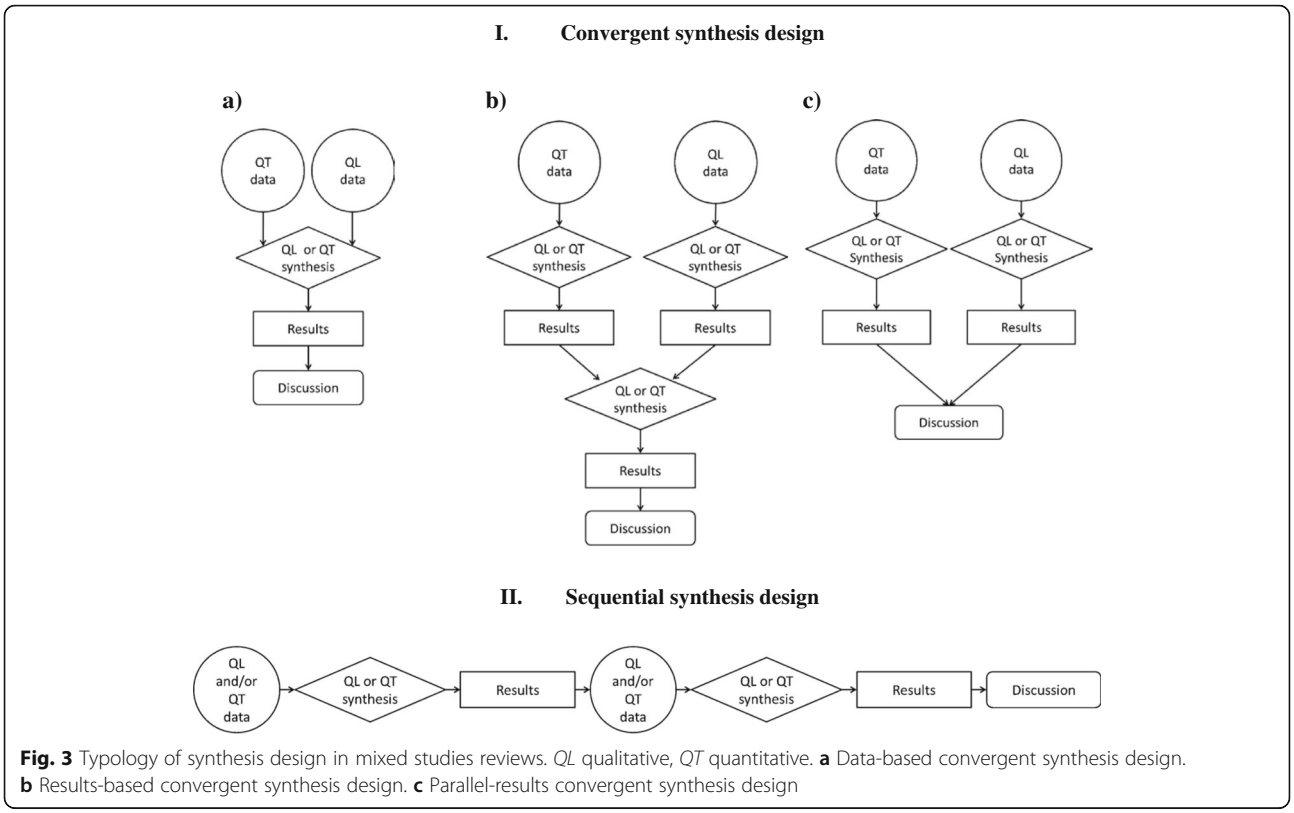


Fig. 3 Typology of synthesis design in mixed studies reviews. *QL* qualitative, *QT* quantitative. **a** Data-based convergent synthesis design. **b** Results-based convergent synthesis design. **c** Parallel-results convergent synthesis design

design, three main objectives were found. The first category sought to describe the findings of the included studies, and the synthesis methods ranged from summarizing each study to grouping main findings. The review questions were generally broad (similar to a scoping review) such as what is known about a specific topic. The second category consisted of SMSRs that sought to identify and define main concepts or themes using a synthesis method such as qualitative content analysis or thematic synthesis. The review questions were generally more specific such as identifying the main barriers and facilitators to the implementation of a program or types of impact. The third category included SMSRs that aimed to establish relationships between the concepts and themes identified from the included studies or to provide a framework/theory.

(b) Results-based convergent synthesis design (Fig. 3b): Nearly 9% of SMSRs were classified in this synthesis design (Table 6). In this design, the qualitative and quantitative evidence is analyzed and presented separately but integrated using another synthesis method. The integration could consist of comparing or juxtaposing the findings of qualitative and quantitative evidence using tables and matrices or reanalyzing evidence in light of the results of both syntheses. For example, Harden and Thomas [6] suggest performing a quantitative synthesis (e.g., meta-analysis) of trials and a qualitative synthesis of studies of people's views (e.g., thematic synthesis). Then, the results of both syntheses are combined in a third synthesis. This

Table 6 Percentages of systematic mixed studies reviews among the 12 synthesis strategies ($n = 459$)

Synthesis	Sequence and integration				Total
	Convergent synthesis design			Sequential synthesis design	
	Data-based	Results-based	Parallel-results		
Qualitative	69.5%	6.3%	12.0%	2.6%	90.4%
Quantitative	0.2%	0%	0%	0%	0.2%
Mixed	0%	2.2%	5.2%	2.0%	9.4%
Total	69.7%	8.5%	17.2%	4.6%	100%

type of design usually addresses an overall review question with subquestions.

(c) Parallel-results convergent design (Fig. 3c): A little over 17% of reviews were classified in this design (Table 6). In this design, qualitative and quantitative evidence is analyzed and presented separately. The integration occurs during the interpretation of results in the Discussion section. Some of these SMSRs included two or more complementary review questions. For example, health technology assessments evaluate several dimensions such as clinical effectiveness, cost-effectiveness, and acceptability of an intervention. The evidence of each dimension is reviewed separately and brought together in the discussion and recommendations.

II. Sequential synthesis design (Fig. 3): This design was found in less than 5% of the reviews (Table 6). It involves a two-phase approach where the data collection and analysis of one type of evidence occur after and are informed by the collection and analysis of the other type. This design usually addressed one overall review question with subquestions and both syntheses complemented each other. For example, in a review aiming at identifying the obstacles to treatment adherence, the qualitative synthesis provided a list of barriers and the quantitative synthesis reported the prevalence of these barriers and knowledge gaps (barriers for which prevalence was not estimated) [37].

Discussion

The number of published SMSRs has considerably increased in the past few years. In a previous review of reviews in 2006, Pluye et al. [9] identified only 17 SMSRs. This shows that there is an increasing interest for this type of review and warrants the need for more methodological development in this field.

In accordance with the literature on mixed methods research, two main types of synthesis designs were identified in this review of reviews: convergent and sequential synthesis designs. Three subtypes of convergent synthesis were found: data-based convergent, results-based, and parallel-results convergent synthesis designs. The data-based convergent design was more frequently used probably because it is easier to perform, especially for a descriptive purpose. The other synthesis designs might be more complex but could allow for greater analytical depth and breadth of the literature on a specific topic. Also, focusing the analysis on the concepts of convergent and sequential designs allowed us to clarify and refine their definitions. Considering that the focus of the analysis was the synthesis process in SMSRs, the literature on process studies especially in the fields of management provides insight into these concepts. First, in line with Langley et al. [38], the convergent design can be defined as a process of gradual, successive, and constant refinements of synthesis and interpretation of the qualitative and quantitative evidence. Researchers are working forward in a non-linear manner guided by a cognitive representation of new data-based synthesis or results-based synthesis or interpretation of results to be created. Second, in line with Van de Ven [39], a sequential synthesis design can be defined, according to a developmental perspective (phase 1 informing phase 2; phase 2 building on the results of phase 1), as a change of focus at the level of data or synthesis over time and as a cognitive transition into a new phase (e.g., from qualitative to quantitative or from quantitative to qualitative).

The synthesis designs found in this review of reviews reflect those suggested by Sandelowski et al. [8] (see Table 1) who used the terms *segregated*, which can be similar to results-based and parallel-results convergent synthesis designs, *integrated*, which is comparable to data-based convergent synthesis design, and *contingent* designs, which could be considered as a form of sequential design. In this review of reviews, we used the mixed methods concepts and terminology because they account for the integration that may be present at the level of data, results, interpretation, or design.

As in Heyvaert et al. [22], the concepts found in the literature on mixed methods research to define the synthesis designs were used; yet, the definition of the synthesis method and integration concepts was somewhat different. In Heyvaert et al. [22], they focused on the relative importance of methods, i.e., whether the qualitative or the quantitative method was dominant or of equal status. This was not done in this review of reviews because measuring or documenting the dominance of a method is difficult given the influences of multiple factors (power, resources, expertise, time, training, and worldviews of each research team member, among other factors). Also, in Heyvaert et al. [22], they considered that integration could be partial (i.e., part of the qualitative and quantitative studies are involved separately in some or all stages) or full (i.e., all the qualitative and quantitative studies are involved in all the stages). In this review of reviews, the focus was put on where the integration occurred. Therefore, this review of reviews resulted in respectively four and three types of synthesis designs and methods, which led to propose 12 synthesis strategies, as compared to 18 in Heyvaert et al. [22].

In Frantzen and Fetters [40], three main types of convergent designs are suggested (see Table 1). Similarly, this review of reviews also found qualitative, quantitative, or mixed convergent synthesis design types. However, no distinction was made during the analysis between SMSRs including only qualitative and quantitative studies (basic

type) and those also including mixed methods studies (advanced type) because this review of reviews aimed at defining ideal types of synthesis designs. The paper written by Frantzen and Fetters [40] went into deeper analysis of convergent design to provide detailed information on the steps to follow to integrate qualitative, quantitative, and mixed methods studies.

Some SMSRs using sequential synthesis design were found in our sample of reviews. Pluye and Hong [10] suggested using the sequential exploratory or explanatory designs. In the exploratory sequential design, a qualitative synthesis is performed first and results inform the subsequent quantitative synthesis. Conversely, in an explanatory sequential design, the quantitative synthesis is done first and informs the subsequent qualitative synthesis. In this review of reviews, the sequence was defined as the results of one phase informing the other (not limited to the order of the syntheses) and no review was classified as sequential explanatory. In addition, 12 SMSRs performing only qualitative syntheses were found and could not be classified as exploratory or explanatory. For the sake of parsimony, we did not make a distinction between exploratory and explanatory sequential synthesis designs.

Implications for conducting and reporting mixed studies reviews

In light of this review of reviews and the literature on mixed methods research, four complementary key recommendations can be made regarding the title, justification, synthesis methods, and the integration of qualitative and quantitative data.

First, researchers should explicitly state in the title that the review included qualitative and quantitative evidence. Various terms are used to designate this type of review. Some SMSRs used the term "mixed" such as mixed systematic review, mixed methods review, mixed research synthesis, or mixed studies review. The term mixed has been used in the mixed methods literature to designate primary research designs combining qualitative and quantitative approaches [23]. In the field of review, mixing qualitative and quantitative evidence can be seen at two levels: study level and synthesis level [22]. Pluye et al. [9] suggested "mixed studies review" referring to a review of studies of different designs. This name focuses on the study level and does not prescribe a specific synthesis method. Others have suggested labelling this type of review as mixed methods review [6, 22] wherein mixing occurs at both the level of the study and the synthesis. Another popular term is integrative review proposed by Whittemore and Knafl [5]. Integrative review is described as a type of literature review to synthesize the results of research, methods, or theories using a narrative analysis [41]. Currently, all these terms are used interchangeably without a clear distinction [40].

Second, researchers should provide a clear justification for performing a SMSR and describe the synthesis design used. In this review of reviews, this information was found in only 24% of the SMSRs. This lack of justification for using qualitative and quantitative evidence is also found in the literature on mixed methods research [42]. The rationale will influence the review questions and the choice of the synthesis design. For example, if quantitative and qualitative evidence is used for corroboration purpose, the convergent synthesis design may be more relevant. On the other hand, when they are used in complementarity such as using the quantitative studies to generalize qualitative findings or using qualitative studies to interpret, explain, or provide more insight to some quantitative findings, the sequential synthesis design may be more appropriate.

Third, results of this review of reviews suggest a need to recommend that researchers describe their synthesis methods and cite methodological references. Only 39% of the SMSRs provided a full description of the synthesis methods with methodological references. Various synthesis methods have been developed over the past decade [13, 32, 33, 43]. Meta-analysis is the best known synthesis method to aggregate findings in reviews, especially for clinical effectiveness questions. However, when this method is not possible, researchers tend to omit describing the synthesis. Researchers should avoid limiting the description to what was not done such as using the sentence "because of the heterogeneity of studies, no meta-analysis was performed and data were analyzed narratively." The term "narrative" can be confusing since it is often used differently by different authors. In some SMSRs, narrative analysis corresponded to summarizing each included study. In others, it consisted in grouping the different findings of included studies into main categories and summarizing the evidence of each category. Still, others followed Popay et al.'s [44] four main elements for narrative synthesis (i.e., develop a theoretical model, preliminary synthesis, relationship, and assess robustness). Hence, in addition to naming the synthesis method, we recommend that reviews should provide a clear description of what was done to synthesize the data and add methodological references. This will improve transparency of the review process, which is an essential quality of systematic reviews.

Fourth, researchers should describe how the data were integrated and discuss the insight gained from this process. Integration is an inherent component of mixed methods research [15], and careful attention must be paid to how integration is done and reported to enhance the value of a review. The synthesis designs outline that can provide guidance on how to integrate data (Fig. 3). Also, the discussion should include more than a simple wrap-up of results. It should clearly reflect on the added

value and insight gained of combining qualitative and quantitative evidence into a review.

Limitations

The search strategy used was not comprehensive; thus, not all SMSRs were identified in this review of reviews. Indeed, the search was limited to six databases mainly in health and no hand searching was performed. As this review of reviews deals with methods, citation tracking of included SMSRs would not have provided additional relevant references. Nonetheless, our sample of included SMSRs was large ($n = 459$) and sufficient to achieve the aim of this review of reviews.

To ensure a manageable sample size, selection of included reviews was limited to peer-reviewed journal articles. We acknowledge that the sample of included reviews might not include some innovative developments in this field, given that some recent SMSRs may be reported in other types of publications (e.g., conference abstracts or gray literature).

Finally, the synthesis methods were not classified as aggregative and configurative [45, 46]. As mentioned in Gough et al. [45], some configurative synthesis can include aggregative component and vice versa. To avoid this confusion, the terms qualitative and quantitative synthesis methods were preferred. Moreover, these terms were used to align with the mixed methods research terminology. Yet, as discussed in the Methods section, the interpretation of some synthesis methods used in this review of reviews can be debatable.

Conclusions

The field of SMSR is still young, though rapidly evolving. This review of reviews focused on how the qualitative and quantitative evidence is synthesized and integrated in SMSRs and suggested a typology of synthesis designs. The analysis of this literature also highlighted a lack of transparency in reporting how data were synthesized and a lack of consistency in the terminology used. Some avenues for future research can be suggested. First, there is a need to reach consensus on the terminology and definition of SMSRs. Moreover, given the wide range of approaches to synthesis, clear guidance and training are required regarding which synthesis methods to use and when and how they should be used. Also, future research should focus on the development, validation, and reliability testing of quality appraisal criteria and standards of high-quality SMSRs. Finally, an adapted PRISMA statement for reporting SMSRs should be developed to help advance the field.

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; QL: Qualitative; QT: Quantitative; RCT: Randomized controlled trial; SMSR: Systematic mixed studies review

Acknowledgements

The authors would like to thank Ms. Vera Granikov, research-embedded health librarian, and Ms. Genevieve Gore, liaison librarian at McGill University, for their help in the development of the search strategy. Also, the authors are grateful to Dr. Paula Bush for the constructive comments on an earlier version of this manuscript. Finally, the authors thank the reviewers for their constructive comments that helped to improve the manuscript.

Funding

QNH holds a doctoral fellowship from the Canadian Institutes of Health Research (CIHR). PP holds a Senior Research Scholar Fellowship from the Fonds de recherche du Québec—Santé (FRQS). MB is supported by the methodological developments component of the Quebec SUPPORT Unit (CIHR). The views expressed in this article are those of the authors and not necessarily those of the funding agencies.

Availability of data and materials

The data are reported in the manuscript.

Authors' contributions

QNH conceived the review design, contributed to the search, selection, and analysis of the reviews, and drafted the first version of the manuscript. PP supervised the whole process, and contributed to the selection (third reviewer in case of disagreement) and the analysis of the reviews. MB and MW contributed to the selection of the reviews. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethical approval and consent to participate

Not applicable.

Publisher's Note

Author details

[1]Department of Family Medicine, McGill University, 5858 Chemin de la Côte-des-Neiges, 3rd Floor, Montreal, QC H3S 1Z1, Canada. [2]Information Technology Primary Care Research Group, Department of Family Medicine, McGill University, Montreal, QC, Canada.

Received: 8 January 2016 Accepted: 13 March 2017
Published online: 23 March 2017

References

1. Bunn F, Trivedi D, Alderson P, Hamilton L, Martin A, Pinkney E, et al. The impact of Cochrane Reviews: a mixed-methods evaluation of outputs from Cochrane Review Groups supported by the National Institute for Health Research. Health Technol Assess. 2015;19(28):1–100. doi:10.1186/2046-4053-3-125.
2. Goldsmith MR, Bankhead CR, Austoker J. Synthesising quantitative and qualitative research in evidence-based patient information. J Epidemiol Community Health. 2007;61(3):262–70. doi:10.1136/jech.2006.046110.
3. Victora CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health. 2004;94(3):400–5. doi:10.2105/AJPH.94.3.400.
4. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Integrative approaches to qualitative and quantitative evidence. London, UK: Health Development Agency; 2004. Report No.: 1842792555.
5. Whittemore R, Knafl K. The integrative review: updated methodology. J Adv Nurs. 2005;52(5):546–53. doi:10.1111/j.1365-2648.2005.03621.x.
6. Harden A, Thomas J. Methodological issues in combining diverse study types in systematic reviews. Int J Soc Res Meth. 2005;8(3):257–71. doi:10.1080/13645570500155078.
7. Heyvaert M, Hannes K, Onghena P. Using mixed methods research synthesis for literature reviews: the mixed methods research synthesis approach. Thousand Oaks, CA: SAGE Publications; 2016.

8. Sandelowski M, Voils CI, Barroso J. Defining and designing mixed research synthesis studies. Res Schools. 2006;13(1):29–40.
9. Pluye P, Gagnon MP, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in mixed studies reviews. Int J Nurs Stud. 2009;46(4):529–46. doi:10.1016/j.ijnurstu.2009.01.009.
10. Pluye P, Hong QN. Combining the power of stories and the power of numbers: mixed methods research and mixed studies reviews. Annu Rev Public Health. 2014;35:29–45. doi:10.1146/annurev-publhealth-032013-182440.
11. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. J Clin Epidemiol. 2011;64(1):79–89. doi:10.1016/j.jclinepi.2010.02.008.
12. Sirriyeh R, Lawton R, Gardner P, Armitage G. Reviewing studies with diverse designs: the development and evaluation of a new tool. J Eval Clin Pract. 2012;18(4):746–52. doi:10.1111/j.1365-2753.2011.01662.x.
13. Mays N, Pope C, Popay J. Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. J Health Serv Res Policy. 2005;10 Suppl 1:6–20. doi:10.1258/1355819054308576.
14. Tricco AC, Antony J, Soobiah C, Kastner M, MacDonald H, Cogo E, et al. Knowledge synthesis methods for integrating qualitative and quantitative data: a scoping review reveals poor operationalization of the methodological steps. J Clin Epidemiol. 2016;73:29–35. doi:10.1016/j.jclinepi.2015.12.011.
15. Plano Clark VL, Ivankova NV. Mixed methods research: a guide to the field. SAGE mixed methods research series. Thousand Oaks: SAGE Publications; 2015.
16. Langley A. Strategies for theorizing from process data. Acad Manage Rev. 1999;24(4):691–710. doi:10.2307/259349.
17. Bramer WM, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016;104(3):240. doi:10.3163/1536-5050.104.3.014.
18. Booth A, Papaioannou D, Sutton A. Systematic approaches to a successful literature review. London: SAGE Publications; 2012.
19. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34. doi:10.1016/j.jclinepi.2015.06.005.
20. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol. 2009;62(10):1013–20. doi:10.1016/j.jclinepi.2008.10.009.
21. Weber M, Freund J, Kamnitzer P, Bertrand P. Économie et société: les catégories de la sociologie. Paris: Pocket; 1995.
22. Heyvaert M, Maes B, Onghena P. Mixed methods research synthesis: definition, framework, and potential. Qual Quant. 2013;47(2):659–76. doi:10.1007/s11135-011-9538-6.
23. Creswell JW, Plano CV. Designing and conducting mixed methods research. 2nd ed. Thousand Oaks: SAGE Publications; 2011.
24. Collins KM, O'cathain A. Introduction: ten points about mixed methods research to be considered by the novice researcher. Int J Mult Res Approaches. 2009;3(1):2–7. doi:10.5172/mra.455.3.1.2.
25. Sutton AJ, Jones DR, Abrams KR, Sheldon TA, Song F. Systematic reviews and meta-analysis: a structured review of the methodological literature. J Health Serv Res Policy. 1999;4(1):49–55. doi:10.1177/135581969900400112.
26. Porta MS, Greenland S, Hernán M, dos Santos SI, Last JM. A dictionary of epidemiology. New York: Oxford University Press; 2014.
27. Neuendorf KA. The content analysis guidebook. Thousand Oaks: SAGE Publications; 2002.
28. Krippendorff K. Content analysis: an introduction to its methodology. Thousand Oaks: SAGE Publications; 2012.
29. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res. 2005;15(9):1277–88. doi:10.1177/1049732305276687.
30. Rihoux B, Marx A. QCA, 25 years after "the comparative method": mapping, challenges, and innovations—mini-symposium. Polit Res Q. 2013;66(1):167–235. doi:10.1177/1065912912468269.
31. Hedges LV, Olkin I. Vote-counting methods in research synthesis. Psychol Bull. 1980;88(2):359. doi:10.1037/0033-2909.88.2.359.
32. Barnett-Page E, Thomas J. Methods for the synthesis of qualitative research: a critical review. BMC Med Res Methodol. 2009;9(59). doi:10.1186/1471-2288-9-59.
33. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: a review of possible methods. J Health Serv Res Policy. 2005;10(1):45–53. doi:10.1258/1355819052801804.
34. Kastner M, Tricco AC, Soobiah C, Lillie E, Perrier L, Horsley T et al. What is the most appropriate knowledge synthesis method to conduct a review? Protocol for a scoping review. BMC Med Res Methodol. 2012;12(114). doi:10.1186/1471-2288-12-114.
35. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs—principles and practices. Health Serv Res. 2013;48(6pt2):2134–56. doi:10.1111/1475-6773.12117.
36. Guetterman TC, Fetters MD, Creswell JW. Integrating quantitative and qualitative results in health science mixed methods research through joint displays. Ann Fam Med. 2015;13(6):554–61. doi:10.1370/afm.1865.
37. Mills EJ, Nachega JB, Bangsberg DR, Singh S, Rachlis B, Wu P, et al. Adherence to HAART: a systematic review of developed and developing nation patient-reported barriers and facilitators. PLoS Med. 2006;3(11):e438. doi:10.1371/journal.pmed.0030438.
38. Langley A, Mintzberg H, Pitcher P, Posada E, Saint-Macary J. Opening up decision making: the view from the black stool. Organ Sci. 1995;6(3):260–79. doi:10.1287/orsc.6.3.260.
39. Van de Ven AH. Suggestions for studying strategy process: a research note. Strat Manag J. 1992;13(5):169–88. doi:10.1002/smj.4250131013.
40. Frantzen KK, Fetters MD. Meta-integration for synthesizing data in a systematic mixed studies review: insights from research on autism spectrum disorder. Qual Quant. 2015:1-27. doi:10.1007/s11135-015-0261-6.
41. Whittemore R, Chao A, Jang M, Minges KE, Park C. Methods for knowledge synthesis: an overview. Heart Lung. 2014;43(5):453–61. doi:10.1016/j.hrtlng.2014.05.014.
42. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy. 2008;13(2):92–8. doi:10.1258/jhsrp.2007.007074.
43. Tricco AC, Soobiah C, Antony J, Cogo E, MacDonald H, Lillie E, et al. A scoping review identifies multiple emerging knowledge synthesis methods, but few studies operationalize the method. J Clin Epidemiol. 2016;73:19–28. doi:10.1016/j.jclinepi.2015.08.030.
44. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews. Lancaster, UK: Lancaster University; 2006.
45. Gough D, Thomas J, Oliver S. Clarifying differences between review designs and methods. Syst Rev. 2012;1(28). doi:10.1186/2046-4053-1-28.
46. Anderson LM, Oliver SR, Michie S, Rehfuess E, Noyes J, Shemilt I. Investigating complexity in systematic reviews of interventions by using a spectrum of methods. J Clin Epidemiol. 2013;66(11):1223–9. doi:10.1016/j.jclinepi.2013.06.014.
47. Abbott A. The causal devolution. Sociol Methods Res. 1998;27(2):148–81. doi:10.1177/0049124198027002002.
48. Gough D, Oliver S, Thomas J. An introduction to systematic reviews. London: SAGE Publications; 2012.
49. Creswell JW. Research design: qualitative, quantitative, and mixed methods approaches. Thousand Oaks: SAGE Publications; 2013.
50. Denzin NK, Lincoln YS. The Sage handbook of qualitative research. Thousand Oaks: SAGE Publications; 2005.
51. Schwandt TA. The Sage dictionary of qualitative inquiry. Thousand Oaks: SAGE Publications; 2015.
52. Glass GV. Primary, secondary, and meta-analysis of research. Educ Res. 1976;5:3–8. doi:10.3102/0013189X005010003.
53. Louis TA, Zelterman D. Bayesian approaches to research synthesis. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russell Sage; 1994. p. 411–22.
54. Larsson R. Case survey methodology: quantitative analysis of patterns across case studies. Acad Manage J. 1993;36(6):1515–46. doi:10.2307/256820.
55. Yin RK, Heald KA. Using the case survey method to analyze policy studies. Adm Sci Q. 1975;20(3):371–81. doi:10.2307/2391997.
56. Rihoux B, Ragin CC. Configurational comparative methods: qualitative comparative analysis (qca) and related techniques, vol. 51. Thousand Oaks: SAGE Publications; 2009.
57. Droitcour J, Silberman G, Chelimsky E. Cross-design synthesis: a new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases. Int J Technol Assess Health Care. 1993;9(03):440–9. doi:10.1017/S0266462300004694.
58. Sutton AJ, Higgins J. Recent developments in meta-analysis. Stat Med. 2008;27(5):625–50. doi:10.1002/sim.2934.
59. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21(11):1559–73. doi:10.1002/sim.1187.

60. Sandelowski M, Barroso J, Voils CI. Using qualitative metasummary to synthesize qualitative and quantitative descriptive findings. Res Nurs Health. 2007;30(1):99–111. doi:10.1002/nur.20176.
61. Light RJ, Smith PV. Accumulating evidence: procedures for resolving contradictions among different research studies. Harv Educ Rev. 1971;41(4):429–71. doi:10.17763/haer.41.4.437714870334w144.
62. Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J et al. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. BMC Med Res Methodol. 2006;6(35). doi:10.1186/1471-2288-6-35.
63. Booth A, Carroll C. How to build up the actionable knowledge base: the role of 'best fit' framework synthesis for studies of improvement in healthcare. BMJ quality & safety. 2015. doi:10.1136/bmjqs-2014-003642.
64. Noblit GW, Hare RD. Meta-ethnography: synthesizing qualitative studies. Thousand Oaks: SAGE Publications; 1988.
65. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O, Peacock R. Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. Soc Sci Med. 2005;61:417–30. doi:10.1016/j.socscimed.2004.12.001.
66. Jensen LA, Allen MN. Meta-synthesis of qualitative findings. Qual Health Res. 1996;6(4):553–60. doi:10.1177/104973239600600407.
67. Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. RAMESES publication standards: realist syntheses. BMC Med. 2013;11(21). doi:10.1186/1741-7015-11-21.
68. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review—a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10 Suppl 1:21–34. doi:10.1258/1355819054308530.
69. Lucas PJ, Baird J, Arai L, Law C, Roberts HM. Worked examples of alternative methods for the synthesis of qualitative and quantitative research in systematic reviews. BMC Med Res Methodol. 2007;7:4. doi:10.1186/1471-2288-7-4.
70. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol. 2008;8:45. doi:10.1186/1471-2288-8-45.

APPENDIX 3. LIST OF CRITICAL APPRAISAL TOOLS USED IN SYSTEMATIC MIXED STUDIES REVIEWS

List of Critical Appraisal Tools Used in Systematic Mixed Studies Reviews

	Tool	Reference	Number of SMSRs that used the tool for:			
			MM	QUAL	QUAN	Several types of studies
1.	Aboelela 2007	Aboelela, S., Stone, P., & Larson, E. (2007). Effectiveness of bundled behavioural interventions to control healthcare-associated infections: A systematic review of the literature. *Journal of Hospital Infection*, 66(2), 101-108.			1	
2.	Abramson 2005	Abramson, J. H., & Abramson, Z. H. (2005). *Survey methods in community medicine*. 5th ed. Edinburgh: Chruchill Livingstone.			1	
3.	Ajetunmob 2002	Ajetunmobi, O. (2002). *Making sense of critical appraisal*. London: Hodder Arnold.			1	
4.	Akesson 2007	Åkesson, K. M., Saveman, B.-I., & Nilsson, G. (2007). Health care consumers' experiences of information communication technology—A summary of literature. *International Journal of Medical Informatics*, 76(9), 633-645.				1
5.	Altman 2001	Altman, D. G. (2001). Systematic reviews in health care: Systematic reviews of evaluations of prognostic variables. *British Medical Journal*, 323(7306), 224.			1	
6.	APA 2006	American Psychological Association (APA). (2006). Evidence based practice in psychology. *American Psychologist*, 61(May–June), 271–285.			1	
7.	APRAC 2006	Australian Department of Health and Ageing and National Health and Medical Research Council (2006). *Guidelines for a palliative approach in residential aged care – Enhanced version*. Canberra, AU: Edith Cowan University.				1
8.	Armstrong 2007	Armstrong, R., Waters, E., Jackson, N. et al. (2007). *Guidelines for systematic reviews of health promotion and public health interventions*. Version 2. Melbourne, AU: Melbourne University.		1		
9.	Arora 2010	Arora, S., Ashrafian, H., Davis, R., Athanasiou, T., Darzi, A., & Sevdalis, N. (2010). Emotional intelligence in medicine: A systematic review through the context of the ACGME competencies. *Medical Education*, 44(8), 749-764.				1
10.	Atkins 2008	Atkins, S., Lewin, S., Smith, H., Engel, M., Fretheim, A., & Volmink, J. (2008). Conducting a meta-ethnography of qualitative literature: Lessons learnt. *BMC Medical Research Methodology*, 8(1), 21.				1
11.	Atkins 2002	Atkins, C., & Sampson, J. (2002). *Critical appraisal guidelines for single case study research*. In: Proceedings of the Tenth European Conference on Information Systems, Gdansk, Poland. London: European Council of International Schools, 100–109.		1		

12.	Beck 2009	Beck, C. T. (2009). Critiquing qualitative research. *AORN Journal*, 90(4), 543-554.		1		
13.	Bennett 2011	Bennett, C., Khangura, S., Brehaut, J. C., Graham, I. D., Moher, D., Potter, B. K., et al. (2011). Reporting guidelines for survey research: An analysis of published guidance and reporting practices. *PLoS Medicine*, 8(8), e1001069.			2	
14.	Bhui 2003	Bhui, K., Stansfeld, S., Hull, S., Priebe, S., Mole, F., & Feder, G. (2003). Ethnic variations in pathways to and use of specialist mental health services in the UK. *British Journal of Psychiatry*, 182(2), 105-116.			1	
15.	Bowling 2009	Bowling, A. (2009). *Research methods in health: Investigating health and health services*. Open University Press, Maidenhead.			1	2
16.	Boyles 1998	Boyle, M. H. (1998). Guidelines for evaluating prevalence studies. *Evidence-Based Mental Health*, 1(2), 37-39.			4	
17.	Campbell & Cochrane Equity Methods Group checklist	Ueffing, E., Tugwell, P., Welch, V., Petticrew, M., Kristjansson, E., Campbell Equity Methods Group. (2011). *Equity checklist for systematic review authors*. Version 2011–11–08. The Cochrane Collaboration: Cochrane and Campbell Equity Methods Group.			1	
18.	Carter 2012	Carter, B. & Goodarce, L. (2012). Using evidence from qualitative studies, in: J. Craig, R. Smyth (Eds.), *The evidence-based practice manual for nurses*, Churchill Livingstone Elsevier (Chapter 4).		1		
19.	CASP	http://www.casp-uk.net/casp-tools-checklists		50	2	19
20.	CEBMA - Centre for evidence based management	Centre for Evidence Based Management. *Critical appraisal of a survey tool*. [Internet]. 2012 [cited 2012 Jan 27]. Available from: http://www.cebma.org/wp-content/uploads/Critical-AppraisalQuestions-for-a-Survey.pdf.			2	
21.	Clark 2003 (RATS Scale)	Clark, J.P. (2003). How to peer review a qualitative manuscript. In T. Jefferson & F. Godlee (Eds.), *Peer review in health sciences* (2nd ed., pp. 219–235). London: BMJ Books.		2		
22.	Classen 2006 (SPIDER)	Classen, S., Garvan, C. W., Awadzi, K., Sundaram, S., Winter, S., Lopez, E. D., et al. (2006). Systematic literature review and model for older driver safety. *Topics in Geriatric Rehabilitation*, 22(2), 87-98.				1
23.	Cochrane Risk of Bias (RoB) Tool	Higgins, J.P.T., & Green, S. (eds). (2009). *Cochrane handbook for systematic reviews of interventions* - Version 5.0.2 [updated September 2009]. The Cochrane Collaboration.			21	
24.	Cochrane Effective Practice and Organisation of Care	Cochrane Effective Practice and Organisation of Care (2009). *Risk of bias criteria for studies with a control group*. http://www.webcitation.org/5oB61IOub.			1	

25.	CONSORT	Schulz, K. F., Altman, D. G., Moher, D., for the CONSORT Group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *PLoS Medicine*, 7(3), e1000251.			2	
26.	COREQ	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.		6		
27.	Centre for Reviews and Dissemination (CRD)	NHS Centre for Reviews and Dissemination (2001). *Systematic reviews: CRD's guidance for undertaking reviews in health care*. 2nd ed. Edited by NHS Centre for Reviews and Dissemination. York: University of York. Khan, K.S. (2001). *Undertaking systematic reviews of research on effectiveness: CRD's guidelines for those carrying out or commissioning reviews*, Volume 2nd. York: University of York.			3	4
28.	Crombie 1996	Crombie, I. (1996). *The pocket guide to critical appraisal*. London: BMJ Publishing.			5	
29.	Cummings and Estrabooks	Cummings, G., & Estabrooks, C. A. (2003). The effects of hospital restructuring that included layoffs on individual nurses who remained employed: A systematic review of impact. *International Journal of Sociology and Social Policy*, 23(8/9), 8-53. Cummings, G., Lee, H., MacGregor, T., Davey, M., Wong, C., Paul, L., et al. (2008). Factors contributing to nursing leadership: A systematic review. *Journal of Health Services Research & Policy*, 13(4), 240-248. Estabrooks, C., Goel, V., Thiel, E., Pinfold, P., Sawka, C., & Williams, I. (2001). Decision aids: Are they worth it? A systematic review. *Journal of Health Services Research & Policy*, 6(3), 170-182. Estabrooks, C. A., Floyd, J. A., Scott-Findlay, S., O'leary, K. A., & Gushta, M. (2003). Individual determinants of research utilization: A systematic review. *Journal of Advanced Nursing*, 43(5), 506-520.			3	
30.	Deeks 2003	Deeks, J., Dinnes, J., D'amico, R., Sowden, A., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7(27), iii-x, 1-186.			1	
31.	Dixon-Wood 2004	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.		2		
32.	Dixon-Woods 2005	Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy*, 10(1), 45-53.		1		

33.	Dixon-Woods 2006	Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., et al. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology*, 6(1), 35.		1		3
34.	Down 1998 (Down & Black tool)	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.			6	1
35.	Drummond 2005	Drummond, M.F., Sculpher, M.J., Torrance, G.W. (2005). *Methods for the economic evaluation of health care programmes*. 3rd ed. Oxford: Oxford University Press.			4	
36.	EPOC	Cochrane Effective Practice and Organisation of Care Review Group. (2002). *The data collection checklist*. Retrieved January 24, 2005, from http://www.epoc.uottawa.ca/checklist2002.doc. Cochrane Effective Practice and Organisation of Care (2009). *Risk of bias criteria for studies with a control group*. Accessed: 2010-03-12 (archived by WebCite1 at www.webcitation.org/5oB61IOub).			3	
37.	Fowkes 1991	Fowkes, F., & Fulton, P. (1991). Critical appraisal of published research: Introductory guidelines. *British Medical Journal*, 302(6785), 1136.			2	
38.	Giacomini 2000	Giacomini, M. K., & Cook, D. J. (2000). Users' guides to the medical literature: XXIII. Qualitative research in health care A. Are the results of the study valid? *Journal of the American Medical Association*, 284(3), 357-362. Giacomini, M. K., & Cook, D. J. (2000). Users' guides to the medical literature: XXIII. Qualitative research in health care B. What are the results and how do they help me care for my patients? *Journal of the American Medical Association*, 284(4), 478-482.		3		
39.	Glaser 2005	Glaser, B. E., & Bero, L. A. (2005). Attitudes of academic and clinical researchers toward financial ties in research: A systematic review. *Science and Engineering Ethics*, 11(4), 553-573.			1	
40.	Glasziou 2001	Glasziou, P., Irwig, L., Bain, C., & Colditz, G. (2001*). Systematic reviews in health care: A practical guide*: Cambridge University Press.			1	
41.	Goldsmith 2007	Goldsmith, M. R., Bankhead, C. R., & Austoker, J. (2007). Synthesising quantitative and qualitative research in evidence-based patient information. *Journal of Epidemiology & Community Health*, 61(3), 262-270.				1
42.	Gough's Weight of Evidence	Gough, D. (2007). Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. In: J. Furlong, A. Oancea A (eds). *Applied and practice-based research*. (pp. 213–228). Special Edition of Research Papers in Education.				2

43.	Greenhalgh 1997	Greenhalgh, T. (1997). How to read a paper: Assessing the methodological quality of published papers. *British Medical Journal*, 315(7103), 305-308. Greenhalgh, T. (2010). *How to read a paper: The basics of evidence-based medicine*. John Wiley & Sons. Greenhalgh, T., & Taylor, R. (1997). How to read a paper: Papers that go beyond numbers (qualitative research). *British Medical Journal*, 315(7110), 740-743.		2		1
44.	Greenhalgh 2004	Greenhalgh T., Robert G., Bate P., Kyriakidou O., Macfarlane F. & Peacock R. (2004). *How to spread good ideas: A systematic review of the literature on diffusion, dissemination and sustainability of innovations in health service delivery and organization*. London: Report for the National Coordinating Centre for NHS Service Delivery and Organisation R&D (NCCSDO). Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., & Peacock, R. (2005). Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science & Medicine*, 61(2), 417-430.				3
45.	Greenwood 2009	Greenwood, N., Mackenzie, A., Cloud, G. C., & Wilson, N. (2009). Informal primary carers of stroke survivors living at home – Challenges, satisfactions and coping: A systematic review of qualitative studies. *Disability and Rehabilitation*, 31(5), 337-351.		1		
46.	Gysels 2007	Gysels, M., Higginson, I.J. (2007). Systematic reviews. In: Addington-Hall J, et al., eds. *Research methods in palliative care*. Oxford: Oxford University Press, 115–1134.			1	
47.	Harden 2009	Harden, A., Brunton, G., Fletcher, A., & Oakley, A. (2009). Teenage pregnancy and social disadvantage: Systematic review integrating controlled trials and qualitative studies. *British Medical Journal*, 339, b4254.		1		
48.	Harden 2004	Harden, A., Garcia, J., Oliver, S., Rees, R., Shepherd, J., Brunton, G., et al. (2004). Applying systematic review methods to studies of people's views: An example from public health research. *Journal of Epidemiology & Community Health*, 58(9), 794-800. Harden, A. (2006). Extending the boundaries of systematic reviews to integrate different types of study: Examples of methods developed within reviews on young people's health. In J. Popay (ed.), *Moving beyond effectiveness in evidence synthesis: Methodological issues in the synthesis of diverse sources of evidence* (pp. 31-40). London: National Institute for Health and Clinical Excellence.		1	1	
49.	Hawker 2002	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.		2		16

50.	Hayden 2006 (QUIPS)	Hayden, J. A., Côté, P., & Bombardier, C. (2006). Evaluation of the quality of prognosis studies in systematic reviews. *Annals of Internal Medicine*, 144(6), 427-437.			1	
51.	Health Evidence Bulletins Wales	Weightman, A. L., Mann, M. K., Sander, L., Turley, R. L. (2004). *Health evidence bulletins wales: A systematic approach to identifying the evidence - Project methodology 5*. Cardiff, Wales: National Public Health Service for Wales.			1	2
52.	Heller 2008	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.			1	
53.	Higginson 2002	Higginson, I. J., Finlay, I., Goodwin, D. M., Cook, A. M., Hood, K., Edwards, A. G., et al. (2002). Do hospital-based palliative teams improve care for patients or families at the end of life? *Journal of Pain and Symptom Management*, 23(2), 96-106.			2	
54.	Jackson 2006 (GATE)	Jackson, R., Ameratunga, S., Broad, J., Connor, J., Lethaby, A., Robb, G., et al. (2006). The GATE frame: Critical appraisal with pictures. *Evidence Based Medicine*, 11(2), 35-38.				1
55.	Jadad 1996	Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, 17(1), 1-12.			6	
56.	JBI	http://joannabriggs.org/research/critical-appraisal-tools.html		5	1	20
57.	Kearney 2001	Kearney M. (2001). Levels and applications of qualitative research evidence. *Research in Nursing & Health*, 24, 145–153.		1		
58.	Kelley 2003	Kelley K., Clark B., Brown V. & Sitzia J. (2003) Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care,* 15, 7261–266.			1	
59.	Kirk 2010	Kirk, S., Bone, M., Callery, P., Milnes, L., Pryjmachuk, S. (2010). *Evaluating self-care support for children and young people with long term conditions*. Available online at www.sdo.nihr.ac.uk/projdetails.php?ref=08-1715-162#: NIHR Service Delivery and Organisation programme.			1	
60.	Kirkevold 1997	Kirkevold, M. (1997). Integrative nursing research–An important strategy to further the development of nursing science and nursing practice. *Journal of Advanced Nursing,* 25(5), 977–984.				1
61.	Koufogiannakis 2006 (ReLIANT)	Koufogiannakis, D., Booth, A., & Brettle, A. (2006). ReLIANT: *Reader's guide to the literature in interventions addressing the need for education and training*. E-Prints in Library and Information Science [Online] Available from: http://eprints.rclis.org/handle/10760/8082 Accessed 02.01.12.				1

62.	Kmet 2004 (QualSyst)	Kmet, L., Lee, R., & Cook, L. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields*. Edmonton, AB: Alberta Heritage Foundation for Medical Research.				17
63.	Leboeuf-Yde 1995	Leboeuf-Yde, C., Lauritsen, J. (1995). The prevalence of low back pain in the literature: A structured review of 26 Nordic studies from 1954 to 1993. *Spine*, 20(19), 2112–2118.			1	
64.	Lethaby 2001	Lethaby, A., Wells, S., Furness, S., Strid, J., Arroll, B. & Milne, R. (2001). *Handbook for the preparation of explicit evidence-based clinical practice guidelines*. New Zealand Guidelines Group, Effective Practice Institute of the University of Auckland, Auckland, New Zealand.			1	
65.	Liverpool Quality Assessment Tool (LQAT)	Liverpool University Quality Assessment Tool (LQAT) in Voss, P. H., & Rehfuess, E. A. (2013). Quality appraisal in systematic reviews of public health interventions: An empirical study on the impact of choice of tool on meta-analysis. *Journal of Epidemiology in Community Health*, 67(1), 98-104.			1	
66.	Loney 2000	Loney, P. L., Chambers, W. L., Bennett, K. J., Roberts, J. G., & Stratford, P. (2000). Critical appraisal of the health research literature: Prevalence or incidence of a health problem. *Chronic Diseases in Canada*, 19(4), 170–176.			1	
67.	Long 2002	Long, A.F., Godfrey, M., Randall, T., Brettle, A., & Grant, M.J. (2003). *HCPRDU evaluation tool for quantitative studies*. University of Salford Health Care Practice Research and Development Unit.			1	1
68.	Malhora 1998	Malhotra, M. K., & Grover, V. (1998). An assessment of survey research in POM: From constructs to theory. *Journal of Operations Management*, 16(4), 407-425.			1	
69.	Malterud 2001	Malterud, K. (2001). Qualitative research: Standards, challenges, and guidelines. *The Lancet*, 358(9280), 483-488.		1		
70.	Mays 1995	Mays, N., & Pope, C. (1995) Rigour and qualitative research. *British Medical Journal*, 311, 109–112. Mays, N., & Pope, C. (2000) Assessing quality in qualitative research. *British Medical Journal*, 320, 50-52.		3		
71.	McCarthy 2008	McCarthy, G., & O'Sullivan, D. (2008). Evaluating the literature. In R. Watson, H. McKenna, S. Cowman & J. Keady (Eds.), *Nursing research: Designs and methods* (pp. 113-123). Edinburgh, Scotland: Churchill Livingstone.				1

72.	McMaster's Critical Review Form	Law, M., Stewart, D., Letts, L., Pollock, N., Bosch, J., & Westmorland, M. (1998). *Guidelines for critical review of qualitative studies*. McMaster University Occupational Therapy Evidence-Based Practice Research Group. Letts, L., Wilkins, S., Law, M., Stewart, D., Bosch, J., & Westmorland, M. (2007). *Guidelines for critical review form: Qualitative studies (Version 2.0)*. McMaster University Occupational Therapy Evidence-Based Practice Research Group.		3	1	9
73.	Miller 2003	Miller, R.M., Wilbourne, P.L., Hettema, J.E. (2003). What works? A summary of alcohol treatment outcome research. In R. K. Hester, W. R. Miller (eds), *Handbook of alcoholism treatment approaches*. Vol. 3rd ed. Boston: Pearson Education; scoring manual available at http://casaa.unm.edu/download/mesa.pdf2003				1
74.	Mills 2006	Mills, E. J., Nachega, J. B., Bangsberg, D. R., Singh, S., Rachlis, B., Wu, P., et al. (2006). Adherence to HAART: A systematic review of developed and developing nation patient-reported barriers and facilitators. *PLoS Medicine*, 3(11), e438.				2
75.	Mills 2005	Mills, E., Jadad, A. R., Ross, C., & Wilson, K. (2005). Systematic review of qualitative studies exploring parental beliefs and attitudes toward childhood vaccination identifies common barriers to vaccination. *Journal of Clinical Epidemiology*, 58(11), 1081-1088.		2		
76.	Mirza 2004	Mirza, I., & Jenkins, R. (2004). Risk factors, prevalence, and treatment of anxiety and depressive disorders in Pakistan: Systematic review. *British Medical Journal*, 328(7443), 794.			1	
77.	Mitchell 2001	Mitchell, E., & Sullivan, F. (2001). A descriptive feast but an evaluative famine: Systematic review of published articles on primary care computing during 1980-97. *British Medical Journal*, 322(7281), 279-282.			1	
78.	Moncrieff 2009 (Quality Rating Scale)	Moncrieff, J., Churchill, R., Drummond, D. C., & McGuire, H. (2001). Development of a quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods in Psychiatric Research*, 10(3), 126-133.				1
79.	Moule 2003	Moule, P., Pontin, D., Gilchrist, M., & Ingram, R. (2003). *Critical appraisal framework*, http://hsc.uwe.ac.uk/dataanalysis/critFrame.asp ©.				1
80.	National Service Framework	Department of Health (2005). *The National Service Framework for long term conditions. Annex 2: Research and evidence*. Available from: https://www.networks.nhs.uk/nhs-networks/vocational-rehabilitation/documents/DoH-NationalServiceFramework.pdf				1
81.	Newcastle Ottawa scale	Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2009). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses*. Ottawa Health Research Institute.			7	

82.	NHMRC	NHMRC. (2000). *How to use the evidence: assessment and application of scientific evidence*. https://www.nhmrc.gov.au/guidelines-publications/cp69. NHMRC. (2006). *Guidelines for a palliative approach to residential aged care: A systematic review of the literature*. Canberra: NHMRC.		1	1	1
83.	NICE	National Institute for Health and Clinical Excellence (NICE) (2009). *The guidelines manual*. London: NICE.		7		1
84.	Noyes 2008	Noyes, J., Popay, J., Pearson, A., et al. (2008) Qualitative research and Cochrane reviews. In J. Higgins & S. Green (eds). *Cochrane handbook for systematic reviews of interventions - Version 5.0.1*. The Cochrane Collaboration.		1		
85.	Ogrinc 2008 (SQUIRE)	Ogrinc, G., Mooney, S., Estrada, C., Foster, T., Goldmann, D., Hall, L. W., et al. (2008). The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: Explanation and elaboration. *Quality and Safety in Health Care*, 17(Suppl 1), i13-i32.				1
86.	PEDro scale	De Morton, N.A. (2009). The PEDro scale is a valid measure of the methodological quality of clinical trials: A demographic study. *Australian Journal of Physiotherapy*, 55, 129–133.			3	
87.	Pengel 2003	Pengel, L. H., Herbert, R. D., Maher, C. G., & Refshauge, K. M. (2003). Acute low back pain: Systematic review of its prognosis. *British Medical Journal*, 327(7410), 323.			1	
88.	Petticrew 2006	Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences*. Blackwell Publishing, Oxford.			5	
89.	Pluye 2009 (MMAT)	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.	7			20
90.	Polit and Beck 2004	Polit, D.F. & Beck, C.T. (2004). *Nursing research. Principles and methods*. Lippincott Williams & Wilkins, Philadelphia.			3	3
91.	Popay 1998	Popay, J., Rogers, A., & Williams, G. (1998). Rationale and standards for the systematic review of qualitative literature in health services research. *Qualitative Health Research*, 8(3), 341-351.		3		
92.	Pope 2002	Pope, C., van Royen, P., & Baker, R. (2002). Qualitative methods in research on healthcare quality. *Quality and Safety in Health Care*, 11(2), 148-152.		1		
93.	Ramsay 2003	Ramsay, C. R., Matowe, L., Grilli, R., Grimshaw, J. M., & Thomas, R. E. (2003). Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. *International Journal of Technology Assessment in Health Care*, 19(4), 613-623.			1	

94.	Reed 2009 (MERSQI)	Reed, D. A., Beckman, T. J., & Wright, S. M. (2009). An assessment of the methodologic quality of medical education research studies published in The American Journal of Surgery. *The American Journal of Surgery*, 198(3), 442-444.			1	2
95.	Rees 2001	Rees, R., Harden, A., Shepherd, J., Brunton, V., Oliver, S., & Oakley, A. (2001). *Young people and physical activity: A systematic review of barriers and facilitators*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.		1		
96.	Rees 2010	Rees, A., Beecroft, C., & Booth, A. (2010). Critical appraisal of the evidence. In Gerrish, K., Lacey, A. *The research process in nursing*. 6th ed. Chichester, UK: Wiley-Blackwell.			2	
97.	Regan 2013	Regan, J. L., Bhattacharyya, S., Kevern, P., & Rana, T. (2013). A systematic review of religion and dementia care pathways in black and minority ethnic populations. *Mental Health, Religion & Culture*, 16(1), 1-15.			1	
98.	Riesenberg 2009	Riesenberg, L. A., Leitzsch, J., Massucci, J. L., Jaeger, J., Rosenfeld, J. C., Patow, C., et al. (2009). Residents' and attending physicians' handoffs: A systematic review of the literature. *Academic Medicine*, 84(12), 1775-1787.				1
99.	Ross 2011 (SAQOR)	Ross, L., Grigoriadis, S., Mamisashvili, L., Koren, G., Steiner, M., Dennis, C. L., et al. (2011). Quality assessment of observational studies in psychiatry: An example from perinatal psychiatric research. *International Journal of Methods in Psychiatric Research*, 20(4), 224-234.				1
100.	Russell 2003	Russell, C. K., & Gregory, D. M. (2003). Evaluation of qualitative research studies. *Evidence-Based Nursing*, 6(2), 36-40.		1		
101.	Sanderson 2007	Sanderson, S., Tatt, I. D., & Higgins, J. (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.			1	
102.	Schneider 2004	Schneider, Z. (2004). Strategies for conducting a critical review. In Z. Schneider, D. Elliott, C. Beanland, G. LoBiondo-Wood & J. Haber (eds). *Nursing research: Methods, critical appraisal and utilization*, 2nd ed (pp. 52-72). Marrickville, AU: Elsevier.				1
103.	Shekelle 2010	Shekelle, P. G., Pronovost, P. J., Wachter, R. M., Taylor, S., Dy, S., Foy, R., et al. (2010). *Assessing the evidence for context-sensitive effectiveness and safety of patient safety practices: Developing criteria* (prepared under Contract No. HHSA-290-2009-10001C): Agency for Healthcare Research and Quality (AHRQ).				1
104.	Shepherd 2006	Shepherd, J., Harden, A., Rees, R., Brunton, G., Garcia, J., Oliver, S., et al. (2006). Young people and healthy eating: A systematic review of research on barriers and facilitators. *Health Education Research*, 21(2), 239-257.		1		1

105.	SIGN50	Scottish Intercollegiate Guidelines Network (SIGN) (2008). *SIGN 50: A guideline developer's handbook*. Edinburgh: Scottish Intercollegiate Guidelines Network.			3	1
106.	Sirriyeh 2012 (QATSDD)	Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice*, 18(4), 746-752.				2
107.	Sitzia 1998	Sitzia, J., & Wood, N. (1998). Response rate in patient satisfaction research: An analysis of 210 published studies. *International Journal for Quality in Health Care*, 10(4), 311-317.			1	
108.	Spencer 2003	Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*. London: National Centre for Social Research.		4		
109.	Stige 2009 (EPICURE)	Stige, B., Malterud, K., & Midtgarden, T. (2009). Toward an agenda for evaluation of qualitative research. *Qualitative Health Research*, 19(10), 1504-1516.		1		
110.	STROBE statement	*STROBE statement* (2013). http://www.strobe-statement.org/index.php?id=strobe-home. Switzerland: University of Bern.			8	1
111.	Thomas 2004	Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., et al. (2004). Integrating qualitative research with trials in systematic reviews. *British Medical Journal*, 328(7446), 1010.			1	
112.	Thomas 2003 (EPHPP)	Thomas, B.H., Ciliska, D., Dobbins, M., & Micucci, S. (2003). *Quality assessment tool for quantitative studies. Effective Public Health Practice Project*. Toronto: McMaster University. Thomas, B., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing*, 1(3), 176-184.			19	
113.	Tracy 2010	Tracy, S. J. (2010). Qualitative quality: Eight ''big-tent'' criteria for excellent qualitative research. *Qualitative Inquiry*, 16(10), 837-851.		1		
114.	Tranter 2012	Tranter, S., Irvine, F., & Collins, E. (2012). Innovations aimed at improving the physical health of the seriously mentally ill: An integrative review. *Journal of Clinical Nursing*, 21(9-10), 1199-1214.				1
115.	Treloar 2000 (CACQRS)	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.		1		

116.	UK health development agency's evidence based	Health Development Agency's website. *Evidence base-quality standards for evidence.* (http://www.HDAonline.org.uk/evidence/eb2000).				1
117.	US Preventive Services Task Force Work Group - AHRQ	Harris, R. P., Helfand, M., Woolf, S. H., Lohr, K. N., Mulrow, C. D., Teutsch, S. M., et al. (2001). Current methods of the US Preventive Services Task Force: A review of the process. *American Journal of Preventive Medicine*, 20(3), 21-35. Agency for Healthcare Research and Quality (2008). *US Preventive Services Task Force procedure manual*. www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.pdf. Agency for Healthcare Research and Quality (2011). *Methods guide for effectiveness and comparative effectiveness reviews*. AHRQ Publication No. 10(11)-EHC063-EF. Agency for Healthcare Research and Quality.			4	
118.	van Tulder Scale	Furlan, A. D., Pennick, V., Bombardier, C., & van Tulder, M. (2009). 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine*, 34(18), 1929-1941.			1	
119.	Wallace 2004	Wallace, A., Croucher, K., Quilgars, D., & Baldwin, S. (2004). Meeting the challenge: Developing systematic reviewing in social policy. *Policy & Politics*, 32(4), 455-470.		1		
120.	Walsh 2006	Walsh, D., & Downe, S. (2006). Appraising the quality of qualitative research. *Midwifery*, 22(2), 108-119.		3		
121.	Webber 2011	Webber, M. (2011). *Evidence-based policy and practice in mental health social work*, 2nd ed. Exeter: Learning Matters.			1	
122.	Weightman 2009	Weightman, A., Urquhart, C., Spink, S., & Thomas, R. (2009). The value and impact of information provided through library services for patient care: Developing guidance for best practice. *Health Information & Libraries Journal*, 26(1), 63-71.			1	
123.	Wendler 2011	Wendler, D., & Rid, A. (2011). Systematic review: The effect on surrogates of making treatment decisions for others. *Annals of Internal Medicine*, 154(5), 336-346.				1
124.	Whiting 2003 (QUADAS)	Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, 3(1), 25.			2	
TOTAL (SMSRs)			**7**	**118**	**163**	**156**

Acronym: MM: mixed methods study; QUAL: qualitative study; QUAN: quantitative study; SMSR: systematic mixed studies review.

APPENDIX 4. LIST OF REVIEWS ON CRITICAL APPRAISAL TOOLS

Reviews on Critical Appraisal Tools

Published reviews	Sources used in the reviews to identify tools	Number of tools identified
1. Armijo-Olivo et al. (2008)*	MEDLINE, Embase, CINAHL, ISI Web of Science, CCTR, Cochrane Library, Best Evidence, 2007, CDSR, ACP Journal Club, DARE, CCTR, Global Health, HealthSTAR (up to March 2007)	21 tools for RCTs
2. Bai, Shukla, Bak, and Wells (2012)	PubMed, MEDLINE, Embase, BIOSIS Previews, and The Cochrane Library (January 2000 to February 2005)	267 tools • 99 for observational studies • 94 for RCTs • 57 for systematic reviews • 17 for multiple designs 60 evidence grading systems
3. Crowe and Sheppard (2011)*	CSA Illumina, EBSCOhost, Gale InfoTrac, Informit ISI Web of Knowledge, JStore, OvidSP (CINAHL and MEDLINE), ProQuest, Scopus, The Cochrane Library (1996-2009)	44 tools • 11 for RCTs • 8 for qualitative studies • 6 for several designs • 5 for quantitative studies • 4 for experimental studies • 4 for systematic reviews • 2 for epidemiological studies • 2 for cohort studies • 1 for survey • 1 for single-case experimental design
4. Deeks et al. (2003)*	MEDLINE, Embase, PsycLit, Science Citation Index, Social Science Citation Index, Index to Scientific and Technical Proceedings, Applied Social Sciences Index and Abstracts, ERIC, British Education Index, Cochrane Review Groups, Citation searches, DARE, CRD and Cochrane Collaboration methodology databases, hand searching of 6 key journals, contact with methodological experts (up to 1999)	194 tools for non-randomized studies

5. Harder et al. (2014)*	Snowballing techniques from Bai et al. (2012) review. Identified 8 systematics reviews on quality appraisal tools (up to June 2012)	21 tools • 11 for observational studies • 4 for experimental studies • 3 for economic studies • 2 for qualitative studies • 1 for diagnostic test accuracy studies • 1 for animal studies
6. Heyvaert, Hannes, Maes, and Onghena (2013)	ASP, Allied and Complementary Medicine, British Education Index, CINAHL, Embase, ERIC, Francis, MEDLINE, PsycINFO, PubMed, Sociological Abstracts, CORDIS Library, Educational Technology and E-Learning, Grey Literature Database of the Canadian Evaluation Society, Index of Conference Proceedings, Index to Theses in Great Britain and Ireland, International Bibliography of the Social Sciences, ProQuest Dissertations & Theses, Social Science Research Network eLibrary, System for Information on Grey Literature in Europe, and Theses Canada, hand search in 10 journals (up to December 2009)	13 frameworks for mixed methods studies
7. Jarde, Losilla, and Vives (2012)*	MEDLINE, PsycINFO, CINAHL, Cochrane Library and Dissertation Abstracts International, Google (first 300 links) (up to beginning of 2010)	74 tools for non-experimental studies
8. Katrak, Bialocerkowski, Massy-Westropp, Kumar, and Grimmer (2004)*	Trip database, Clinical Evidence, Physiotherapy Evidence Database, OT Seeker, McMaster University Evidence-Based Practice Group, MEDLINE, Embase, CINAHL, Current Contents, The Cochrane Library, CDSR, DARE, CCTR, SIGN, NICE, NH&MRC, Google, Yahoo, MSN, Reference lists, contact content experts	120 tools • 45 for experimental studies • 26 for systematic reviews • 10 for all study designs • 19 for observational studies • 7 for diagnostic studies • 7 for qualitative studies • 6 for experimental and observational studies
9. Moher et al. (1999); Moher et al. (1995)*	MEDLINE, contact authors of scales and checklists (1966-1995)	34 tools for RCTs • 25 scales • 9 checklist
10. Neyarapally, Hammad, Pinheiro, and Iyasu (2012)	MEDLINE, Embase, Web of Science, Google Scholar (first 50 hits)	61 tools for pharmacoepidemiological safety studies

11. Sanderson et al. (2007)*	MEDLINE, Embase, Dissertation Abstracts, Google (up to March 2005)	86 tools for observational quantitative studies
12. Saunders, Soomro, Buckingham, Jamtvedt, and Raina (2003)*	MEDLINE (1966 to March 1999)	18 tools for non-randomized intervention
13. Shamliyan, Kane, and Dickinson (2010)*	MEDLINE, PubMed, Cochrane Library and Working Groups, WorldCat, Scirus (1966 to June 2008)	97 tools for observational studies on the incidence or prevalence
14. Wendt and Miller (2012)*	CINAHL, ERIC, LLBA, MEDLINE, PsycINFO, search engines and publisher specific databases including Google Scholar, Ixquick, ScienceDirect, Scirus, Scopus, SpringerLink	7 tools for single-subject experimental research
15. Walsh and Downe (2006)	Iterative process	8 tools for qualitative research
16. West et al. (2002)*	MEDLINE (1995 to mid-2000)	106 tools • 49 for RCTs • 20 for systematic reviews • 19 for observational studies • 18 for diagnostic studies
17. Zeng et al. (2015)	PubMed, Reference lists of published articles, Google (first 300 links), Cochrane Handbook, JBI Reviewers Manual, and the CRD guidance (up to May 2014)	21 tools • 6 for RCTs • 3 for analytical studies • 2 for non-randomized intervention studies • 2 for diagnostic accuracy studies • 1 for case series • 3 for animal studies • 3 for systematic reviews • 1 for guidelines

* Review provided information on measurement properties of the critical appraisal tools.

Acronyms: ACP: American College of Physicians Journal Club; ASP: Academic Search Premier; CCTR: Cochrane Central Register of Controlled Trials; CDSR: Cochrane Database of Systematic Review; CINAHL: Cumulative Index to Nursing and Allied Health Literature; CSA: Cambridge Scientific Abstracts; DARE: Database of Abstracts of Reviews of Effects; EBM: Evidence-Based Medicine Reviews; Embase: Excerpta Medica dataBASE; ERIC: Educational Resource Information Centre database; LLBA: Linguistics and Language Behaviour Abstracts; RCT: randomized controlled trial.

References

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.

Bai, A., Shukla, V. K., Bak, G., & Wells, G. (2012). *Quality assessment tools project report*. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health.

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

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., et al. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment, 7*(27), i-186.

Harder, T., Takla, A., Rehfuess, E., Sanchez-Vivar, A., Matysiak-Klose, D., Eckmanns, T., et al. (2014). Evidence-based decision-making in infectious diseases epidemiology, prevention and control: Matching research questions to study designs and quality appraisal tools. *BMC Medical Research Methodology, 14*(69), 1-16.

Jarde, A., Losilla, J.-M., & Vives, J. (2012). Methodological quality assessment tools of non-experimental studies: A systematic review. *Anales de Psicología, 28*(2), 617-628.

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

Moher, D., Cook, D., Jadad, A., Tugwell, P., Moher, M., Jones, A., et al. (1999). Assessing the quality of reports of randomised trials: Implications for the conduct of meta-analyses. *Health Technology Assessment, 3*(12), i-iv, 1-98.

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.

Neyarapally, G. A., Hammad, T. A., Pinheiro, S. P., & Iyasu, S. (2012). Review of quality assessment tools for the evaluation of pharmacoepidemiological safety studies. *BMJ Open, 2*(5), e001362.

Sanderson, S., Tatt, I. D., Higgins, J. P., 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.

Saunders, L. D., Soomro, G. M., Buckingham, J., Jamtvedt, G., & Raina, P. (2003). Assessing the methodological quality of nonrandomized intervention studies. *Western Journal of Nursing Research, 25*(2), 223-237.

Shamliyan, T., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology, 63*(10), 1061-1070.

Walsh, D., & Downe, S. (2006). Appraising the quality of qualitative research. *Midwifery, 22*(2), 108-119.

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*(2), 235-268.

West, S. L., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., et al. (2002). *Systems to rate the strength of scientific evidence*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ).

Zeng, X., Zhang, Y., Kwong, J. S., Zhang, C., Li, S., Sun, F., et al. (2015). The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: A systematic review. *Journal of Evidence-Based Medicine, 8*(1), 2-10.

APPENDIX 5. LIST OF CRITICAL APPRAISAL TOOLS WITH VALIDITY AND RELIABILITY TESTING

Critical appraisal tools with validity and reliability studies identified in the literature review

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
1. Al-Jader 2002	Prevalence surveys IV	Tool developed to rate genetic prevalence survey. Number of items: 8 on 5 categories (degree of ascertainment, population, cases, year, rate). Rating: Six items are rated 0 (inadequate) or 10 (exhaustive). Two items are rated 0 (inadequate), 10 (intermediate) and 20 (exhaustive). Overall score: 0 to 100.	x			x			(Al-Jader et al., 2002)
2. Bizzini scale	RCT IV	Tool to assess RCTs on nonoperative treatments for patellofemoral pain syndrome. Number of items: 15 on 4 categories (population, intervention, effect size, data presentation and analysis). Rating: yes, no. Overall score: 0 to 100.	x			x			(Bizzini et al., 2003)
3. Boejie checklist	QUAL III	Tool adjusted existing CATs for QUAL. Number of items: 10. Rating: 0 (item is absent), 1 (item is dealt with but weak), and 2 (item is satisfactorily dealt with). Overall score: 0 to 20.	x	x		x			(Boeije et al., 2011)
4. CCAT (Crowe Critical Appraisal Tool)	Generic I	Tool developed for appraising all types of research designs. Number of items: 22 on 8 categories (preliminaries, introduction, design, sampling, data collection, ethical matters, results, discussion). Each item can have several item descriptors. A total of 54 items descriptors were developed. Rating: present, absent, N/A. Overall score: 0 to 40.	x	x	x	x			(Crowe & Sheppard, 2011a, 2011b; Crowe et al., 2011, 2012)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
		User guide available at https://conchra.com.au/2015/12/08/crowe-critical-appraisal-tool-v1-4/							
5. Chalmers 1981	RCT IV	Tool to evaluate the design, implementation, and analysis of RCT. Number of items: 36 on 4 categories (basic description, study protocol, statistical analysis, presentation of results). Rating: different scales. Overall score: 0 to 1 (total score divided by the total possible score, i.e., by removing N/A items scores from the denominator).		x		x			(Bérard et al., 2000; Chalmers et al., 1981; Detsky et al., 1992)
6. Cho and Bero 1994	Observational and experimental studies III and IV	Two tools developed to assess the methodological quality and clinical relevance of drug studies. Number of items: 24 on methodological quality and 7 on clinical relevance. Rating: yes (2), partial (1), no (0), N/A. Overall score: 0 to 1. Time: approximately 30 min.		x		x			(Cho & Bero, 1994)
7. COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments checklist)	Studies on measurement properties of health measurement instruments IV	Tool developed to assess the methodological quality of articles on health measurement instruments. Number of items: contains 12 boxes: internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing, cross-cultural validity, criterion validity, responsiveness, interpretability, generalizability, and item response theory. The number of items varies in each box (ranging from 5 to 18). Rating: yes, no, ? (a 4-point scale was also developed: excellent, good, fair, poor).	x			x			(Mokkink et al., 2009, 2010a, 2010b, 2010c)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
		Overall score: variable depending on the boxes. User guide available at www.cosmin.nl/							
8. Delphi list	RCT IV	Tool developed from a 3-round Delphi study. Number of items: 9 on 3 dimensions (internal validity, external validity, statistical considerations). Rating: yes, no, don't know.	x	x		x			(Verhagen et al., 2000; Verhagen et al., 1998)
9. Detsky Scale	RCT IV	Tool to measure quality variation of RCT. Number of items: 5 categories (randomization, measure, selection criteria, intervention, statistical analysis). Each category includes 2 to 4 items. Rating: different scales, mainly yes/no. Overall score: 1 to 15.		x		x			(Colle et al., 2002; Detsky et al., 1992; Morrison et al., 2006)
10. DIAD (Design and Implementation Assessment Device)	Intervention studies III	Tool for assessing the quality of the design and implementation of a study in social sciences. Number of items: 4 global questions, 8 composite questions and 32-34 design and implementation questions. Rating: different scales, mainly yes/no and fully reasonable range/limited/not at all.	x			x			(Valentine & Cooper, 2008)
11. Downs & Black Quality Index	RCT and NRS III and IV	Tool developed for RCT and NRS. Number of items: 27 on 5 categories (reporting, external validity, bias, confounding, power). Rating: yes, no, unable to determine. Overall score: 0 to 31. Time: mean of 20 to 25 min (range: 10 to 45 min).	x	x		x	x	x	(Aubut et al., 2013; Downs & Black, 1998)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
12. EAI (Epidemiological Appraisal Instrument)	Epidemiological studies III	Tool for evaluating the methodological quality of existing or new ergonomic epidemiological studies. Epidemiological studies include cohort (prospective and retrospective), intervention (randomized and non-randomized), case-control, cross-sectional and hybrid (e.g., nested case-control). Number of items: 43 on 5 categories (reporting, subject/record selection, measurement quality, data analysis, generalization of results). Rating: yes (2), partial (1), no or unable to determine (0), N/A. Overall score: %	x	x	x	x		x	(Genaidy et al., 2007)
13. EPHPP (Effective Public Health Practice Project quality assessment tool)	RCT and NRS III and IV	Tool for appraising different designs of intervention studies for public health services. Number of items: 20 on 8 categories (selection bias, study design, confounders, blinding, data collection and methods, withdrawals and drop-outs, intervention integrity, analysis). Rating: different scales. Overall score: strong, moderate, weak.	x	x		x	x		(Armijo-Olivo et al., 2012; Thomas et al., 2004)
14. Hoy 2012	Prevalence studies IV	Tool to assess the risk of bias of prevalence studies on low back and neck pain. Number of items: 10. Rating: yes (low risk) or no (high risk). Overall: low, moderate or high risk of bias.	x			x			(Hoy et al., 2012)
15. IHE QA (Institute of Health Economics Quality Assessment)	Case series studies IV	Tool developed from a 3-round Delphi process with health technology assessment experts. Number of items: 20. Rating: yes, partial/unclear, no. Time: median of 15 min (range: 5 to 110 min).	x		x	x			(Guo et al., 2016; Moga et al., 2012)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
16. Imperiale 1990	Trials IV	Tool for the assessment of internal validity and reproducibility of trials on the impact of corticosteroids on the mortality of alcoholic hepatitis. Number of items: 5. Rating: + (specific); - (nonspecific or vague); 0.5+ (intermediate); ? (indeterminate). Overall score: 0 to 5.		x		x			(Colle et al., 2002; Imperiale & McCullough, 1990)
17. Jadad Scale	RCT IV	Tool developed to measure the likelihood of bias in pain research reports. Number of items: 3. Rating: 0 (no) or 1 (yes). Overall score: 0 to 5.	x	x	x	x	x		(Clark et al., 1999; Colle et al., 2002; Jadad et al., 1996; Oremus et al., 2012; Verhagen et al., 2000)
18. Maastricht list	RCT IV	Tool for assessing the methodological quality of RCT. Number of items: 15 main items divided into 47 subitems measuring 3 dimensions (internal validity, external validity and statistical consideration). Rating: + (presented and adequately done), - (presented but not adequately done or leading to bias), ? (presented but unclear), 0 (not presented). Weights were assigned to all items to reflect their relative importance. Overall score: 0 to 100 points.		x		x			(Brockow et al., 2000; de Vet et al., 1997; Verhagen et al., 2000)
19. MacLeashose 2000	RCT and NRS III and IV	Tool to appraise the quality of reporting, external validity, and internal validity (bias and confounding). Number of items: 23 with 8 items having subitems. Rating: different scales. Overall score: 0 to 18.	x		x	x		x	(MacLehose et al., 2000)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
20. MERSQI (Medical Education Research Study Quality Instrument)	QUAN III	Tool developed in the field of medical education and designed for experimental, quasi-experimental, and observational studies. Number of items: 10 on 6 domains (study design, sampling, type of data (subjective or objective), validity, data analysis, outcomes) Rating: A maximal score of 3 for each domain. Overall score: 0 to 18.	x	x	x	x	x	x	(Cook & Reed, 2015; Reed et al., 2007)
21. MetaQAT (Meta-tool for Quality Appraisal for Public Health Evidence)	Generic I	This tool consists in a meta-tool for public health. Authors specified that is not a critical appraisal tool but a quality assessment process. A companion tool was assembled from existing critical appraisal tools to provide study design-specific guidance on validity appraisal. Number of items: 9 on 4 domains: relevancy, reliability, validity, and applicability. The validity can be appraised using signalling questions or with existing appraisal tools. Rating: The scale is optional (yes, no, unclear, N/A).	x	x		x			(Rosella et al., 2016; Savage et al., 2016)
22. MINORS (Methodological Index for Non-Randomized Studies)	NRS III	Tool developed for surgery studies for non-randomized studies (both comparative and non-comparative). Number of items: 12. Rating: 0 (not reported), 1 (reported but inadequate) or 2 (reported and adequate). Overall score: 0 to 24.	x		x	x	x	x	(Slim et al., 2003)
23. MMAT (Mixed Methods Appraisal Tool)	RCT, NRS, descriptive, QUAL and mixed methods studies	Tool developed for use in systematic mixed studies reviews. Includes items for RCT, NRS, descriptive, qualitative and mixed methods studies. Number of items: 19 and 2 screening questions.	x			x			(Pace et al., 2012; Pluye et al., 2009; Souto et al., 2015)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
	III and IV	Rating: yes, no, can't tell. Overall score: 0, 25, 50 or 100% or stars (up to 4). Time: mean of 14 min (range: 4 to 40 min). User guide available at http://mixedmethodsappraisaltoolpublic.pbworks.com							
24. Moncrieff 2001	Controlled trials IV	Tool developed for psychiatric research. It covers aspects of both internal validity (or control of bias) and external validity (or generalizability). Number of items: 23 covering different aspects of quality including objective formulation, design, presentation of results, analysis and quality of conclusions. Rating: Two scales: 2-point (0, 2) and 3-point (0, 1, 2). Overall score: 0 to 46. Time: between 15 and 20 min.	x		x	x		x	(Moncrieff et al., 2001)
25. MORE (Methodological Evaluation of Observational REsearch) MEVORECH (Methodological EValuation of Observational REsearCH)	Observational studies of incidence, prevalence or risk factors IV	Developed two tools for the quality of observational studies (cohort, cross-sectional, and case-control studies) of incidence/prevalence or risk factors of chronic diseases. Number of items: The tool for studies of incidence or prevalence of chronic disease has 6 items for external validity and 5 for internal validity. The tool for risk factor studies had 6 criteria for external validity, 13 items for internal validity, and 2 aspects of causality. Rating: different response choices.	x		x	x			(Shamliyan et al., 2011)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
26. NOS (Newcastle Ottawa Scale)	Case-control and cohort studies IV	Tool for assessing the quality of case-control studies and cohort studies in meta-analyses. Number of items: 8 for case-control studies and 8 for cohort studies. Rating: Different scales. Overall score: stars system, maximum of 9 points.	x	x		x	x		(Cook & Reed, 2015; Hartling et al., 2013b; Lo et al., 2014; Oremus et al., 2012; Wells et al., 2000)
27. PEDro scale (Physiotherapy Evidence-based Database scale)	RCT IV	Tool to appraise the quality of RCTs in physiotherapy. Number of items: 10. Rating: yes, no. Overall score: 0 to 10.	x	x	x	x			(Aubut et al., 2013; de Morton, 2009; Foley et al., 2006; Maher et al., 2003; Moseley et al., 2002; Sherrington et al., 2000)
28. Psychotherapy Quality Rating Scale	RCT IV	Tool to assess the quality of RCTs of psychotherapy. Number of items: 25 on 6 domains (description of subjects, definition and delivery of treatment, outcome measures, data analysis, treatment assignment, overall quality of study). Rating: 0 (poor execution or description), 1 (moderately described and executed), 2 (well described and executed). Omnibus rating: 1=exceptionally poor, 2=very poor, 3=moderately poor, 4=average, 5=moderately good, 6=very good, 7=exceptionally good.	x	x	x	x		x	(Kocsis et al., 2010)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
29. QAREL (Quality Appraisal tool for studies of diagnostic RELiability checklist)	Studies of diagnostic reliability IV	Tool for appraising studies reporting the reliability of examination procedure. Number of items: 11. Rating: yes, no, unclear, N/A	x			x			(Lucas et al., 2013; Lucas, Macaskill, Irwig, & Bogduk, 2010)
30. QATSDD (Quality Assessment Tool for Studies with Diverse Designs)	Diverse studies II	Tool developed to be applied to a methodologically diverse set of research articles. Number of items: 16, in which 2 are specific to QUAL and 2 are specific to QUAN. When appraising a mixed methods research, all 16 items are rated. If only QUAN or QUAL, only 14 items are used. Rating: not at all, very slightly, moderately, complete. Overall score: 0 to 42.	x			x	x		(Sirriyeh et al., 2012)
31. QATSO (Quality assessment tool for systematic reviews of observational studies)	Observational designs III	Tool for assessing the quality of observational studies concerning HIV prevalence/risk behaviours among men having sex with men. Number of items: 5. Rating: different scales. Overall score: 0 to 100% (0-33%: bad; 34%-66%: satisfactory; 67%-100%: good)	x			x			(Wong et al., 2008)
32. Q-Coh (Quality of cohort studies)	Cohort studies IV	Tool for assessing the methodological quality of cohort studies in systematic reviews. Number of items: 26 items and 7 inferences on 7 domains (study design, representativeness, comparability of the groups, exposure measure, maintenance of comparability, outcome measure, attrition, statistical analyses). Rating: different scales.	x	x	x	x			(Jarde et al., 2013)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
		Overall score: good, acceptable, low. User guide available at http://www.tdx.cat/bitstream/handle/10803/116205/aj1de1.pdf?sequence=1							
33. QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies)	Diagnostic accuracy studies IV	This tool consists of an update of QUADAS. Number of items: 4 key domains of bias. Each domain has a set of signalling questions to help reach the judgments regarding bias and applicability. Rating: low, high or unclear risk of bias. User guide available at www.quadas.org	x			x			(Hollingworth et al., 2006; Mann et al., 2009; Schueler et al., 2012; Whiting et al., 2011, 2003; 2006)
34. QUIPS (Quality In Prognosis Studies Tool)	Prognosis studies IV	Tool developed for prognosis studies of low back pain using a modified Delphi approach and nominal group techniques. Number of items: 6 bias domains. Rating: yes, partly, no, unsure. Overall score: low, moderate or high risk of bias. Time: median of 20 min.	x			x			(Hayden et al., 2013)
35. RAC (Research Appraisal Checklist)	QUAN III	Tool designed for use with quantitative research reports. Number of items: 51 on 10 categories (title, abstract, problem, literature review, methodology, subjects, instruments, design, data analysis, form and style). Rating: 1 to 6 (1 or 2=not met; 3 or 4 = partially met; 5 or 6 = completely met), N/A. Overall score: 0 to 306. Time: approximately 30 min.	x			x		x	(Duffy, 1985, 2001)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
36. Reis 2007	QUAL III	Tool to assess the quality of qualitative work within the context of meta-ethnography or meta-synthesis. Number of items: 15. Rating: 0 (unable to rate); 1 (low/barely); 2 (moderate/moderately); 3 (high/clearly). Overall score: one item on global rating.	x			x			(Reis et al., 2007)
37. Reisch scale	Intervention studies III	Tool developed to facilitate the evaluation of the design and performance of therapeutic studies in medicine. Number of items: 34 on 13 general categories (purpose of study, experimental design, sample size determination, description and suitability of subjects, randomization and stratification, control, procedures for treatment, blinding, subject attrition, evaluation of subjects and treatment, presentation and analysis of data, recommendations and conclusions, overall study design and performance). Rating: yes, no, unclear or unknown, or N/A. Overall score: %		x		x			(Colle et al., 2002; Reisch et al., 1989; Tyson et al., 1983)
38. RoB (Cochrane Risk of Bias Tool)	RCT IV	Tool to assess the risk of bias in RCT. It focuses on the assessment of internal validity of studies. Number of items: 6 domains of bias (selection bias, performance bias, detection bias, attrition bias, reporting bias, and other bias). Rating: Yes, no, unclear. Overall score: low, high or uncertain risk of bias. User guide available at https://sites.google.com/site/riskofbiastool	x	x		x			(Armijo-Olivo et al., 2012, 2014; Hartling et al., 2009, 2013a; Vale et al., 2013)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
39. RoBANS (Risk of Bias Assessment tool for Non-randomized Studies)	NRS III	Tool developed for use in all study design except RCT. Use a domain-based evaluation approach. Number of items: 6 domains for risk of bias. Rating: low, high, unclear risk of bias. Time: mean of 9.50 min.	x	x	x	x			(Kim et al., 2013)
40. ROBINS-I (Risk Of Bias In Non-randomized Studies - of Interventions)	NRS III	This tool was previously named ACROBAT-NRSI (A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions). Number of items: 34 signalling questions on 7 domains of bias (confounding, selection of participants into the study, classification of the interventions, deviations from intended interventions, missing data, measurement of outcomes, selection of the reported result). Rating: yes, probably yes, no, probably no, no information. Overall score: low, moderate, serious and critical risk of bias. User guide available at: https://sites.google.com/site/riskofbiastool//	x			x			(Couto et al., 2015; Sterne et al., 2016)
41. RoBINT (Risk of Bias in N-of-1 Trials)	Single-case experimental design IV	This tool is an update of the SCED (Single-Case Experimental Design Scale). Number of items: 15. Rating: 0, 1, or 2. Time: mean of 26.2 min.	x		x	x			(Perdices et Tate, 2009; Tate et al, 2008; Tate et al, 2013)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
42. RTI-IB (Research Triangle Institute - Item Bank)	Observational studies of interventions or exposures III	Tool to appraise the quality of studies examining the outcomes of interventions, treatments, or exposures (cohort studies, case-control studies, case series, and cross-sectional studies). Number of items: 29 on 12 domains (background/context, sample definition and selection, interventions/ exposure, outcomes, creation of treatment groups, blinding, soundness of information, follow-up, analysis comparability, analysis outcome, interpretation, presentation and reporting). Rating: different scales. Time: mean of 48 min (range: 17 to 90 min).	x			x			(Viswanathan & Berkman, 2011, 2012)
43. SAQOR (Systematic Appraisal of Quality for Observational Research)	Observational studies III	Tool for the quality assessment of observational studies in reproductive psychiatry. Number of items: 19 on 5 categories (sample, control/comparison group, quality of measurements and outcomes, follow-up, distorting influences). Rating: yes, no, unclear, NA. Overall score: high, moderate, low, very low.	x			x			(Ross et al., 2011)
44. Shay 1972	Generic I	Tool used to appraise the quality of research articles, regardless of the methodology employed. Number of items: 25. Rating: completely incompetent, poor, mediocre, good and excellent.			x	x			(Shay et al., 1972)
45. Sindhu 1997	RCT IV	Tool developed from a 4-round Delphi study to rate the methodological quality of RCTs to be included in a meta-analysis. Number of items: 53 on 15 dimensions. Rating: Yes/No, weighting of the items. Overall score: 0 to 100.	x	x	x	x			(Sindhu et al., 1997)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
46. SPIDER (Systematic Process for Investigating and Describing Evidence-Based Research)	Several studies II	Tool developed for etiological systematic reviews in occupational therapy and other health-related disciplines. Number of items: 15 quality indicators classified into 9 quality themes (sampling and participation, statistical analysis, outliers/missing data, diagnostics, model fit, author limitations, validity, reliability, rationale). Rating: 1 (evidence existed), 0 (absence of evidence). Overall score: 1 (very poor) to 10 (excellent).	x	x	x	x	x		(Classen et al., 2008)
47. Van Tulder Scale	RCT IV	Tool developed by the Cochrane group on back pain. A first version of the tool was developed in 1997. The tool was updated in 2003. Number of items: 11 in the version published in 2003. Rating: yes, no, don't know.	x	x		x			(Colle et al., 2002; Van Tulder et al., 1997, 2003)
48. Vermeire 2002	Focus group IV	Tool for focus group research articles in primary healthcare. Number of items: 13. Rating: yes, no, can't tell. Overall score: high quality (yes or no). Time: median of 30 min, mean of 68 min.	x			x			(Vermeire et al., 2002)
49. Wells-Parker 1995	Intervention studies III	Tool for assessing methodological adequacy of studies on drinking/driving offenders. Number of items: 4 dimensions. Rating: 1 (ideal methods) to 7 (erroneous, invalid methods).	x		x	x			(Wells-Parker & Bangert-Drowns, 1990; Wells-Parker et al., 1995)

Tool	Type of studies Type of CATs*	Description	Validity			Reliability			References
			Content	Criterion	Construct	Interrater	Test-retest	Internal consistency	
50. Yang 2009	Case series IV	Tool to assess the quality of case series studies on herbal medicines. Number of items: 13 on 4 factors (study aims and design, description of treatment protocol, description of methods and therapeutic/side-effects, conduct of the study). Rating: 0 or 1. Time: 15 min. Overall score: 0 to 13.	x		x	x		x	(Yang et al., 2009)
51. Yates 2005	RCT IV	Tool for assessing the quality of reports of RCTs for psychological treatments. Number of items: 2 grids: one on treatment quality (6 items) and one on quality of study design and methods (20 items). Rating : adequate, partial, inadequate Overall score: 0 to 9 on the treatment quality scale and 0 to 26 on quality of study design and methods.	x	x	x	x			(Yates et al., 2005)
52. Zaza 2000	Intervention studies III	Tool developed to collect and evaluate the quality of execution of studies of intervention effectiveness. Number of items: 23 on 6 categories (descriptions, sampling, measurement, analysis, interpretation of results, other). Rating: yes, no, N/A. Time: 2 to 3 hours. Presents a study design algorithm.	x			x			(Zaza et al., 2000)

*Type of CATs: I – generic; II – generic with specific criteria, III – specific for category of studies; IV – specific for study designs.
Acronyms: CAT: critical appraisal tool; min: minutes; N/A: not applicable; NRS: non-randomized study; QUAL: qualitative study; QUAN: quantitative study; RCT: randomized controlled trial.

References

Al-Jader, L., Newcombe, R., Hayes, S., Murray, A., Layzell, J., & Harper, P. (2002). Developing a quality scoring system for epidemiological surveys of genetic disorders. *Clinical Genetics, 62*(3), 230-234.

Armijo-Olivo, S., Ospina, M., da Costa, B. R., Egger, M., Saltaji, H., Fuentes, J., et al. (2014). Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane Risk of Bias Tool in physical therapy trials. *PLoS One, 9*(5), e96920.

Armijo-Olivo, S., Stiles, C. R., Hagen, N. A., Biondo, P. D., & Cummings, G. G. (2012). Assessment of study quality for systematic reviews: A comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: Methodological research. *Journal of Evaluation in Clinical Practice, 18*(1), 12-18.

Aubut, J.-A. L., Marshall, S., Bayley, M., & Teasell, R. W. (2013). A comparison of the PEDro and Downs and Black quality assessment tools using the acquired brain injury intervention literature. *NeuroRehabilitation, 32*(1), 95-102.

Bérard, A., Andreu, N., Tétrault, J.-P., Niyonsenga, T., & Myhal, D. (2000). Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. *Annals of Epidemiology, 10*(8), 498-503.

Bizzini, M., Childs, J. D., Piva, S. R., & Delitto, A. (2003). Systematic review of the quality of randomized controlled trials for patellofemoral pain syndrome. *Journal of Orthopaedic & Sports Physical Therapy, 33*(1), 4-20.

Boeije, H. R., van Wesel, F., & Alisic, E. (2011). Making a difference: Towards a method for weighing the evidence in a qualitative synthesis. *Journal of Evaluation in Clinical Practice, 17*(4), 657-663.

Brockow, T., Hausner, T., Dillner, A., & Resch, K. (2000). Clinical evidence of subcutaneous CO2 insufflations: A systematic review. *The Journal of Alternative and Complementary Medicine, 6*(5), 391-403.

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

Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *Journal of the American Medical Association, 272*(2), 101-104.

Clark, H. D., Wells, G. A., Huët, C., McAlister, F. A., Salmi, L. R., Fergusson, D., et al. (1999). Assessing the quality of randomized trials: Reliability of the Jadad scale. *Controlled Clinical Trials, 20*(5), 448-452.

Classen, S., Winter, S., Awadzi, K. D., Garvan, C. W., Lopez, E. D., & Sundaram, S. (2008). Psychometric testing of SPIDER: Data capture tool for systematic literature reviews. *American Journal of Occupational Therapy, 62*(3), 335-348.

Colle, F., Rannou, F., Revel, M., Fermanian, J., & Poiraudeau, S. (2002). Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. *Archives of Physical Medicine and Rehabilitation, 83*(12), 1745-1752.

Cook, D. A., & Reed, D. A. (2015). Appraising the quality of medical education research methods: The Medical Education Research Study Quality Instrument and the Newcastle–Ottawa Scale-Education. *Academic Medicine, 90*(8), 1067-1076.

Couto, E., Pike, E., Torkilseng, E. B., & Klemp, M. (2015). *Inter-rater reliability of the Risk Of Bias Assessment Tool: For Non-Randomized Studies of Interventions (ACROBAT-NRSI)*. Paper presented at the 2015 Cochrane Colloquium Vienna.

Crowe, M., & Sheppard, L. (2011a). A general critical appraisal tool: An evaluation of construct validity. *International Journal of Nursing Studies, 48*(12), 1505-1516.

Crowe, M., & Sheppard, L. (2011b). A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89.

Crowe, M., Sheppard, L., & Campbell, A. (2011). 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.

Crowe, M., Sheppard, L., & Campbell, A. (2012). Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. *Journal of Clinical Epidemiology, 65*(4), 375-383.

de Morton, N. A. (2009). The PEDro scale is a valid measure of the methodological quality of clinical trials: A demographic study. *Australian Journal of Physiotherapy, 55*(2), 129-133.

de Vet, H. C., de Bie, R. A., van der Heijden, G. J., Verhagen, A. P., Sijpkes, P., & Knipschild, P. G. (1997). Systematic reviews on the basis of methodological criteria. *Physiotherapy, 83*(6), 284-289.

Detsky, A. S., Naylor, C. D., O'Rourke, K., McGeer, A. J., & L'Abbé, K. A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology, 45*(3), 255-265.

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 and Health Care, 6*(10), 538-540.

Duffy, M. E. (2001). Research appraisal checklist. *Measurement of Nursing Outcomes, 1*, 323-330.

Foley, N. C., Bhogal, S. K., Teasell, R. W., Bureau, Y., & Speechley, M. R. (2006). Estimates of quality and reliability with the physiotherapy evidence-based database scale to assess the methodology of randomized controlled trials of pharmacological and nonpharmacological interventions. *Physical Therapy, 86*(6), 817-824.

Genaidy, A., Lemasters, G., Lockey, J., Succop, P., Deddens, J., Sobeih, T., et al. (2007). An epidemiological appraisal instrument – A tool for evaluation of epidemiological studies. *Ergonomics, 50*(6), 920-960.

Guo, B., Moga, C., Harstall, C., & Schopflocher, D. (2016). A principal component analysis is conducted for a case series quality appraisal checklist. *Journal of Clinical Epidemiology, 69*, 199-207.e192.

Hartling, L., Hamm, M. P., Milne, A., Vandermeer, B., Santaguida, P. L., Ansari, M., et al. (2013a). Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *Journal of Clinical Epidemiology, 66*(9), 973-981.

Hartling, L., Milne, A., Hamm, M. P., Vandermeer, B., Ansari, M., Tsertsvadze, A., et al. (2013b). Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. *Journal of Clinical Epidemiology, 66*(9), 982-993.

Hartling, L., Ospina, M., Liang, Y., Dryden, D. M., Hooton, N., Krebs Seida, J., et al. (2009). Risk of bias versus quality assessment of randomised controlled trials: Cross sectional study. *British Medical Journal, 339*(7728), b4012.

Hayden, J. A., van der Windt, D. A., Cartwright, J. L., Côté, P., & Bombardier, C. (2013). Assessing bias in studies of prognostic factors. *Annals of Internal Medicine, 158*(4), 280-286.

Hollingworth, W., Medina, L. S., Lenkinski, R. E., Shibata, D. K., Bernal, B., Zurakowski, D., et al. (2006). Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool: A preliminary assessment. *Academic Radiology, 13*(7), 803-810.

Hoy, D., Brooks, P., Woolf, A., Blyth, F., March, L., Bain, C., et al. (2012). Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. *Journal of Clinical Epidemiology, 65*(9), 934-939.

Imperiale, T. F., & McCullough, A. J. (1990). Do corticosteroids reduce mortality from alcoholic hepatitis?: A meta-analysis of the randomized trials. *Annals of Internal Medicine, 113*(4), 299-307.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J. M., Gavaghan, D. J., et al. (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials, 17*(1), 1-12.

Jarde, A., Losilla, J.-M., Vives, J., & Rodrigo, M. F. (2013). Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology, 13*(2), 138-146.

Kim, S. Y., Park, J. E., Lee, Y. J., Seo, H.-J., Sheen, S.-S., Hahn, S., et al. (2013). Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology, 66*(4), 408-414.

Kocsis, J. H., Gerber, A. J., Milrod, B., Roose, S. P., Barber, J., Thase, M. E., et al. (2010). A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Comprehensive Psychiatry, 51*(3), 319-324.

Lo, C. K.-L., Mertz, D., & Loeb, M. (2014). Newcastle-Ottawa Scale: Comparing reviewers' to authors' assessments. *BMC Medical Research Methodology, 14*(5), 1-5.

Lucas, N., Macaskill, P., Irwig, L., Moran, R., Rickards, L., Turner, R., et al. (2013). The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Medical Research Methodology, 13*(111), 1-6.

Lucas, N. P., Macaskill, P., Irwig, L., & Bogduk, N. (2010). The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology, 63*(8), 854-861.

MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., & Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment, 4*(34), 1-154.

Maher, C. G., Sherrington, 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.

Mann, R., Hewitt, C. E., & Gilbody, S. M. (2009). Assessing the quality of diagnostic studies using psychometric instruments: Applying QUADAS. *Social Psychiatry and Psychiatric Epidemiology, 44*(4), 300.

Moga, C., Guo, B., Schopflocher, D., & Harstall, C. (2012). *Development of a quality appraisal tool for case series studies using a modified delphi technique*. Edmonton AB: Institute of Health Economics.

Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010a). Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Medical Research Methodology, 10*(82), 1-11.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010b). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology,* 10(22), 1-8.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010c). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539-549.

Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*(3), 313-333.

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

Morrison, L. J., Brooks, S., Sawadsky, B., McDonald, A., & Verbeek, P. R. (2006). Prehospital 12-lead electrocardiography impact on acute myocardial infarction treatment times and mortality: A systematic review. *Academic Emergency Medicine, 13*(1), 84-89.

Moseley, A. M., Herbert, R. D., Sherrington, C., & Maher, C. G. (2002). Evidence for physiotherapy practice: A survey of the Physiotherapy Evidence Database (PEDro). *Australian Journal of Physiotherapy, 48*(1), 43-49.

Oremus, M., Oremus, C., Hall, G. B., McKinnon, M. C., ECT, & Team, C. S. R. (2012). Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and Newcastle–Ottawa Scales. *BMJ Open, 2*(4), e001368.

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., et al. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies, 49*(1), 47-53.

Perdices, M., & Tate, R. L. (2009). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation, 19*(6), 904-927.

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.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *Journal of the American Medical Association, 298*(9), 1002-1009.

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.

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

Rosella, L., Bowman, C., Pach, B., Morgan, S., Fitzpatrick, T., & Goel, V. (2016). The development and validation of a meta-tool for quality appraisal of public health evidence: Meta Quality Appraisal Tool (MetaQAT). *Public Health, 136*, 57-65.

Ross, L., Grigoriadis, S., Mamisashvili, L., Koren, G., Steiner, M., Dennis, C. L., et al. (2011). Quality assessment of observational studies in psychiatry: An example from perinatal psychiatric research. *International Journal of Methods in Psychiatric Research, 20*(4), 224-234.

Savage, R. D., Rosella, L. C., Brown, K. A., Khan, K., & Crowcroft, N. S. (2016). Underreporting of hepatitis A in non-endemic countries: A systematic review and meta-analysis. *BMC Infectious Diseases, 16*(281), 1-12.

Schueler, S., Schuetz, G. M., & Dewey, M. (2012). The revised QUADAS-2 tool. *Annals of Internal Medicine, 156*(4), 323.

Shamliyan, T. A., Kane, R. L., Ansari, M. T., Raman, G., Berkman, N. D., Grant, M., et al. (2011). Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: Pilot study of new checklists. *Journal of Clinical Epidemiology, 64*(6), 637-657.

Shay, C. B., Zimmerman, W. S., & Michael, W. B. (1972). The factorial validity of a rating scale for the evaluation of research articles. *Educational and Psychological Measurement, 32*(2), 453-457.

Sherrington, C., Herbert, R., Maher, C., & Moseley, A. (2000). PEDro. A database of randomized trials and systematic reviews in physiotherapy. *Manual Therapy, 5*(4), 223-226.

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.

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2012). Reviewing studies with diverse designs: The development and evaluation of a new tool. *Journal of Evaluation in Clinical Practice, 18*(4), 746-752.

Slim, K., Nini, E., Forestier, D., Kwiatkowski, F., Panis, Y., & Chipponi, J. (2003). Methodological index for non-randomized studies (MINORS): Development and validation of a new instrument. *ANZ Journal of Surgery, 73*(9), 712-716.

Souto, R. Q., Khanassov, V., Hong, Q. N., Bush, P. L., Vedel, I., & Pluye, P. (2015). Systematic mixed studies reviews: Updating results on the reliability and efficiency of the mixed methods appraisal tool. *International Journal of Nursing Studies, 52*(1), 500-501.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal, 355*(i4919).

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.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., et al. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. Neuropsychological Rehabilitation, 23(5), 619-638.

Thomas, B., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176-184.

Tyson, J. E., Furzan, J. A., Reisch, J. S., & Mize, S. G. (1983). An evaluation of the quality of therapeutic studies in perinatal medicine. *The Journal of Pediatrics, 102*(1), 10-13.

Vale, C. L., Tierney, J. F., & Burdett, S. (2013). Can trial quality be reliably assessed from published reports of cancer trials: Evaluation of risk of bias assessments in systematic reviews. *British Medical Journal, 346(* f1798), 1-10.

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.

Van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., & Editorial Board of the Cochrane Collaboration Back Review Group (2003). Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine, 28*(12), 1290-1299.

van Tulder, M. W., Assendelft, W. J., Koes, B. W., Bouter, L. M., & Editorial Board of the Cochrane Collaboration Back Review Group (1997). Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for spinal disorders. *Spine, 22*(20), 2323-2330.

Verhagen, A. P., De Bie, R. A., Lenssen, A. F., De Vet, H. C., Kessels, A. G., Boers, M., et al. (2000). Quality assessment of trials: A comparison of three criteria lists. *Physical Therapy Reviews, 5*(1), 49-58.

Verhagen, A. P., de Vet, H. C., de Bie, R. A., Kessels, A. G., Boers, M., Bouter, L. M., et al. (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.

Vermeire, E., Van Royen, P., Griffiths, F., Coenen, S., Peremans, L., & Hendrickx, K. (2002). The critical appraisal of focus group research articles. *The European Journal of General Practice, 8*(3), 104-108.

Viswanathan, M., & Berkman, N. D. (2011). *Development of the RTI item bank on risk of bias and precision of observational studies*. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ).

Viswanathan, M., & Berkman, N. D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology, 65*(2), 163-178.

Wells-Parker, E., & Bangert-Drowns, R. (1990). Meta-analysis of research on DUI remedial interventions. *Alcohol, Drugs & Driving, 6*(3-4), 147-160.

Wells-Parker, E., Bangert-Drowns, R., McMillen, R., & Williams, M. (1995). Final results from a meta-analysis of remedial interventions with drink/drive offenders. *Addiction, 90*(7), 907-926.

Wells, G., Shea, B., O'connell, D., Peterson, J., Welch, V., Losos, M., et al. (2000). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Retrieved April 16, 2016, from /www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf.

Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 3*(25), 1-13.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., et al. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*(8), 529-536.

Whiting, P. F., Weswood, M. E., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. N., & Kleijnen, J. (2006). Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology, 6*(9), 1-8.

Wong, W. C., Cheung, C. S., & Hart, G. J. (2008). Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. *Emerging Themes in Epidemiology, 5*(1), 1-4.

Yang, A. W., Li, C. G., Da Costa, C., Allan, G., Reece, J., & Xue, C. C. (2009). Assessing quality of case series studies: Development and validation of an instrument by herbal medicine CAM researchers. *The Journal of Alternative and Complementary Medicine, 15*(5), 513-522.

Yates, S. L., Morley, S., Eccleston, C., & Williams, A. C. d. C. (2005). A scale for rating the quality of psychological trials for pain. *Pain, 117*(3), 314-325.

Zaza, S., Wright-De Agüero, L. K., Briss, P. A., Truman, B. I., Hopkins, D. P., Hennessy, M. H., et al. (2000). Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *American Journal of Preventive Medicine, 18*(Suppl 1), 44-74.

APPENDIX 6. ETHICS CERTIFICATE AND CONSENT FORMS

- Letter of approval from the Institutional Review Board of the Faculty of Medicine at McGill University
- Consent form of phase 1 of the project (qualitative descriptive study)
- Consent form of phase 2 of the project (modified e-Delphi study)

Consent form of phase 1 of the project (qualitative descriptive study)

You are being invited to participate in a research study titled 'Update and Measurement Properties of the Mixed Methods Appraisal Tool (MMAT)'. Before you decide, it is important for you to understand why the research is being done, how your information will be used, what the study will involve and the possible benefits and risks. Please take time to read the following information carefully and once you have been fully informed about the study and had any questions answered, you will be asked to agree if you wish to participate.

The purpose of this part of the research study is to know what you think about this tool based on your experience and if changes need to be made. This study is being done by Dr. Pierre Pluye and Ms. Quan Nha Hong from the McGill University, Canada. You were selected to participate in this study because you used the MMAT in a systematic review.

If you agree to take part in this study, you will be asked to participate to an interview that will last between 30 and 60 minutes.

We believe there are no known risks, side effects or disadvantage associated with this research study. Your participation in the study will contribute to improve and clarify the MMAT.

We will ensure the confidentiality of the information collected at the time of your participation in the study. The interview will be audio recorded only to facilitate the later analyses. The information provided, in no case, will be transmitted to other persons not involved in this project. Moreover, the recordings and transcriptions will be stored on a secure server during the period of the project. When the project is finished, the recordings will be destroyed. Your name will be preserved for ten years in a file separate from your research data and only accessible to the persons in charge of the project. A coded number will be used and no personally identifiable information will be presented in the data file used for analysis and reporting of results.

Your participation in this study is completely voluntary and you can withdraw at any time.

If you have questions about this project or if you have a research related problem, you may contact the researchers, Dr. Pierre Pluye or Ms. Quan Nha Hong at (514) 398-8483.

By clicking "I agree" below you are indicating that you are at least 18 years old, have read and understood this consent form and agree to participate in this research study. Please print a copy of this page for your records.

- ☐ I agree
- ☐ I don't agree

Consent form of phase 2 of the project (modified e-Delphi study)

The purpose of this study is to identify the most relevant methodological criteria for assessing the quality of studies in systematic reviews. This study is part of a project entitled 'Update and Measurement Properties of the Mixed Methods Appraisal Tool (MMAT)'. This study is being done by Dr. Pierre Pluye (full professor) and Ms. Quan Nha Hong (PhD candidate) from McGill University, Canada. The MMAT is designed for use in systematic reviews including qualitative, quantitative, and mixed methods studies. It includes criteria for assessing the quality of quantitative, qualitative, and mixed methods studies. The criteria identified in this study will inform the update of the MMAT.

You are being asked to participate in this study because of your expertise and experience in research methods. Your name was retained because you published methodological papers.

This study consists of an e-Delphi study composed of 2 to 3 rounds of web-based questionnaires. You will be asked to complete questions on the level of relevance of methodological quality criteria identified in a literature review. Each questionnaire should take around 30 minutes. After all panel members have completed the first round questionnaire, we will pool all responses anonymously, and use this information to develop the second questionnaire for you to complete. If necessary, we will repeat this one last time for a final consensus among experts.

We believe there are no known risks, side effects or disadvantage associated with this research study. Your participation in the study will contribute to identify the most relevant methodological criteria that need to be assessed when conducting a systematic review.

We will ensure the confidentiality of the information collected at the time of your participation in the study. The information provided, in no case, will be transmitted to other persons not involved in this project. Your questionnaire information will be coded, and no personally identifiable information will be present in the data file used for analysis and reporting of results. All the responses will remain anonymous to the panel. In the publications related with this Delphi study, your name will appear in the acknowledgement section for your contribution to the Delphi panel.

Your participation in this study is completely voluntary and you can withdraw at any time. If you have questions about this project or if you have a research-related problem, you may contact the researchers, Dr. Pierre Pluye (pierre.pluye@mcgill.ca) or Ms. Quan Nha Hong (quan.nha.hong@mail.mcgill.ca). For any questions about the rights of research participants, please contact the McGill Institutional Board: Ms. Ilde Lepore (ilde.lepore@mcgill.ca).

By clicking "I agree" below you are indicating that you have read and understood this consent form and agree to participate in this research study.

- ☐ I agree
- ☐ I don't agree

I agree to have my name stated in acknowledgment of my participation in this Delphi study in related publications.

- ☐ Yes
- ☐ No

APPENDIX 7. PROJECT PHASE 1 – INVITATION EMAIL AND INTERVIEW GUIDE

- Invitation email
- Interview guide

Invitation email

Subject: Invitation to participate in a research on the Mixed Methods Appraisal Tool (MMAT)

Dear…..,

We are currently conducting interviews as part of a study on the Mixed Methods Appraisal Tool (MMAT), a tool developed by Dr. Pierre Pluye and collaborators for appraising the methodological quality of qualitative, quantitative and mixed methods studies in systematic reviews. Briefly, the study aims to update the MMAT and test its validity and reliability.

As a researcher who had used the MMAT in a review, you are in an ideal position to give us valuable firsthand information on this tool. Your participation is important to help us understand the changes to be made to the MMAT.

The interview will take around 30 minutes. We are interested to know what you think about this tool (e.g., things you like and dislike about the MMAT, difficulties encountered, changes made or suggested). Your responses to the questions will be kept confidential. Prior to the interview, we will send you the consent form approved by the Institutional Review Board of McGill University.

If you are willing to participate, please let us know when would be a convenient time for you. If you have any questions, please do not hesitate to contact us: quan.nha.hong@mail.mcgill.ca.

We hope you will consider participating in this study.

Quan Nha

_____
Quan Nha HONG, OT, MSc
PhD candidate | CIHR doctoral fellow
Department of Family Medicine, McGill University
5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1
http://mixedmethodsappraisaltoolpublic.pbworks.com
http://toolkit4mixedstudiesreviews.pbworks.com

Interview guide

Thank you very much for taking to time to participate to this study on the Mixed Methods Appraisal Tool (or MMAT short name). Let me quickly introduce myself. I am a PhD student at McGill University and my supervisor is Dr. Pierre Pluye. He developed the MMAT nearly 10 years ago. We would like to update this tool and believe that the best starting point is to interview those who have used it. All comments you might have on this tool (good and bad) will be very valuable to have a tool that is relevant for the users. Do you have any questions of the project or consent form before I start the recording?

- My first questions aim to describe our study population and are about your experience in research.
    - What are your main research interests?
    - Around how many years of experience in research?
    - What study approaches do you usually use? Mainly qualitative, quantitative, mixed?
    - What is your experience with systematic review?
- What is your experience using the MMAT?
    - How did you find out the MMAT? (e.g., website, conference, colleagues, …)
    - In how many systematic reviews did you use the MMAT?
    - Around how many papers did you appraise using the MMAT?
    - The MMAT has 5 dimensions. Which criteria of the MMAT did you used?
    - What did you do with the results of the MMAT?
- What are the things you like about the MMAT and why?
    - On the items
    - On the scale (yes, no, can't tell)
    - On the instructions/guide/tutorial
- What are the things you dislike about the MMAT and why?
- Did you encounter any problems when using the MMAT?
    - If yes → What are the problems you encountered and how did you deal with them?
- Did you make any change to the tool during your project?
    - If yes → Could you describe the changes that were made?
    - If not → Would you suggest any changes to the MMAT?
- Were you able to appraise all the papers included in your reviews with the MMAT?
    - What type of papers could not be appraised with the MMAT?
- I think I have covered the questions I wanted to ask. Do you have any other comments on this tool?

If you think of any comments you might have on this tool that have not been addressed during the interview, please send me an email. All the comments on this tool will be useful for us to improve it.

Thank you very much for your time and participation!

APPENDIX 8. PROJECT PHASE 2 – INVITATION EMAILS AND QUESTIONNAIRES

- Invitation emails
- Questionnaires Round 1
    - Mixed methods research
    - Qualitative research
    - Survey research
- Questionnaires Round 2
    - Mixed methods research
    - Qualitative research
    - Survey research

Round 1 - Invitation email (sent to all potential participants)

Subject: Delphi study - Your expertise is needed for 30 minutes

Dear Dr. «LASTNAME»,

I am writing to invite you to participate in a Delphi study. Our objective is to identify the most relevant criteria for assessing the methodological quality of * research in systematic reviews (please see the attached letter for further information).

Your experience in * research gives you a unique expertise and perspective on this issue, and we would be very grateful if you would consider participating in this Delphi study.

Each expert on the panel will be asked to complete two to three rounds of questionnaires. Each questionnaire should take you no more than 30 minutes.

To participate, please send an email to my PhD candidate, Quan Nha Hong (quan.nha.hong@mail.mcgill.ca), and she will forward the instructions for the first round of the Delphi study.

Thank you very much for your consideration!

Yours sincerely,

Pierre Pluye, MD, PhD
Full Professor, FRQS Senior Research Scholar
Director, Methodological Developments, Quebec Support Unit
Department of Family Medicine, McGill University
5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1
Phone: 514-398-8483

*: depending of the experts, it was either written ‘mixed methods’, ‘qualitative’ or ‘survey’.

Round 1 - Invitation letter (sent to all potential participants)

Dr. «FIRSTNAME» «LASTNAME»
«INSTITUTION»

Subject: Invitation to participate in a Delphi study on the quality of * research

Dear Dr. «LASTNAME»,

I am writing to invite you to participate in a Delphi study that aims to identify the most relevant criteria for assessing the methodological quality of * research in systematic reviews. This study is part of a larger project aiming to update and validate a tool for appraising the methodological quality of studies in systematic reviews combining qualitative and quantitative evidence, the Mixed Methods Appraisal Tool (MMAT).

More and more mixed methods studies are included in systematic reviews. Yet, still few appraisal tools exist and the existing ones rarely include criteria for appraising the mixed methods component; they mainly have criteria for appraising the qualitative and quantitative components. We would like to gather a group of experts to identify the methodological criteria that characterize best the mixed methods component.

Your experience in * research gives you a unique expertise and perspective on this issue, and we would be very grateful if you would consider participating in this Delphi study. Your name was retained because you published methodological papers/textbooks on * research.

Each expert on the panel will be asked to complete two to three rounds of questionnaires to reach consensus. The first online questionnaire should take no more than 30 minutes. We will pool the responses to this first questionnaire anonymously (round 1 of the Delphi) and create a second version of the questionnaire for you to complete (round 2 of the Delphi). If necessary, we will conduct a third round. This study was approved by the ethic committee of McGill University (#A05-E26-15B) and all the responses will remain anonymous to the panel. In the publications related with this Delphi study, we will acknowledge your contribution as expert in * research.

To participate, please send an email to my PhD candidate, Quan Nha Hong (quan.nha.hong@mail.mcgill.ca), and she will forward the instructions for the first round of the Delphi.

Thank you very much for your consideration!

Yours sincerely,

Pierre Pluye, MD, PhD
Full Professor, FRQS Senior Research Scholar
Director, Methodological Developments, Quebec Support Unit

Round 1 – First reminder of invitation email (sent to non-respondents)

Subject: Reminder: Delphi study - Your expertise is needed for 30 minutes

Dear Dr. «LASTNAME»,

We recently sent you an invitation email regarding a Delphi study on the quality of * research in systematic reviews. As an expert in the field, your participation would be invaluable for this research; please consider participating. Here are the answers to some questions you might have about this study.

Why do a Delphi study on this topic? There is currently no common agreement on the most important criteria for appraising the quality of * research in systematic reviews. Delphi is a group technique suitable for establishing consensus among experts. It consists of sequential rounds of online questionnaires. After the first round, an anonymous summary of the group's responses will be provided and a second questionnaire will be sent.

Will it really take no more than 30 minutes? The experts who have completed the questionnaire so far took between 5 and 30 minutes with a mean of 15 minutes. Our research team preselected items from a literature review on critical appraisal tools. The questionnaire has 20 short methodological quality statements and a box for comments.

What is your timeline? Our plan is to complete Round 1 of the Delphi study by the end of March 2017 and start Round 2 in April 2017. If a consensus is not reached after two rounds, we will add one last round in May 2017.

Do I need to have experience in systematic reviews? No, you do not need to have experience in systematic reviews. You are being contacted as an expert in * research.

How is the confidentiality of participants assured? This study was approved by McGill University Institutional Review Board. Your questionnaire information will be coded, and no personally identifiable information will be present in the data file used for analysis and reporting of results. All the responses will remain anonymous to the panel. If you accept, your name will appear in the acknowledgement section for your contribution as expert to the Delphi panel in the publications related with this study.

How can I participate? Simply reply to this email. Quan Nha Hong (PhD candidate) will send you the link of the questionnaire.

Please do not hesitate to contact us if you have any further questions. Thank you very much for your time and consideration.

Sincerely,

Pierre Pluye, MD, PhD
Full Professor, FRQS Senior Research Scholar
Director, Methodological Developments, Quebec Support Unit

Round 1 – Second reminder of invitation email (sent to non-respondents)

Subject: Last reminder: Thanks for taking 15 minutes of your time to complete a questionnaire on the quality of * research

Dear Dr. «LASTNAME»,

We invited you to participate in a Delphi study on the quality of * research. Your expertise is very important to us. Thanks for considering helping us with this important study that aims to identify the most relevant criteria for appraising * research. This study is part of a PhD project.

Systematic reviews use explicit and rigorous methods to review the literature. One important step in systematic reviews is to appraise the quality of included studies. Traditionally, these reviews have mainly included randomized controlled trials. Over the past decade, more and more systematic reviews have included other types of studies such as * research.

It should take no more than 15 minutes to respond to the questionnaire, which consists of 20 short methodological statements. The questionnaire for the first round will remain open until March 31, 2017.

Please click on this web link to participate: SURVEYLINK

Your responses to this study will be kept anonymous to the panel. Your participation in this study is voluntary and you may withdraw at any time. We will acknowledge you, if you wish, in the publications related with this study.

If you have any questions about this study, please do not hesitate to contact us.

Best regards,

Pierre Pluye, MD, PhD
Full Professor, FRQS Senior Research Scholar
Director, Methodological Developments, Quebec Support Unit
Department of Family Medicine, McGill University

Round 1 - Reminder to complete the questionnaire (sent to those who have accepted to participate but have not completed the questionnaire yet)

Dear Dr. «LASTNAME»,

We recently sent you the link to the questionnaire for Round 1 of the Delphi study on the quality of * studies in systematic reviews. We noticed that you have not yet responded, and wish to remind you that it is still available. The experts who have completed the questionnaire so far took between 5 and 30 minutes with a mean of 15 minutes. We would greatly appreciate if you can take some time to complete the questionnaire before March 31, 2017.

To participate, please click on the link below: SURVEYLINK

Please do not hesitate to contact us if you have any questions.

Thank you very much for your time.

Sincerely,

Quan Nha

Quan Nha HONG, OT, MSc
PhD candidate | CIHR doctoral fellow
Department of Family Medicine, McGill University
5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1

Round 2 - Invitation email (sent to all participants of Round 1)

Subject: Delphi – Round 2 – Quality of * research

Dear Dr. «LASTNAME»,

Thank you very much for completing Round 1 questionnaire on the quality of * studies in systematic reviews. A total of XX experts in * research participated in Round 1 and provided very interesting comments on the criteria.

You will find attached a PDF with the group's responses and comments for each criterion, as well as a reminder of your responses (in blue). Based on the results of Round 1, we have removed XX criteria, clarified XX criteria and added XX new criteria.

The Round 2 questionnaire includes a total of XX criteria to rate. The format of this questionnaire is different from Round 1. Each question includes: the original criterion, group's responses from Round 1 (%), comments provided by the participants in Round 1, suggested revision, and question to answer. We estimate that it should take around 30 minutes to complete it.

In systematic reviews, the appraisal of the included studies is performed to judge whether the quality of studies is good enough. The results of this process are mainly used to: inform the synthesis, identify the strengths and limits of the included studies, determine how much confidence to have in the findings, and ensure that the recommendations and conclusions properly reflect the quality of studies.

Systematic reviews can include studies with different * research designs. In this Delphi project, we are seeking to identify the core generic criteria that are the most relevant. The identified criteria will be included in an appraisal tool. In addition, a manual guide will be developed to explain each criterion and provide hints on how to judge it.

Please click on this web link to start Round 2 questionnaire: SURVEYLINK

If you have any questions about this study, please do not hesitate to contact us.

We would greatly appreciate if you could complete Round 2 questionnaire by May 12, 2017.

Yours sincerely,

Quan Nha

Quan Nha HONG, OT, MSc
PhD candidate | CIHR doctoral fellow
Department of Family Medicine, McGill University
5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1

Round 2 – First reminder (sent to non-respondents)

Reminder: Delphi - Round 2 - Quality of * research

Dear Dr. «LASTNAME»,

We recently sent you the questionnaire for Round 2 of the Delphi study* on the quality of * research. We noticed that you have not yet completed the questionnaire, and wish to remind you that it is still available should you wish to take part.

Results of Round 1 helped to remove, clarify and add criteria as well as obtain group response. In Round 2, participants are asked to (re)rate the criteria using the group's comments and responses. Your participation is very important to reach a group consensus on the most relevant items for appraising the quality of * studies in systematic reviews.

To participate, please click here: SURVEYLINK

We would greatly appreciate if you could complete the Round 2 questionnaire by May 5, 2017. If more time is needed, please let us know.

Thank you for your time and participation,

Quan Nha

**Note: The Delphi technique consists of a group research approach to reach a consensus among a group of experts on an important issue that has limited or contradictory evidence. Typically, it is characterized by two rounds with controlled feedback, statistical group response, and anonymity.*

Quan Nha HONG, OT, MSc
PhD candidate | CIHR doctoral fellow
Department of Family Medicine, McGill University
5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1

Round 2 – Second reminder (sent to non-respondents experts in survey)

Last reminder: Round 2 of Delphi study on the quality of survey research

Dear Dr. «LASTNAME»,

Last month, we sent you the questionnaire for Round 2 of the Delphi study on the quality of survey research. In Round 2, participants are asked to (re)rate the criteria using the group's comments and responses. Thank you for helping us with this important study that aims to identify the most relevant criteria for appraising survey research. This study is part of a PhD project.

Participants who completed the questionnaire spent on average 20 minutes. The questionnaire for the
Round 2 will remain open until June 30, 2017.

Please click on this web link to participate: SURVEYLINK

Thank you in advance,

Pierre Pluye MD, PhD
Full Professor, FRQS Senior Research Scholar
Director, Methodological Developments, Quebec Support Unit
Department of Family Medicine, McGill University

5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1
Phone/tel: 514-398-8483
Email: pierre.pluye@mcgill.ca

Final email (sent to all participants of Round 2)

Dear Dr. «LASTNAME»,

We would like to inform you that the Delphi study is now completed. Thank you for your contribution. This is truly appreciated. Your responses will inform the update of a quality appraisal tool (Mixed Methods Appraisal Tool).

If you have agreed in the consent form, we will acknowledge your contribution to the Delphi panel in the related publications.

Wishing you a great summer!

Pierre and Quan Nha

Pierre Pluye, MD, PhD
Director, Method Development, Quebec SPOR Support Unit
FRQS Senior Research Scholar, Full Professor

Quan Nha HONG, OT, MSc
PhD candidate | CIHR doctoral fellow

Department of Family Medicine, McGill University
5858 Côte-des-Neiges, Suite 300
Montréal, QC, Canada, H3S 1Z1

Round 1 - Questionnaires

Delphi study – Round 1 – Quality of mixed methods studies

We are interested in identifying the criteria that are the most relevant for assessing the quality of mixed methods studies in systematic reviews.

In systematic reviews, quality appraisal is performed to judge the trustworthiness of included studies. The appraisal is generally based on methodological criteria.

You will find below a list of 20 methodological criteria. These criteria come from a recent literature review on the quality of mixed methods research (Fàbregues, S. & Molina-Azorin, J. (2016). Quality & Quantity, doi:10.1007/s11135-016-0449-4) and discussions among our research team.

Please rate the level of relevance of each of the following criteria for appraising the quality of mixed methods studies in systematic reviews.

Criteria	Not at all relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
1. A mixed methods research question (or purpose statement) is formulated.					
2. A rationale is provided for using a mixed methods design to address the research problem and questions.					
3. Key literature on mixed methods is reviewed in support of the mixed methods approach chosen by the authors.					
4. The mixed methods design is linked to the study aims and research questions.					
5. The mixed methods design matches the rationale given for combining quantitative and qualitative components.					
6. The mixed methods design is consistent with the epistemological assumptions of the study.					
7. Methods were selected to minimize shared bias.					
8. Quantitative and qualitative components of the study are effectively integrated.					

Criteria	Not at all relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
9. The type of integration of the quantitative and qualitative components matches the mixed methods design.					
10. The epistemological, ontological and teleological stances of the researcher that underlie the quantitative and quantitative approaches are successfully combined.					
11. Strategies for integrating phases, results and/or data are adequately performed.					
12. Methods are implemented in a way that remains true to the mixed methods design.					
13. The qualitative and quantitative components are linked in a cohesive and logical manner.					
14. Divergences and inconsistencies between quantitative and qualitative results are adequately addressed.					
15. Inferences derived from the quantitative and qualitative results are adequately incorporated in the meta-inferences regarding the entire study.					
16. Meta-inferences regarding the entire study are consistent with the rationale given for using a mixed methods design.					
17. The study contributes to advancing the field of mixed methods research.					
18. The added value gained from using a mixed methods design in this study is described.					
19. The strengths and weaknesses of methods optimize the breadth and depth of the study.					
20. Threats to the validity of quantitative, qualitative and mixed methods are identified and adequately addressed.					

If you have any further criteria that you believe are important for the appraisal of mixed methods studies in systematic reviews, please list below. Also, if you have any comments on the 20 criteria listed above, please mention them here.

Delphi study – Round 1 – Quality of qualitative studies

We are interested in identifying the criteria that are the most relevant for assessing the quality of qualitative studies in systematic reviews.

In systematic reviews, quality appraisal is performed to judge the trustworthiness of included studies. The appraisal is generally based on methodological criteria.

You will find below a list of 20 methodological criteria. These criteria come from a literature review on critical appraisal tools and discussions among our research team.

Please rate the level of relevance of each of the following criteria for appraising the quality of mixed methods studies in systematic reviews.

Criteria	Not at all relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
1. The research question can be answered using qualitative methodology and methods.					
2. The methods were adapted to fit the context of the study.					
3. The roles of the researchers in the data collection are adequately defined.					
4. The time, extent, and nature of the researcher's involvement in the data collection/analysis is appropriate for the method used.					
5. The sampling strategy is appropriately justified.					
6. The sample size is justified.					
7. The sample represents the diversity of the population for whom the research question is relevant.					
8. The characteristics of the participants relevant to the interpretation of the data are adequately described.					
9. The sites of recruitment are appropriate for addressing the purpose of the study.					
10. The sources of qualitative data (archives, documents, informants, observations) are relevant to address the research question.					
11. The data collection methods are					

Criteria	Not at all relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
appropriate to address the research question.					
12. The qualitative data analysis adequately addresses the research question.					
13. Appropriate explanation is given of how themes, concepts and categories were derived from the data.					
14. The data sources and processes of data collection, analysis and interpretation are coherent.					
15. Suitable strategies are used to verify the findings.					
16. The features of the sample critical to understand findings are described.					
17. Appropriate consideration is given to how findings relate to the context.					
18. Sufficient description of the data is given to allow understanding of the results, including the relevance of the context.					
19. The influence of the researchers on the data collection and analysis, results and interpretation is adequately considered.					
20. The interpretation of results is plausible and sufficiently substantiated with data.					

If you have any further criteria that you believe are important for the appraisal of qualitative studies in systematic reviews, please list below. Also, if you have any comments on the 20 criteria listed above, please mention them here.

Delphi study – Round 1 – Quality of survey research

We are interested in identifying the criteria that are the most relevant for assessing the quality of survey research in systematic reviews.

In systematic reviews, quality appraisal is performed to judge the trustworthiness of included studies. The appraisal is generally based on methodological criteria.

You will find below a list of 20 methodological criteria. These criteria come from a literature review on critical appraisal tools and discussions among our research team.

Please rate the level of relevance of each of the following criteria for appraising the quality of mixed methods studies in systematic reviews.

Criteria	Not at all relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
1. The target population is clearly defined.					
2. The study participants and the setting are described in detail.					
3. The list from which the sample is drawn is appropriate for answering the research question.					
4. The sampling strategy is relevant to address the research question.					
5. The study participants are adequately sampled.					
6. The sample is representative of the target population.					
7. The sample size is adequate.					
8. The sample size is based on pre-study considerations of statistical power.					
9. The same methods of data collection are used for all participants.					
10. Objective or standard criteria are used for the measurement of the parameter of interest.					
11. The choice of variables is based on their relevance and representativeness (content validity).					

Criteria	Not at all relevant	Slightly relevant	Moderately relevant	Very relevant	Extremely relevant
12. The variables are measured using known "gold standard", or using validated methods.					
13. The survey instrument has been piloted.					
14. The survey instrument has been tested for reliability.					
15. The survey instrument has been validated.					
16. The statistical analysis is appropriate to answer the research question.					
17. The sampling bias is adequately addressed in the analysis.					
18. All important confounding factors/subgroups/differences are identified and accounted for in the analysis.					
19. The response rate is adequate (if not, the low response rate is managed appropriately).					
20. The likelihood of nonresponse bias is minimal.					

If you have any further criteria that you believe are important for the appraisal of survey research in systematic reviews, please list below. Also, if you have any comments on the 20 criteria listed above, please mention them here.

Round 2 – Questionnaires

Delphi - Round 2 - Quality of mixed methods research

Thank you very much for participating in this study aimed to identify the most relevant methodological criteria for appraising the quality of mixed methods studies in systematic reviews.

The questionnaire of Round 2 is based on the feedback received in Round 1. Two criteria were removed and one was revised. Also, the participants have suggested new criteria. We have retained three generic methodological criteria (i.e., not specific to one mixed methods design or to a specific context).

In light of the results of Round 1, please rate the relevance of 21 criteria (18 original and 3 new) for appraisal in systematic reviews.

Note that we added the response category "I cannot tell", as requested by some participants.

The format of this questionnaire is different from Round 1. Each question is presented on one page and includes:

- Original criterion
- Group's responses (%)
- Comments provided by the participants
- Suggested revision, when applicable
- Question to answer (in bold)

Question 1

Group's responses and comments

1. A mixed methods research question (or purpose statement) is formulated.			
Answer choice	Count	Percentage	Comments
Not at all relevant	3	11.54%	• Many research questions could be addressed by either a mixed methods or a single method study. I don't think the mixing must be evident in the questions posed. • Questions are methods neutral. • I consider it very important that both research problem and questions or aims are clearly formulated - however, I do not wholly share the understanding proposed by some MM authors that there has to be QUAN and QUAL and MM research questions separately proposed.
Slightly relevant	2	7.69%	
Moderately relevant	3	11.54%	
Very relevant	8	30.77%	
Extremely relevant	10	38.46%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **A mixed methods research question (or purpose statement) is formulated.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 2

This question concerns the criteria 2, 4 and 5.

Group's responses and comments

2. A rationale is provided for using a mixed methods design to address the research problem and questions.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 2 and 4 and 5 seem quite similar. I think they could be usefully combined. OR if they are making important but separate, distinct points, then each needs clarification to signal its primary point.
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	8	30.77%	
Extremely relevant	17	65.38%	
4. The mixed methods design is linked to the study aims and research questions.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 2 and 4 and 5 seem quite similar. I think they could be usefully combined. OR if they are making important but separate, distinct points, then each needs clarification to signal its primary point. • Ambivalent about this (similar to #1)
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	5	19.23%	
Extremely relevant	20	76.92%	
5. The mixed methods design matches the rationale given for combining quantitative and qualitative components.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 2 and 4 and 5 seem quite similar. I think they could be usefully combined. OR if they are making important but separate, distinct points, then each needs clarification to signal its primary point.
Slightly relevant	1	3.85%	
Moderately relevant	2	7.69%	
Very relevant	4	15.38%	
Extremely relevant	19	73.08%	

Suggested revision

- We agree that these criteria are very similar and suggest retaining only criterion #2. In criterion #2, we added the term "clear" before "rationale".

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **A clear rationale is provided for using a mixed methods design to address the research problem and questions.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 3

Group's responses and comments

3. Key literature on mixed methods is reviewed in support of the mixed methods approach chosen by the authors.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Lots of good MM studies done without reference to MM literature.
Slightly relevant	4	15.38%	
Moderately relevant	10	38.46%	
Very relevant	5	19.23%	
Extremely relevant	7	26.92%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Key literature on mixed methods is reviewed in support of the mixed methods approach chosen by the authors.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 4

Group's responses and comments

6. The mixed methods design is consistent with the epistemological assumptions of the study.			
Answer choice	Count	Percentage	Comments
Not at all relevant	2	7.69%	• Although epistemological (etc.) issues are critical, I am assuming that systematic reviewers are working with published works that typically come with limited space. Hence, I expect few articles will dedicate space to these matters unless a philosophical focus is adopted. • Not clear. The MM design could include mixing at the paradigm or epistemological level, so the design is not really separate from the study's epistemological assumptions. Rather these assumptions can be part of the mix.
Slightly relevant	5	19.23%	
Moderately relevant	4	15.38%	
Very relevant	11	42.31%	
Extremely relevant	4	15.38%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The mixed methods design is consistent with the epistemological assumptions of the study.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 5

Group's responses and comments

7. Methods were selected to minimize shared bias.			
Answer choice	Count	Percentage	Comments
Not at all relevant	2	7.69%	• I was not entirely clear what you mean by the term "shared bias"
Slightly relevant	4	15.38%	• Sounds vague because it is unclear what is meant by "shared bias" (researchers', participants', or epistemological approach?)
Moderately relevant	8	30.77%	• • Time to let this go. This was an original rationale for mixing. But, given how many different types of and rationales for mixing we currently have, minimizing bias is not a universal criterion for a good MM study.
Very relevant	11	42.31%	
Extremely relevant	1	3.85%	• This aspect might not be relevant for all type of MM studies

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Methods were selected to minimize shared bias.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 6

Group's responses and comments

8. Quantitative and qualitative components of the study are effectively integrated.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• This relates to item 8 about the effective integration of qualitative and quantitative methods, but you might have also asked about whether either qualitative or quantitative methods in the research are given priority over the other. Or to word it as you would have above: One type of method (qualitative or quantitative) is not given excessive priority over the other in conducting the study and in discussing the results. • Related to many criteria, we need to get away from using the language of "qualitative and quantitative." First this language is imprecise; it could refer to types of data or methods, or to distinct inquiry paradigms. These are very different kinds of mixes. And the labels of "q" and"q" do not communicate effectively just what is being mixed in a mixed methods study. I rated these criteria (those with "q and q" language) within a more general framework about the character of the actual mix that takes place.
Slightly relevant	2	7.69%	
Moderately relevant	3	11.54%	
Very relevant	5	19.23%	
Extremely relevant	16	61.54%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Quantitative and qualitative components of the study are effectively integrated.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 7

Group's responses and comments

9. The type of integration of the quantitative and qualitative components matches the mixed methods design.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	3.85%	• I would like an option that states something like " relevance depends on research question." Not all methods should or can be integrated or need to be integrated." Methods can inform one another and not be integrated. I think it's important to define what you mean by integration since it is really important to many of your integration questions above. • Not sure what this one means (problem is in what is 'the MM design', especially as design often evolves).
Slightly relevant	1	3.85%	
Moderately relevant	3	11.54%	
Very relevant	8	30.77%	
Extremely relevant	13	50.00%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The type of integration of the quantitative and qualitative components matches the mixed methods design.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 8

Group's responses and comments

10. The epistemological, ontological and teleological stances of the researcher that underlie the quantitative and quantitative approaches are successfully combined.			
Answer choice	Count	Percentage	Comments
Not at all relevant	4	15.38%	• Very different philosophical stances may well be 'un-combinable.' So I think that is the wrong word. Perhaps respected or honored? Or perhaps the main point should not be combining or integrating two different paradigms, but rather that the stances of each are respected in the study, and even that respectful conversation across paradigms can take place. • No idea what teleological stances are. These are issues that most researchers don't think about once they are involved in the nitty-gritty of the study. So, whatever their basic philosophy is, is what implicitly will be influencing their approach to analysis and integration. • 10 was fuzzy as it referred to the single researcher, but often there are teams in action, etc.
Slightly relevant	5	19.23%	
Moderately relevant	8	30.77%	
Very relevant	7	26.92%	
Extremely relevant	2	7.69%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The epistemological, ontological and teleological stances of the researcher that underlie the quantitative and quantitative approaches are successfully combined.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 9

Group's responses and comments

11. Strategies for integrating phases, results and/or data are adequately performed.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	3.85%	No comment
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	8	30.77%	
Extremely relevant	16	61.54%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Strategies for integrating phases, results and/or data are adequately performed.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 10

Group's responses and comments

12. Methods are implemented in a way that remains true to the mixed methods design.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 12 and 13: this seems obvious for the most part. But, some of the best MM (and other) studies are ones that take advantage of a puzzle in the data and pursue it to an unexpected end and these studies can be quite messy rather than 'cohesive and logical.' • 12 and 16: MM studies are quite likely to diverge from original design, in the light of increasing understanding that develops during the research.
Slightly relevant	3	11.54%	
Moderately relevant	3	11.54%	
Very relevant	8	30.77%	
Extremely relevant	12	46.15%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Methods are implemented in a way that remains true to the mixed methods design.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 11

Group's responses and comments

13. The qualitative and quantitative components are linked in a cohesive and logical manner.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 12 and 13: this seems obvious for the most part. But, some of the best MM (and other) studies are ones that take advantage of a puzzle in the data and pursue it to an unexpected end and these studies can be quite messy rather than 'cohesive and logical.'
Slightly relevant	1	3.85%	
Moderately relevant	0	0.00%	
Very relevant	9	34.62%	• I think it is also important for the links between the different components to be made explicit, whether those links be at the level of data (e.g. common participants) and/or analysis (e.g. qualitising/quantising data) and/or interpretation.
Extremely relevant	16	61.54%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The qualitative and quantitative components are linked in a cohesive and logical manner.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 12

Group's responses and comments

14. Divergences and inconsistencies between quantitative and qualitative results are adequately addressed.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	No comment
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	14	53.85%	
Extremely relevant	11	42.31%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Divergences and inconsistencies between quantitative and qualitative results are adequately addressed.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 13

Group's responses and comments

15. Inferences derived from the quantitative and qualitative results are adequately incorporated in the meta-inferences regarding the entire study.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Meta-inferences are critical when a qualitative-quantitative binary is assumed; however, some full integration approaches may reject this assumption and meta-inferences could then be less important. I don't think there is consensus on this issue but do know that research questions and application as are as varied as the imagination of researchers who conduct studies. Hence, I would focus less on meta-inference per se and more on the logic of mixing. • I think that a joint display, including metainferences, is a state of the art procedure for integrating mixed methods data. I know this concerns systematic reviews, but I am wondering if there is a way to include this in the review process.
Slightly relevant	1	3.85%	
Moderately relevant	1	3.85%	
Very relevant	13	50.00%	
Extremely relevant	11	42.31%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Inferences derived from the quantitative and qualitative results are adequately incorporated in the meta-inferences regarding the entire study.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 14

Group's responses and comments

16. Meta-inferences regarding the entire study are consistent with the rationale given for using a mixed methods design.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• My concern with criteria 16 is that sometimes unanticipated findings and meta-inferences may emerge. It is possible that they were not consistent with what was originally conceived. • 12 and 16: MM studies are quite likely to diverge from original design, in the light of increasing understanding that develops during the research.
Slightly relevant	2	7.69%	
Moderately relevant	5	19.23%	
Very relevant	13	50.00%	
Extremely relevant	6	23.08%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Meta-inferences regarding the entire study are consistent with the rationale given for using a mixed methods design.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 15

Group's responses and comments

17. The study contributes to advancing the field of mixed methods research.			
Answer choice	Count	Percentage	Comments
Not at all relevant	4	15.38%	• Although this is relevant for methodologists, I do not see this idea of contributing to mixed methods within my areas of research. Researchers outside of methodology circles use these designs to address questions and problems in their fields of Study from pragmatic perspectives. I don't think they're often concerned with the field of mixed methods as they are with their own areas of research being pushed further. With that in mind I don't think this criteria can be universally used to assess rigor in systematic reviews in all disciplines. • I do NOT feel that making a methodological contribution is a required component of an excellent systematic review. It is desirable and laudable, but not required. • The statement may not be applicable to all MMR studies because the study may help advance the knowledge on a specific research topic but not necessarily use MMR in an innovative manner to advance the field of MMR. • We do not put this burden on practitioners of other methodologies -- like surveys or quasi-experimentation. I would not favor this for the MM field at this time. • Not relevant - assuming the study is of a substantive (rather than methodological) topic.
Slightly relevant	9	34.62%	
Moderately relevant	7	26.92%	
Very relevant	5	19.23%	
Extremely relevant	1	3.85%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The study contributes to advancing the field of mixed methods research.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 16

Group's responses and comments

18. The added value gained from using a mixed methods design in this study is described.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	No comment
Slightly relevant	1	3.85%	
Moderately relevant	8	30.77%	
Very relevant	12	46.15%	
Extremely relevant	5	19.23%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The added value gained from using a mixed methods design in this study is described.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 17

Group's responses and comments

19. The strengths and weaknesses of methods optimize the breadth and depth of the study.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I think we should make judgments about the strengths and weaknesses of how the methods were applied in the study vs. their "inherent" weaknesses. • I have a problem with describing methods as having weaknesses (cf Sandelowski, in T&T handbook, 2003) – they are just differences.
Slightly relevant	3	11.54%	
Moderately relevant	9	34.62%	
Very relevant	12	46.15%	
Extremely relevant	2	7.69%	

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The strengths and weaknesses of methods optimize the breadth and depth of the study.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 18

Group's responses and comments

20. Threats to the validity of quantitative, qualitative and mixed methods are identified and adequately addressed.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	3.85%	• A key and important point. But using the language of "validity" restricts the meaning of this to a post-positivist standpoint. So, as written, this is not a viable criterion by which to assess MM quality.
Slightly relevant	1	3.85%	
Moderately relevant	3	11.54%	
Very relevant	8	30.77%	
Extremely relevant	13	50.00%	

Suggested revision

- We suggest replacing the term "validity" with "trustworthiness".

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Threats to the trustworthiness of quantitative, qualitative and mixed methods are identified and adequately addressed.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 19 - New criterion suggested

Comments provided by the participants:

- Another criterion I believe is important is the use of rigorous qualitative and quantitative components. Each should be well-articulated and follow acceptable procedures.
- Having done systematic review of mixed methods, a primary concern I continue to observe is weakness in individual phases. For instance, some researchers rely entirely on descriptive stats in a QUAN phase or open ended questions after a survey in a QUAL phase. This needs to somehow be addressed in discussions about rigour more explicitly so editors and authors alike understand that mixed methods is not reducible to an afterthought of throwing one under-developed method alongside another. I think this can be made more explicit in item 19 - where you address strength and weakness optimization.
- The quality of the quantitative and qualitative components.
- Transparency of the quantitative and qualitative parts.
- Use rigorous and systematic procedures for data collection and analysis in quantitative and qualitative study phases to address weakness minimization.
- Apply validation strategies recommended for quantitative and qualitative research approaches in quantitative and qualitative study phases.

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **Rigorous procedures for data collection and analysis are used in quantitative and qualitative components.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 20 - New criterion suggested

Comments provided by the participants:

- I recommend the addition of items that recognize the idea of what could be referred to as interpretive comprehensive. That refers to purposefully seeking out diverse perspective. It's a construct that is consistent with Greene's (2007) mixed methods way of thinking. It embraces the idea of mixing paradigms, rather than being afraid of it.

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The study purposefully seek out diverse perspectives (interpretive comprehension).**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 21 - New criterion suggested

Comments provided by the participants:

- I would also add a criterion that says something like, "The study generated findings and insights that would not have been possible with a mono-method study."

Please rate the level of relevance of this criterion for appraising the quality of mixed methods studies in systematic reviews: **The mixed methods study generated findings and insights that would not have been possible with a mono-method study.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Thank you very much for completing the questionnaire!

Delphi study - Round 2 - Quality of qualitative research

Thank you very much for participating in this study aimed to identify the most relevant methodological criteria for appraising the quality of qualitative studies in systematic reviews.

The questionnaire of Round 2 is based on the feedback received in Round 1. Two criteria were removed and 15 were modified. Also, participants suggested new criteria and we retained the generic methodological criteria (i.e., not on a specific design or topic). Some criteria were integrated with the existing criteria and two new criteria were added.

In light of the results of Round 1, please rate the relevance of 21 criteria (19 original and 2 new).

Note that we added the response category "I cannot tell", as requested by some participants.

The format of this questionnaire is different from Round 1. Each question is presented on one page and includes:

- Original criterion
- Group's responses from Round 1 (%)
- Comments provided by the participants in Round 1
- Suggested revision, when applicable
- Criterion to rate (in bold)

Question 1

Group's responses and comments:

1. The research question can be answered using qualitative methodology and methods.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Two pronged question – methodology, and methods. What if the study only has qual methods, without a methodology – many do (unfortunately). Similarly, you can have a qual methodology with both QUAL and QUAN methods (e.g., ethnography). • The questions in your criteria seem to presuppose a fixed research question from the start and unidirectional research process. It's far better to start with a very general open-ended question and then to refine the research question to fit the setting or problem as the researcher becomes intimately familiar with it. Qualitative research is an emergent process and the most valuable findings often are not anticipated. For example, following leads in the data can require adding or changing research sites or participants.
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	3	11.54%	
Extremely relevant	21	80.77%	
I cannot answer	1	3.85%	

Suggested revision:

- We revised this criterion and removed "methodology and methods".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **A qualitative approach is appropriate to answer the research question.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 2

Group's responses and comments:

2. The methods were adapted to fit the context of the study.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I do not understand completely the intent of questions #2, 4, 12, and 14. • More needs to be asked, for example, about the contextual relations between the researcher and the subjects/materials of research, as these are key to the interpretations being offered in the research.
Slightly relevant	2	7.69%	
Moderately relevant	3	11.54%	
Very relevant	5	19.23%	
Extremely relevant	16	61.54%	
I cannot answer	0	0.00%	

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The methods were adapted to fit the context of the study.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 3

Group's responses and comments:

3. The roles of the researchers in the data collection are adequately defined.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• The use of the term appropriate (above too - #4): Why is it not in #6? And, why is 'adequate' ok for #3? Do you want it 'justified' or 'appropriately justified' or adequate? • New criterion suggested: The role(s) of researcher(s) are discussed in terms of their assumptions, biases, and position as insider/outsider relative to phenomenon, participants, and/or setting. • Another weak question concerns the 'roles' of researchers - the notion of role and what is implied by this does not seem to sufficiently respond to the central importance of the researcher in the analysis and interpretation; it doesn't adequately tap into what theoretical and epistemological orientation the researchers bring to the questions asked and to the interpretations achieved. Rather I would ask if the interpretations offered and concepts generated align coherently with the theorization of the subject matter, and if the theorization is clear and convincing.
Slightly relevant	3	11.54%	
Moderately relevant	6	23.08%	
Very relevant	8	30.77%	
Extremely relevant	8	30.77%	
I cannot answer	1	3.85%	

Suggested revision:

- We replaced this criterion with the one suggested by a participant.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The role(s) of researcher(s) are discussed in terms of their assumptions and position as insider/outsider relative to the phenomenon, participants, and/or setting.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 4

Group's responses and comments:

4. The time, extent, and nature of the researcher's involvement in the data collection/analysis is appropriate for the method used.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I do not understand completely the intent of questions #2, 4, 12, and 14. • This question has a three-pronged question (time, extent, and nature) and a two-pronged one (data collection and analysis). It cannot be answered with just one response.
Slightly relevant	2	7.69%	
Moderately relevant	2	7.69%	
Very relevant	10	38.46%	
Extremely relevant	11	42.31%	
I cannot answer	1	3.85%	

Suggested revision:

- We removed "time, extend and nature" from the criterion.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The researcher's involvement in the data collection and analysis is appropriate for the method used.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 5

Group's responses and comments:

5. The sampling strategy is appropriately justified.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	3.85%	• "The use of the term appropriate (above too - #4): Why is it not in #6? And, why is 'adequate' ok for #3? Do you want it 'justified' or 'appropriately justified' or adequate? • Many times in qualitative research, especially ethnography, the "sample" comes first and then the research questions and themes are determined based upon what the context or sample makes available. As such, typical criteria like "was the sample appropriate for the research question" should be flipped on their head to read something like, "Were the findings and research direction appropriate given the sample or context."
Slightly relevant	2	7.69%	
Moderately relevant	2	7.69%	
Very relevant	8	30.77%	
Extremely relevant	12	46.15%	
I cannot answer	1	3.85%	

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The sampling strategy is appropriately justified.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 6

Group's responses and comments:

6. The sample size is justified.			
Answer choice	Count	Percentage	Comments
Not at all relevant	4	15.38%	• "The sample size is justified" is an odd way to address sample size. We don't actually talk about "sample size" in QUAL the way one does in QUAN. The number of participant interviews, observations or documents is determined by saturation, that is, how long one needs to observe or how many people need to be interviewed is unique to a particular study. Sample size is determined by the emerging findings feeling "saturated" - that is, continued data collection and analysis (which should be simultaneous) reveals no new information. • The use of the term appropriate (above too - #4): Why is it not in #6? And, why is 'adequate' ok for #3? Do you want it 'justified' or 'appropriately justified' or adequate? • Emphasis is on the sample (size, characteristics) when often it is not the subject-person involved that is relevant as much as the situations/contexts/institutions that situate individual practices/conceptions. a better question would concern whether the SAMPLING UNIT is the correct one. (may not be the individual for example). Meaning and sense (often the objects of inquiry in qualitative research) are socially produced making the situation the key issue of sampling rather than attributes of the individual.
Slightly relevant	2	7.69%	
Moderately relevant	5	19.23%	
Very relevant	8	30.77%	
Extremely relevant	7	26.92%	
I cannot answer	0	0.00%	

Suggested revision:

- We uniformized the terminology and used the term "appropriate". We also added "for the research design".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The sample size is appropriate for the research design.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 7

Group's responses and comments:

7. The sample represents the diversity of the population for whom the research question is relevant.			
Answer choice	Count	Percentage	Comments
Not at all relevant	6	23.08%	• The use of the term 'represents' suggests a quantitative sample (and not participant selection, as is more relevant to qual research) • Item 7 is not quite clear: do you mean representative for the population that the study addresses or the people for whom the study might be relevant? These are different or can be different. • In qualitative research you would never talk about 7. The sample "representing" the diversity of the population
Slightly relevant	1	3.85%	
Moderately relevant	9	34.62%	
Very relevant	4	15.38%	
Extremely relevant	5	19.23%	
I cannot answer	1	3.85%	

Suggested revision:

- We replaced the "population for whom the research question is relevant" with "people for whom the study is relevant".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The sample represents the diversity of the people for whom the study is relevant.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 8

This question concerns criteria #8 and #16.

Group's responses and comments:

8. The characteristics of the participants relevant to the interpretation of the data are adequately described.			
Answer choice	Count	Percentage	Comments
Not at all relevant	3	11.54%	• A number of qualifiers (adequate, appropriate, sufficiently, ...) are used.
Slightly relevant	3	11.54%	
Moderately relevant	3	11.54%	
Very relevant	5	19.23%	
Extremely relevant	12	46.15%	
I cannot answer	0	0.00%	

16. The features of the sample critical to understand findings are described.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	3.85%	• The sample is not usually 'critical' for the findings. Also, what do you mean by 'features'?
Slightly relevant	1	3.85%	
Moderately relevant	3	11.54%	
Very relevant	10	38.46%	
Extremely relevant	10	38.46%	
I cannot answer	1	3.85%	

Suggested revision:

- These two criteria overlap. We combined them.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The characteristics of the sample relevant to the interpretation of the findings are appropriately described.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 9

Group's responses and comments:

9. The sites of recruitment are appropriate for addressing the purpose of the study.			
Answer choice	Count	Percentage	Comments
Not at all relevant	2	7.69%	• The issue of 'sites' seems like a red herring. If you use snowballed sampling, for example, you don't have a 'site'. • Criterion #9 is also a bit puzzling ("sites of recruitment" are appropriate). In qual we talk about "purposive" or "purposeful" sampling - one "samples" sites or participants from which the most can be learned about the phenomenon of interest. Of course that would mean they are "appropriate" but I've never seen the phrase "sites of recruitment" in a report of qual research.
Slightly relevant	2	7.69%	
Moderately relevant	7	26.92%	
Very relevant	8	30.77%	
Extremely relevant	6	23.08%	
I cannot answer	1	3.85%	

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The sites of recruitment are appropriate for addressing the purpose of the study.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 10

Group's responses and comments:

10. The sources of qualitative data (archives, documents, informants, observations) are relevant to address the research question.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• "Clarify please: do you mean: e.g., archives, documents, informants, observations. Or, are these the only options? (e.g., missing is: participant-observation). Also, how are 'archives' different from documents? Informant is a dated term – participant is more respectful. "
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	10	38.46%	
Extremely relevant	14	53.85%	
I cannot answer	1	3.85%	

Suggested revision:

- We added the term "such as" in the criterion. Also, we replaced the term "relevant" with "appropriate".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The sources of qualitative data (such as archives, documents, participant observation, etc.) are appropriate to address the research question.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 11

Group's responses and comments:

11. The data collection methods are appropriate to address the research question.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I can't answer this unless I know what that the methods align with the methodology. • It is hard not to think any of the above are not really important. I look for all of these when reviewing a manuscript. I like to think that you should start with the stories and research questions in choosing a method for addressing the question. Item 11 is close, but might be more specific if it were phrased as "qualitative data collection are most appropriate" -- you can use qualitative data to address a question that might be better addressed with quantitative or mixed methods.
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	5	19.23%	
Extremely relevant	19	73.08%	
I cannot answer	1	3.85%	

Suggested revision:

- We modified the criterion as suggested by a participant.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The qualitative data collection methods are most appropriate to address the research question.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 12

Group's responses and comments:

12. The qualitative data analysis adequately addresses the research question.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I do not understand completely the intent of questions #2, 4, 12, and 14. • Analysis does not 'address' a research question. It makes sense of data which is gathered based upon a design with flows from a research question. • New criterion suggested: Methods of data analysis properly addressed i.e. Not just quoting a software program.
Slightly relevant	0	0.00%	
Moderately relevant	2	7.69%	
Very relevant	11	42.31%	
Extremely relevant	12	46.15%	
I cannot answer	1	3.85%	

Suggested revision:

- As suggested by a participant, we replaced "adequately addresses the research question" with "methods are appropriately addressed".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The qualitative data analysis methods are appropriately addressed.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 13

Group's responses and comments:

13. Appropriate explanation is given of how themes, concepts and categories were derived from the data.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Themes, concepts and categories: what is the difference? Each researcher uses these terms differently. • #13 seems to be specific to grounded theory than to qualitative research in general.
Slightly relevant	1	3.85%	
Moderately relevant	2	7.69%	
Very relevant	9	34.62%	
Extremely relevant	13	50.00%	
I cannot answer	1	3.85%	

Suggested revision:

- We replaced "themes, concepts and categories" with "findings".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **Appropriate explanation is given for how findings (such as themes, concepts, categories, etc.) were derived from the data.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 14

Group's responses and comments:

14. The data sources and processes of data collection, analysis and interpretation are coherent.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Too many constructs in one sentence. • I do not understand completely the intent of questions #2, 4, 12, and 14. • I don't really know what "coherent" means, in #14--or how it would be judged. • Another issue concerns questions about the 'fit' between the research question and the data collection strategy and the analysis; this is not a binary, straightforward matter. Often the most insightful and useful qualitative research is that that has been capable of discovering in the course of the research that the question being asked is not the most important, relevant or productive one.
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	9	34.62%	
Extremely relevant	15	57.69%	
I cannot answer	1	3.85%	

Suggested revision:

- The main construct of this criterion is "coherence". We revised this criterion.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **There is coherence between qualitative data sources, collection, analysis and interpretation.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 15

Group's responses and comments:

15. Suitable strategies are used to verify the findings.			
Answer choice	Count	Percentage	Comments
Not at all relevant	2	7.69%	• Verify? • #15. I marked this as moderately relevant as I was unsure what was meant by verify. If this refers to being able to determine the "truth" of findings, whether an incident that a participant described "really happened," I view this as less important (slightly to moderately), as the point of qualitative studies is often not to identify a single truth, or verify accuracy of events, but to describe multiple truths and realities, acknowledging that participants are providing constructions of and their perspectives of events, which will vary from one person to the next. If "verify" refers to properly substantiating findings with data and evidence from study, then I view this as extremely important. • New criterion suggested: Triangulation of qualitative data sources is used to strengthen findings • New criterion suggested: Any use of multiple coders, analysts, and/or interpreters to increase confidence in findings is explained.
Slightly relevant	3	11.54%	
Moderately relevant	4	15.38%	
Very relevant	6	23.08%	
Extremely relevant	10	38.46%	
I cannot answer	1	3.85%	

Suggested revision:

- We modified this criterion as suggested by one participant. The word "verify was removed and examples of strategies were added.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **Strategies (such as prolonged engagement, peer review, etc.) are used to strengthen the findings.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 16

This question concerns criteria #17 and #18.

Group's responses and comments:

17. Appropriate consideration is given to how findings relate to the context.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	3.85%	• I think context is important but handled differently in qual. • #17 and #18 overlap. But findings is replaced by results. • Criterion #17, "how findings relate to the context" might be better stated as "how findings illuminate the phenomenon under study" (or something like this). If your study is about an internal process such as transformative learning, or meditation, or love for example, "context" doesn't seem to fit.
Slightly relevant	2	7.69%	
Moderately relevant	4	15.38%	
Very relevant	11	42.31%	
Extremely relevant	7	26.92%	
I cannot answer	1	3.85%	

18. Sufficient description of the data is given to allow understanding of the results, including the relevance of the context.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• #17 and #18 overlap. But findings is replaced by results.
Slightly relevant	2	7.69%	
Moderately relevant	1	3.85%	
Very relevant	8	30.77%	
Extremely relevant	14	53.85%	
I cannot answer	1	3.85%	

Suggested revision:

- We agree with the comment that these two criteria overlap. We retained criterion #17. We added an example of what is meant by context.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **Appropriate consideration is given to how findings relate to the context (such as the setting where the data were collected, etc.).**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 17

Group's responses and comments:

19. The influence of the researchers on the data collection and analysis, results and interpretation is adequately considered.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Too many constructs • New criterion suggested: Qualitative investigator's experience, knowledge, qualifications, reflexivity, and relationship to the study is included. • More needs to be asked, for example, about the contextual relations between the researcher and the subjects/materials of research, as these are key to the interpretations being offered in the research.
Slightly relevant	4	15.38%	
Moderately relevant	2	7.69%	
Very relevant	7	26.92%	
Extremely relevant	12	46.15%	
I cannot answer	1	3.85%	

Suggested revision:

- We replaced the term "adequately" with "appropriately".

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The influence of the researcher(s) on the data collection and analysis, results and interpretation is appropriately considered.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 18

Group's responses and comments:

20. The interpretation of results is plausible and sufficiently substantiated with data.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Two-pronged question • On item 20, I would add that it is important that the interpretation is not only plausible and well substantiated, but that it directly relates to the conceptual framework, and that it does not go beyond what the data supports - so many authors try to stretch the conclusions beyond what can truly be supported. This is also true in post-positivist research, so I think it is endemic to academic work in general. • Q.20 asks two different questions. They look independent to me rather than one being an aspect of the other.
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	3	11.54%	
Extremely relevant	21	80.77%	
I cannot answer	1	3.85%	

Suggested revision:

- We divided this criterion into two questions.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The interpretation of results is plausible.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 19

Group's responses and comments:

20. The interpretation of results is plausible and sufficiently substantiated with data.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Two-pronged question • On item 20, I would add that it is important that the interpretation is not only plausible and well substantiated, but that it directly relates to the conceptual framework, and that it does not go beyond what the data supports - so many authors try to stretch the conclusions beyond what can truly be supported. This is also true in post-positivist research, so I think it is endemic to academic work in general. • Q.20 asks two different questions. They look independent to me rather than one being an aspect of the other.
Slightly relevant	0	0.00%	
Moderately relevant	1	3.85%	
Very relevant	3	11.54%	
Extremely relevant	21	80.77%	
I cannot answer	1	3.85%	

Suggested revision:

- We divided this criterion into two questions.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews: **The interpretation of results is sufficiently substantiated with data.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 20 – New criterion suggested

Comments provided by the participants:

- Any relevant epistemological or theoretical framework used is appropriately explained and justified (phenomenology, social construction, grounded theory, etc).

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews:

Any relevant epistemological or theoretical framework used is appropriately explained and justified.

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 21 – New criterion suggested

Comments provided by the participants:

- More needs to be asked, for example, about the contextual relations between the researcher and the subjects/materials of research, as these are key to the interpretations being offered in the research.

Please rate the level of relevance of the following criterion for appraising the quality of qualitative studies in systematic reviews:

The contextual relations between the researcher(s) and the participants (and/or materials) of research are appropriately addressed.

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Thank you very much for completing the questionnaire!

Delphi study - Round 2 - Quality of survey research

Thank you very much for participating in this study aimed to identify the most relevant methodological criteria for appraising the quality of survey research in systematic reviews.

The questionnaire of Round 2 is based on the feedback received in Round 1. Three criteria were removed and 9 were modified. Also, participants suggested new criteria and three criteria specific to survey were retained.

In light of the results of Round 1, please rate the relevance of 20 criteria (17 original and 3 new).

Note that we added the response category "I cannot tell", as requested by some participants.

The format of this questionnaire is different from Round 1. Each question is presented on one page and includes:

- Original criterion
- Group's responses from Round 1 (%)
- Comments provided by the participants in Round 1
- Suggested revision, when applicable
- Criterion to rate (in bold)

Question 1

Group's responses and comments:

1. The target population is clearly defined.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	No comment
Slightly relevant	0	0.00%	
Moderately relevant	0	0.00%	
Very relevant	4	19.05%	
Extremely relevant	17	80.95%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The target population is clearly defined.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 2

Group's responses and comments:

2. The study participants and the setting are described in detail.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	No comment
Slightly relevant	0	0.00%	
Moderately relevant	1	4.76%	
Very relevant	10	47.62%	
Extremely relevant	10	47.62%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The study participants and the setting are described in detail.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 3

Group's responses and comments:

3. The list from which the sample is drawn is appropriate for answering the research question.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	No comment
Slightly relevant	0	0.00%	
Moderately relevant	0	0.00%	
Very relevant	6	28.57%	
Extremely relevant	15	71.43%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The list from which the sample is drawn is appropriate for answering the research question.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 4

This question concerns criteria #4 and #5.

Group's responses and comments:

4. The sampling strategy is relevant to address the research question.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I actually do not know what 4 and 5 mean in practice. I took them to mean that subjects are chosen by probabilistic methods and not self-selected, but the standards really are ambiguous. • I don't understand well the difference between question 4 and 5.
Slightly relevant	0	0.00%	
Moderately relevant	1	4.76%	
Very relevant	6	28.57%	
Extremely relevant	14	66.67%	

5. The study participants are adequately sampled.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I actually do not know what 4 and 5 mean in practice. I took them to mean that subjects are chosen by probabilistic methods and not self-selected, but the standards really are ambiguous. • I don't understand well the difference between question 4 and 5. • 'adequately sampled' is a vague term.
Slightly relevant	0	0.00%	
Moderately relevant	0	0.00%	
Very relevant	5	23.81%	
Extremely relevant	16	76.19%	

Suggested revision:

- We agree that these criteria are similar and suggest retaining criterion #4.

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The sampling strategy is relevant to address the research question.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 5

Group's responses and comments:

6. The sample is representative of the target population.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• I did not answer #6 since this is not a correct statement. It is impossible to speak about "representatively" without specifying according to what variables. The minimum requirement to be able to answer such a "vague" statement is to add at least "representative for the main relevant covariates". And even then is it not possible to answer statement 6 in de given response scale since it depends of the kind of study, for example: quasi-experiments, in this case the subsamples should be comparable... • 'representative' (there are many different definitions)
Slightly relevant	1	4.76%	
Moderately relevant	1	4.76%	
Very relevant	7	33.33%	
Extremely relevant	12	57.14%	

Suggested revision:

- As suggested by a participant, we added "for the main relevant variables" at the end of the criterion.

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The sample is representative of the target population for the main relevant variables.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 6

Group's responses and comments:

7. The sample size is adequate.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• The size of the sample if always a bit arbitrary, so it is hard to set standards for how big a sample should be. It can be keyed to one or more critical tests or estimates, but the standards chosen are often pretty arbitrary. So I am not sure that is a meaningful standard. • The term is vague. What is 'adequate'?
Slightly relevant	0	0.00%	
Moderately relevant	4	19.05%	
Very relevant	9	42.86%	
Extremely relevant	8	38.10%	

Suggested revision:

- We modified this criterion and added "considering the population under study".

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The sample size is appropriate considering the population under study (such as population size, expected response rate, etc.).**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 7

Group's responses and comments:

8. The sample size is based on pre-study considerations of statistical power.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	4.76%	No comment
Slightly relevant	3	14.29%	
Moderately relevant	8	38.10%	
Very relevant	6	28.57%	
Extremely relevant	3	14.29%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The sample size is based on pre-study considerations of statistical power.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 8

Group's responses and comments:

9. The same methods of data collection are used for all participants.			
Answer choice	Count	Percentage	Comments
Not at all relevant	1	4.76%	• Some of these would require a "subject to appropriate documentation", for example "The same methods of data collection are used for all participants" is not necessary (e.g., mixed-mode studies) but should be documented as one of the features.
Slightly relevant	6	28.57%	
Moderately relevant	5	23.81%	
Very relevant	6	28.57%	
Extremely relevant	3	14.29%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The same methods of data collection are used for all participants.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 9

This question concerns criteria #10 and #12.

Group's responses and comments:

10. Objective or standard criteria are used for the measurement of the parameter of interest.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• Not sure what this means. Is this that people answered standardized questions or were measured in some consistent fashion rather than the data emanating from someone's opinion or judgment?
Slightly relevant	1	4.76%	
Moderately relevant	4	19.05%	
Very relevant	7	33.33%	
Extremely relevant	9	42.86%	

12. The variables are measured using known "gold standard", or using validated methods.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 12. I worry about this, I don't know what "gold standard measures" means. Often it seems if a measure has been used in the past, it is considered a good one. As I note below, what I am interested in is the quality of the evidence that the measures measure what they are intended to measure. If there are data on that topic and they point in the right direction, that is a good thing. • 12. The variables are measured using known "gold standard", or using validated methods. - Asking 2 things at once. Very difficult to answer.
Slightly relevant	3	14.29%	
Moderately relevant	8	38.10%	
Very relevant	8	38.10%	
Extremely relevant	2	9.52%	

Suggested revision:

- We agree that these criteria are not clear and double-barreled. We merged them and focused the criterion on the use of standard instrument.

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **Standard instruments are used for the measurement of the variables.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 10

Group's responses and comments:

11. The choice of variables is based on their relevance and representativeness (content validity).			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• By the way, items above (e.g., 11) are double-barrelled and hence not valid.
Slightly relevant	1	4.76%	
Moderately relevant	2	9.52%	
Very relevant	14	66.67%	
Extremely relevant	4	19.05%	

Suggested revision:

- We removed the terms "relevance and representative".

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The choice of variables is based on their content validity.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 11

Group's responses and comments:

13. The survey instrument has been piloted.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 13, 14 and 15 are also not clear to me. A survey instrument often involves multiple measures. 13. Piloting is a good thing, but it can cover a wide variety of things. Just having a few people answer the questions per se doesn't make the instrument any better. I personally would like most questions to be "cognitively tested" as well as pretested to gather evidence that the questions are interpreted to mean what the investigators think they mean and that answers reflect what the investigators are trying to measure.
Slightly relevant	2	9.52%	
Moderately relevant	4	19.05%	
Very relevant	9	42.86%	
Extremely relevant	6	28.57%	

Suggested revision:

- As suggested by a participant, we replaced "piloted" with "pretested".

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The survey instrument was pretested.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 12

Group's responses and comments:

14. The survey instrument has been tested for reliability.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 13, 14 and 15 are also not clear to me. A survey instrument often involves multiple measures. 14. Testing for reliability is a good idea, but it again is not clear what the standards are. I mainly would just like the results reported. Also, results may be different for different measures in a survey.
Slightly relevant	1	4.76%	
Moderately relevant	10	47.62%	
Very relevant	7	33.33%	
Extremely relevant	3	14.29%	

Suggested revision:

- We replaced the terms "has been tested for reliability" with "reliable".

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The survey instrument is reliable.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 13

Group's responses and comments:

15. The survey instrument has been validated.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 13, 14 and 15 are also not clear to me. A survey instrument often involves multiple measures. 15. "Validated" in particular is not a clear term. Studies of validity are often useful, but validity is not a state of a measure (like beatification). It just means there is some evidence under some circumstances that to some degree some questions measure what they are supposed to measure.
Slightly relevant	1	4.76%	
Moderately relevant	11	52.38%	
Very relevant	7	33.33%	
Extremely relevant	2	9.52%	

Suggested revision:

- We replaced the terms "has been validated" with "valid".

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The survey instrument is valid.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 14

Group's responses and comments:

16. The statistical analysis is appropriate to answer the research question.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	No comment
Slightly relevant	0	0.00%	
Moderately relevant	1	4.76%	
Very relevant	5	23.81%	
Extremely relevant	15	71.43%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The statistical analysis is appropriate to answer the research question.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 15

Group's responses and comments:

17. The sampling bias is adequately addressed in the analysis.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 17. Not sure you can adequately address "sampling bias" in the analysis. If people are left out, either due to the sampling frame or nonresponse, there may be no real way to "adjust for" are estimate what data the missing people would have provided.
Slightly relevant	0	0.00%	
Moderately relevant	1	4.76%	
Very relevant	12	57.14%	
Extremely relevant	8	38.10%	

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The sampling bias is adequately addressed in the analysis.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 16

Group's responses and comments:

18. All important confounding factors/subgroups/differences are identified and accounted for in the analysis.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 18. Accounting for all important potential confounders is a high bar and may not be reasonable. • Also, some of these have quite high requirements, e.g., "All important confounding factors/subgroups/differences are identified and accounted for in the analysis" - it may be unlikely that you are able to account for all confounders in the analysis, but you should at least be aware of potential confounders which may be identified but which cannot be accounted for in the analysis. • #18 is something of a double barrelled question. It is impossible in most cases to account for all confounding factors—but all should be identified regardless of whether or not they can be controlled.
Slightly relevant	1	4.76%	
Moderately relevant	7	33.33%	
Very relevant	8	38.10%	
Extremely relevant	5	23.81%	

Suggested revision:

- We removed "All important" and "subgroups/difference"

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **Confounding factors are identified and accounted for in the analysis.**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 17

This question concerns criteria #19 and #20.
Group's responses and comments:

19. The response rate is adequate (if not, the low response rate is managed appropriately).			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 19-20: High response rates are good. Mainly, the more information the investigators provide about how well the sample corresponds to the target population with respect to issues relevant to the subject of the study, the better. Again, if the folks who respond are not representative of the target population, there is pretty good evidence that it is hard to fix that. Weighting can only take you so far, if you have inadequate samples of key subgroups. • 19. The response rate is adequate (if not, the low response rate is managed appropriately). - Asking 2 things are once. Very difficult to answer. • What is 'adequate'?
Slightly relevant	1	4.76%	
Moderately relevant	6	28.57%	
Very relevant	10	47.62%	
Extremely relevant	4	19.05%	

20. The likelihood of nonresponse bias is minimal.			
Answer choice	Count	Percentage	Comments
Not at all relevant	0	0.00%	• 19-20: High response rates are good. Mainly, the more information the investigators provide about how well the sample corresponds to the target population with respect to issues relevant to the subject of the study, the better. Again, if the folks who respond are not representative of the target population, there is pretty good evidence that it is hard to fix that. Weighting can only take you so far, if you have inadequate samples of key subgroups.
Slightly relevant	0	0.00%	
Moderately relevant	7	33.33%	
Very relevant	11	52.38%	
Extremely relevant	3	14.29%	

Suggested revision:

- Concerning the response rate, some appraisal tools used in systematic reviews have suggested cut-off values ranging from 60% to 75%. We have replaced criteria #19 and #20 with a new one including a cut-off value.

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews: **The response rate is acceptable (60% or above).**

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer
Please enter your comment here:

Question 18 - New criterion suggested

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews:

There is no significant difference in relevant sociodemographic characteristics between the respondents and the non-respondents.

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 19 - New criterion suggested

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews:

Weighting for nonresponse is carried out.

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Question 20 - New criterion suggested

Please rate the level of relevance of the following criterion for appraising the quality of survey research in systematic reviews:

A clear justification for using survey method is provided.

☐ Not at all relevant
☐ Slightly relevant
☐ Moderately relevant
☐ Very relevant
☐ Extremely relevant
☐ I cannot answer

Please enter your comment here:

Thank you very much for completing the questionnaire!

APPENDIX 9. RESULTS OF THE MAPPING OF CRITERIA IN CRITICAL APPRAISAL TOOLS ON RANDOMIZED CONTROLLED TRIALS AND NON-RANDOMIZED STUDIES

Mapping of criteria from critical appraisal tools for randomized controlled trials (n=23)

Item	CATs specific to RCT														CATs for RCT and NRS									TOTAL (n)	FREQUENCY (%)
	Bizzini scale[1]	Chalmers[2]	Delphi List[3]	Detsky scale[4]	Imperiale[5]	Jadad scale[6]	Mastrischt list[7]	Moncrieff[8]	PEDro scale[9]	PQRS[10]	RoB[11]	Sindhu[12]	Van Tulder[13]	Yates[14]	Cho[15]	DIAD[16]	Down[17]	EAI[18]	EPHPP[19]	MacLeahose[20]	Reisch scale[21]	Wells-Parker[22]	Zaza[23]		
SAMPLE																									
Participants in the group receiving the intervention comparable to participants in the control group/baseline balance/baseline equivalence		x	x		x		x	x	x		x	x	x	x		x		x	x		x	x		15	65
Clear selection criteria (inclusion and exclusion)	x	x	x	x	x		x	x	x	x				x	x			x		x	x		x	15	65
Clear description of the characteristics of the participants								x				x		x			x	x		x	x		x	8	35
Description of the numbers of subjects screened, included, and excluded		x		x				x		x										x	x			6	26
Subjects representative of the target population								x								x	x		x				x	5	22
Subjects and control recruited from the same population										x						x	x			x				4	17
Subjects and control recruited from the same period of time										x							x	x		x				4	17
Sampling frame/source clearly described								x										x					x	3	13
Subjects appropriate															x						x			2	9
Control subjects appropriate	x														x									2	9
Participation rate adequate																		x	x					2	9
Restriction to a homogeneous study population	x						x																	2	9

Adequate correction for baseline differences							x	x																2	9
Selection bias reported												x												1	4
Documentation or demonstration of reliability of diagnostic methodology										x														1	4
Description of relevant comorbidities										x														1	4
Assess whether the units of analyses were comparable prior to exposure to the intervention																							x	1	4
SAMPLING/ALLOCATION																									
Allocation sequence random/subjects assigned randomly/method of randomization performed	x	x	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x	x			19	83
Treatment allocation concealed		x	x				x	x	x		x	x	x	x			x	x						11	48
Testing randomization/adequate allocation procedure/valid method		x					x					x	x											4	17
Description of random selection and allocation method															x				x					2	9
Use stratification																					x			1	4
BLINDING																									
Blinding of assessors/blinding outcome assessment	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x	x	x	x	x			19	83
Subjects blinded to the intervention		x	x			x	x	x	x	x	x	x	x		x	x	x	x	x	x	x			17	74
Care providers blinded to the intervention		x	x	x			x		x		x	x	x			x				x	x			11	48
Evaluation of blinding/successful blinding		x					x													x				3	13
Blinding of statistician		x																						1	4
Blinding of authorities (e.g., parents, teachers, case managers)																x								1	4
Blinding of trial personnel											x													1	4
Assessment of the outcome likely to be influenced by knowledge of intervention received											x													1	4
Reasons/discussion for not blinding												x												1	4

INTERVENTION/EXPOSURE																								
Interventions well described (e.g., duration, intensity, number of sessions)	x	x		x			x	x		x				x			x	x		x	x		x	12	52
Compliance/adherence with the intervention assessed, acceptable, presented		x					x	x			x	x	x	x			x		x	x	x			11	48
Cointervention avoided or similar between groups	x				x		x			x	x	x	x			x			x		x			10	43
The intended intervention was administered as designed/deviations from the intended intervention beyond what would be expected in usual practice/intervention implemented successfully										x	x					x			x			x		5	22
Intervention suitable to answer questions under investigation/ intervention appropriate					x															x	x			3	13
Similarity of the therapy across trials/treatments comparable		x			x																			2	9
Exposure variables valid/reliable measures of the intervention under study																		x					x	2	9
Staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients received																	x			x				2	9
Therapist training and level of experience in the treatment(s) under investigation										x				x										2	9
Biological equivalence		x																						1	4
Deviations from intended intervention similar between groups and not likely to have affected the outcome											x													1	4
Subjects received treatment or control as allocated									x															1	4

Good, well-matched alternative intervention group used														x									1	4
Therapist supervision while treatment is being provided										x													1	4
Appropriate consideration of therapist and site effects (either discussed or considered statistically)										x													1	4
Balance of allegiance to types of treatment by practitioners										x													1	4
Consideration of effects of patients preferences and expectations of treatment that may affect the outcome																			x				1	4
Inclusion of variation on important characteristics of the target setting															x								1	4
Intervention tested for its effect within important subgroups of participants, settings, and outcomes															x								1	4
Attempt to measure exposure to the intervention																						x	1	4
Methods of assessing the exposure variables similar for each group																	x						1	4
MEASUREMENT																								
Valid, reliable, and/or standard measures								x		x		x		x		x	x	x	x	x	x	x	12	52
Clear description/justification of outcomes measured				x				x									x	x		x	x		7	30
Report/discuss adverse events		x					x	x		x							x	x		x			7	30
Relevant/appropriate outcomes	x											x				x						x	4	17
Measurement bias accounted for/minimized														x	x								2	9
Objective measurement				x								x											2	9
Measurement equivalence																		x				x	2	9
Outcome measure specified in advance								x		x													2	9
Prospective evaluation																					x		1	4

Appropriate evaluation methods to answer questions																					x			1	4
Independent outcomes												x												1	4
Number of outcomes												x												1	4
Measure of the outcome at a time appropriate for capturing the intervention's effect																x								1	4
Study conducted during a time frame appropriate for extrapolating to current conditions																x								1	4
Observations taken over the same time for all groups																		x						1	4
Form of measurement stated												x												1	4
Measure of any sustainable change between the treatment and control groups														x										1	4
All important clinical information reported																					x			1	4
ANALYSIS																									
Appropriate/adequate statistical analysis	x	x		x				x	x	x		x		x	x		x		x	x	x		x	14	61
Inclusion of all subjects in analyses ('intention-to-treat' analysis)/ participants analyzed in the group they were assigned to	x	x	x				x	x	x	x	x	x	x	x					x	x				13	57
Power calculation reported/ prior estimate of numbers/sample size justification		x		x				x				x		x	x	x	x	x		x	x			11	48
Report both point measures and measures of variability/study provide estimates of the random variability in the data for the main outcomes		x	x				x	x	x						x	x	x	x		x				10	43
Adequate sample size	x						x	x		x		x		x		x					x			8	35
Statistical tests adequately reported		x						x				x			x	x		x			x			7	30

Analyses not planned at the outset of the study clearly indicated (especially for nonsignificant results)/post-hoc analysis		x		x										x		x			x				5	22
Sample size reported												x									x			2	9
Results of between-group statistical comparisons are reported for at least one key outcome									x															1	4
Multiple looks considered		x																						1	4
Analyses that were planned at the outset subject to bias																			x					1	4
Effect sizes and their standard errors accurately estimated															x									1	4
Prior history of disease and/or symptoms collected and included in the analysis																	x							1	4
Appropriate interpretation of statistical results																					x			1	4
Computation errors or contradictions identified																					x			1	4
If no intention-to-treat: potential for a substantial impact (on the estimated effect of intervention) of analyzing participants in the wrong group											x													1	4
The overall significance level reported protected against inflation due to multiple testing												x												1	4
Statistical assumption hold												x												1	4
Other problems with the data analysis																							x	1	4
ATTRITION/FOLLOW-UP																									
The number, reasons and/or characteristics of subjects lost to follow-up described (lost after entry or not participating)						x	x	x		x	x		x	x	x		x	x	x	x	x			13	57

Low lost to follow-up/outcome data available for all, or nearly all, participants		x					x		x		x	x	x			x			x	x			x	10	43
Drop-out accounted for (subject losses or unavailable records after entry into the study)	x	x															x	x						4	17
Follow-up period adequate	x						x	x										x						4	17
Follow-up timing comparable/similar in all groups							x						x							x				3	13
Analyses adjust for different lengths of follow-up																	x	x						2	9
Evidence that results are robust to the presence of missing outcome data/any loss of subjects or their records likely to bias the results of the study											x										x			2	9
Differential attrition between groups																x								1	4
Specific procedures established to minimize loss of subjects from the study																					x			1	4
Assessment of long-term post termination outcome										x														1	4
CONFOUNDERS																									
Confounders accounted for in analysis												x			x		x	x	x	x			x	7	30
Confounders identified, described, discussed												x					x	x	x	x			x	6	26
Confounders accounted for in design															x				x	x				3	13
Valid and reliable confounding variables																				x				1	4
REPORTING																									
Selective reporting/complete reporting of findings											x	x			x									3	13

Acronyms: CAT: critical appraisal tool; NRS: non-randomized study; RCT: randomized controlled trial.

Mapping of criteria from critical appraisal tools for non-randomized studies (n=19)

Item	CATs for RCT and NRS									CATs for NRS										TOTAL (n)	FREQUENCY (%)
	Cho[15]	DIAD[16]	Down[17]	EAI[18]	EPHPP[19]	MacLeahose[20]	Reisch scale[21]	Wells-Parker[22]	Zaza[23]	MEVORECH[24]	MINORS[25]	NOS[26]	QATSO[27]	Q-Coh[28]	QUIPS[29]	RBOANS[30]	ROBINS-I[31]	RTI-Bank[32]	SAQOR[33]		
SAMPLE																					
Clear selection criteria (inclusion and exclusion)	x			x		x	x		x					x	x			x	x	9	47
Subjects/sample representative of the target population		x	x		x							x	x	x	x				x	8	42
Participants in the group receiving the intervention comparable to participants in the control group/baseline balance/baseline equivalence		x		x	x		x	x			x	x								7	37
Clear description of the baseline characteristics of the participants			x	x		x	x		x						x					6	32
Participation/response rate adequate				x	x					x			x		x					5	26
Subjects and control recruited from the same/comparable population		x	x			x						x				x				5	26
Sampling frame/source clearly described and appropriate				x					x	x					x				x	5	26
Subjects appropriate	x						x		x			x								4	21
Control subjects appropriate/adequate comparison group	x										x	x						x		4	21
Subjects and control recruited from the same period of time			x	x		x					x									4	21
Description of the numbers of subjects screened, included, and excluded/subject flow						x	x			x										3	16
Adequate correction for baseline differences/analysis control for baseline differences/statistical differences between cases and controls have been controlled for																		x	x	2	11
Eligibility criteria applied uniformly to all comparison groups														x				x		2	11

Same recruitment strategy across groups																	x		1	5
Number of non-participants is small, or non-participants have something in common or give similar reasons for refusing to participate in the study													x						1	5
Participants selected based on their characteristics observed after the start of the intervention (if yes, were the post-intervention variables that influenced selection likely to be associated with intervention or outcome)																x			1	5
Newly incident cases taken into account				x															1	5
Assess whether the units of analyses were comparable prior to exposure to the intervention								x											1	5
Adequate description of the period and place of recruitment														x					1	5
Eligibility criteria measured using valid and reliable measures																	x		1	5
Same response rate for both groups											x								1	5
SAMPLING/ALLOCATION																				
Random allocation	x	x	x	x	x	x	x						x						8	42
Description of sampling method	x				x					x								x	4	21
Attempt to balance the allocation between groups (stratification, matching, propensity score)							x			x							x		3	16
Treatment allocation concealed			x	x															2	11
Inclusion of consecutive patients											x				x				2	11
Control group is included																		x	1	5
Control group is easily identifiable																		x	1	5
Controls are matched or randomized																		x	1	5
Adjustment techniques used to correct the presence of selection bias																x			1	5
Assessment of sampling bias										x									1	5
Sampling bias addressed in the analysis										x									1	5
Adequate allocation procedure/valid method		x																	1	5
BLINDING																				
Blinding of assessors/blinding outcome assessment/assessment of the outcome likely to be influenced by knowledge of intervention received	x	x	x	x	x	x	x			x	x	x		x		x	x	x	14	74

Subjects blinded to the intervention	x	x	x	x	x	x	x				x			x						9	47
Care providers blinded to the intervention		x				x	x													3	16
Blinding of authorities (e.g., parents, teachers, case managers)		x																		1	5
Evaluation of blinding/successful blinding						x														1	5
INTERVENTION/EXPOSURE																					
Intervention well described (e.g., duration, intensity, number of sessions)/clear definition of exposure			x	x		x	x		x	x				x	x		x	x		10	53
Intervention/exposures assessed using valid and reliable measures				x					x	x				x	x			x		6	32
Intended intervention was administered as designed/intervention implemented successfully		x			x			x									x	x		5	26
Compliance/adherence with the intervention			x		x	x	x										x			5	26
Exposure measurement equivalence/same method of ascertainment of cases and controls (case-control study)				x								x		x	x					4	21
Cointervention/coexposures avoided or similar between groups		x			x		x										x			4	21
Ascertainment of exposure/adequate assessment of exposure												x				x			x	3	16
Staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive		x	x			x														3	16
Isolate impact from concurrent intervention or unintended exposure that might bias results																		x		1	5
Intervention suitable to answer questions under investigation							x													1	5
Intervention defined at the start (from sources that could not have been affected by subsequent outcomes)																	x			1	5
Classification of intervention status affected by knowledge of the outcome or risk of the outcome																	x			1	5
Deviation from intended intervention balanced between groups and unlikely to affect outcome																	x			1	5
Exposure conducted at a time prior to the occurrence of disease or symptoms				x																1	5

Inclusion of variation on important characteristics of the target setting		x																	1	5
Intervention tested for its effect within important subgroups of participants, settings, and outcomes		x																	1	5
Consideration of effects of patient preferences and expectations of treatment that may affect the outcome						x													1	5
Intervention suitable to answer questions under investigation/intervention appropriate						x													1	5
Attempt to measure exposure to the intervention									x										1	5
MEASUREMENT																				
Valid, reliable, and/or standard outcome measures		x	x	x	x	x	x	x	x	x				x	x			x		12	63
Clear description/definition of outcomes measured			x	x		x	x		x	x				x	x					8	42
Measurement equivalence between groups (way, time, context)				x				x						x	x		x			5	26
Report/discuss adverse events/harms			x	x		x												x		4	21
Relevant outcome/appropriate		x						x			x									3	16
Prospective evaluation/data collection							x				x					x				3	16
Objective measurement													x							1	5
Measurement bias accounted for	x																			1	5
Appropriate evaluation methods to answer questions							x													1	5
Outcome measure specified in advance																		x		1	5
Outcome not present at the start of the study												x								1	5
Adequate measure of outcomes																			x	1	5
Sources (e.g., registries, database, self-reported)										x										1	5
Systematic errors in the measurement of the outcome related to intervention received																	x			1	5
Measure of the outcome at a time appropriate for capturing the intervention's effect		x																		1	5
Study conducted during a time frame appropriate for extrapolating to current conditions		x																		1	5
Observations taken over the same time for all groups				x																1	5
Continuous variables are reported or appropriate cut point are used														x						1	5
All important clinical information reported							x													1	5

ANALYSIS																					
Appropriate/adequate statistical analysis	x		x		x	x	x		x	x	x			x	x			x		11	58
Power calculation reported/sample size justification	x	x	x	x		x	x			x	x									8	42
Report both point measures and measures of variability (data clearly and accurately presented with CI)	x		x	x		x				x									x	6	32
Adequate/appropriate sample size		x					x											x	x	4	21
Statistical tests adequately/sufficiently reported	x			x			x							x						4	21
Analyses not planned at the outset of the study clearly indicated (especially for nonsignificant results)/post-hoc analysis			x			x														2	11
Inclusion of all subjects in analyses (‘intention-to-treat’ analysis)/participants analyzed in the group they were assigned to					x	x														2	11
Appropriate analysis used to estimate the effect of starting and adhering to the intervention	x															x				2	11
Prior history of disease and/or symptoms collected and included in the analysis				x																1	5
Sample size reported							x													1	5
Appropriate interpretation of statistical results							x													1	5
Computation errors or contradictions identified							x													1	5
Analyses that were planned at the outset subject to bias						x														1	5
Appropriate methods of imputation are used for missing data														x						1	5
Appropriate statistical methods used to assess the main harm or adverse event outcome																		x		1	5
Exclusion rate from the analysis										x										1	
ATTRITION/FOLLOW-UP																					
Complete outcome data/lost to follow-up unlikely to introduce bias/low lost to follow-up/attrition unlikely to bias conclusions		x			x	x			x	x	x	x			x	x	x	x		11	58
The number, reasons and/or characteristics of subjects lost to follow-up described (lost after entry or not participating)	x		x	x	x	x	x								x				x	8	42
Differential attrition between groups/dropout rates/reasons similar in all group		x								x				x	x	x	x	x		7	37

Follow-up period adequate/appropriate/sufficient/long enough				x							x	x						x		4	21
Follow-up timing/length comparable/similar in all groups						x								x				x		3	16
Missing data/drop-out accounted for			x	x											x					3	16
Analyses adjusted for different lengths of follow-up/same time period between the intervention and outcome same for cases and controls			x	x																2	11
Lost to follow-up not associated with key characteristics															x					1	5
Explanation of missing data is given																			x	1	5
Impact of high lost to follow-up assessed (e.g., sensitivity analysis)																		x		1	5
Follow-up and start of intervention coincide for most participants																	x			1	5
Participants excluded due to missing data on intervention status or other variables needed for the analysis																	x			1	5
Evidence that results were robust to the presence of missing data																	x			1	5
Effect sizes and their standard errors accurately estimated		x																		1	5
Evidence that results are robust to the presence of missing outcome data/any loss of subjects or their records likely to bias the results of the study							x													1	5
Specific procedures established to minimize loss of subjects from the study							x													1	5
Description of attempts to collect information on participants who dropped out															x					1	5
CONFOUNDERS																					
Confounders accounted for/taken into account in analysis/use appropriate analysis method that controlled for all important confounding	x		x	x	x	x			x				x	x	x	x	x	x	x	13	68
Confounders accounted for/taken into account in the design	x				x	x								x	x	x		x	x	8	42

Confounders identified, described, assessed, discussed/clear definition of confounders			x	x	x	x			x	x					x					7	37
Confounding/effect modifying variables assessed using valid and reliable measures across all study participants/adequately confirmed						x				x					x	x	x	x		6	32
All important potential confounders are measured/taken into account														x	x					2	11
Potential for confounding of the effect of intervention in the study																	x			1	5
Time-varying confounder: the analysis based on splitting participants' follow-up time according to intervention received; intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome																	x			1	5
Control for post-intervention variable that could have been affected by the intervention																	x			1	5
Method and setting of confounding measurement are the same for all study participants														x						1	5
Appropriate methods are used if imputation is used for missing confounder data														x						1	5
REPORTING																					
Selective reporting/complete reporting of data	x									x					x	x	x	x		6	32

Acronyms: CAT: critical appraisal tool; NRS: non-randomized study; RCT: randomized controlled trial.

References

1. Bizzini M, Childs JD, Piva SR, Delitto A. Systematic review of the quality of randomized controlled trials for patellofemoral pain syndrome. *Journal of Orthopaedic & Sports Physical Therapy* 2003;33:4-20.
2. Chalmers TC, Smith H, Blackburn B, et al. A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials* 1981;2:31-49.
3. Verhagen AP, de Vet HC, de Bie RA, et al. 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* 1998;51:1235-41.
4. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology* 1992;45:255-65.
5. Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis?: A meta-analysis of the randomized trials. *Annals of Internal Medicine* 1990;113:299-307.
6. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials* 1996;17:1-12.
7. de Vet HC, de Bie RA, van der Heijden GJ, Verhagen AP, Sijpkes P, Knipschild PG. Systematic reviews on the basis of methodological criteria. *Physiotherapy* 1997;83:284-9.
8. Moncrieff J, Churchill R, Drummond DC, McGuire H. Development of a quality assessment instrument for trials of treatments for depression and neurosis. *International Journal of Methods in Psychiatric Research* 2001;10:126-33.
9. Sherrington C, Herbert R, Maher C, Moseley A. PEDro. A database of randomized trials and systematic reviews in physiotherapy. *Manual Therapy* 2000;5:223-6.
10. Kocsis JH, Gerber AJ, Milrod B, et al. A new scale for assessing the quality of randomized clinical trials of psychotherapy. *Comprehensive Psychiatry* 2010;51:319-24.
11. Higgins JPT, Sterne JAC, Savović J, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, eds. *Cochrane Methods. Cochrane Database of Systematic Reviews*, Issue 10 (Suppl 1); 2016.
12. Sindhu F, Carpenter L, Seers K. Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. *Journal of Advanced Nursing* 1997;25:1262-8.
13. Van Tulder M, Furlan A, Bombardier C, Bouter L, Editorial Board of the Cochrane Collaboration Back Review Group. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine* 2003;28:1290-9.
14. Yates SL, Morley S, Eccleston C, Williams ACdC. A scale for rating the quality of psychological trials for pain. *Pain* 2005;117:314-25.
15. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *Journal of the American Medical Association* 1994;272:101-4.
16. Valentine JC, Cooper H. 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* 2008;13:130-149.
17. Downs SH, Black N. 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* 1998;52:377-84.

18. Genaidy A, Lemasters G, Lockey J, et al. An epidemiological appraisal instrument – A tool for evaluation of epidemiological studies. *Ergonomics* 2007;50:920-60.
19. Thomas B, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing* 2004;1:176-84.
20. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* 2000;4:1-154.
21. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. *Pediatrics* 1989;84:815-27.
22. Wells-Parker E, Bangert-Drowns R. Meta-analysis of research on DUI remedial interventions. *Alcohol, Drugs & Driving* 1990;6: 147-160.
23. Zaza S, Wright-De Agüero LK, Briss PA, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *American Journal of Preventive Medicine* 2000;18:44-74.
24. Shamliyan TA, Kane RL, Ansari MT, et al. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: Pilot study of new checklists. *Journal of Clinical Epidemiology* 2011;64:637-57.
25. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): Development and validation of a new instrument. *ANZ Journal of Surgery* 2003;73:712-6.
26. Wells G, Shea B, O'connell D, et al. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.* Ottawa 2000. Available from: http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf accessed April 16 2016.
27. Wong WC, Cheung CS, Hart GJ. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. *Emerging Themes in Epidemiology* 2008;5:1.
28. Jarde A, Losilla J-M, Vives J, Rodrigo MF. Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology* 2013;13:138-46.
29. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Annals of Internal Medicine* 2013;158:280-6.
30. Kim SY, Park JE, Lee YJ, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology* 2013;66:408-14.
31. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal* 2016;355:i4919.
32. Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology* 2012;65:163-78.
33. Ross L, Grigoriadis S, Mamisashvili L, et al. Quality assessment of observational studies in psychiatry: An example from perinatal psychiatric research. *International Journal of Methods in Psychiatric Research* 2011;20:224-34.

APPENDIX 10. MIXED METHODS APPRAISAL TOOL (MMAT), VERSION 2018

MIXED METHODS APPRAISAL TOOL (MMAT) VERSION 2018

User guide

Prepared by
Quan Nha HONG[a], Pierre PLUYE[a,], Sergi FÀBREGUES[b], Gillian BARTLETT[a], Felicity BOARDMAN[c], Margaret CARGO[d], Pierre DAGENAIS[e], Marie-Pierre GAGNON[f], Frances GRIFFITHS[c], Belinda NICOLAU[a], Alicia O'CATHAIN[g], Marie-Claude ROUSSEAU[h], & Isabelle VEDEL[a]

[a]McGill University, Montréal, Canada; [b]Universitat Oberta de Catalunya, Barcelona, Spain; [c]University of Warwick, Coventry, England; [d]University of Canberra, Canberra, Australia; [e]Université de Sherbrooke, Sherbrooke, Canada; [f]Université Laval, Québec, Canada; [g]University of Sheffield, Sheffield, England; [h]Institut Armand-Frappier Research Centre, Laval, Canada

McGill

Department of Family Medicine | Département de médecine de famille

Academic excellence and innovation in care, teaching and research

Innovation et excellence académique dans les soins, l'enseignement et la recherche

What is the MMAT?

The MMAT is a critical appraisal tool that is designed for the appraisal stage of systematic mixed studies reviews, i.e., reviews that include qualitative, quantitative, and mixed methods studies. It permits to appraise the methodological quality of five categories to studies: qualitative studies, randomized controlled trials, non-randomized studies, quantitative descriptive studies, and mixed methods studies.

How was the MMAT developed?

The MMAT was developed in 2006 (Pluye et al., 2009a) and was revised in 2011 (Pace et al., 2012). The present version 2018 was developed on the basis of findings from a literature review of critical appraisal tools, interviews with MMAT users, and an e-Delphi study with international experts (Hong, 2018). The MMAT developers are continuously seeking for improvement and testing of this tool. Users' feedback is always appreciated.

What the MMAT can be used for?

The MMAT can be used to appraise the quality of empirical studies, i.e., primary research based on experiment, observation or simulation (Abbott, 1998; Porta et al., 2014). It cannot be used for non-empirical papers such as review and theoretical papers. Also, the MMAT allows the appraisal of most common types of study methodologies and designs. However, some specific designs such as economic and diagnostic accuracy studies cannot be assessed with the MMAT. Other critical appraisal tools might be relevant for these designs.

What are the requirements?

Because critical appraisal is about judgment making, it is advised to have at least two reviewers independently involved in the appraisal process. Also, using the MMAT requires experience or training in these domains. For instance, MMAT users may be helped by a colleague with specific expertise when needed.

How to use the MMAT?

This document comprises two parts: checklist (Part I) and explanation of the criteria (Part II).

1. Respond to the two screening questions. Responding 'No' or 'Can't tell' to one or both questions might indicate that the paper is not an empirical study, and thus cannot be appraised using the MMAT. MMAT users might decide not to use these questions, especially if the selection criteria of their review are limited to empirical studies.
2. For each included study, choose the appropriate category of studies to appraise. Look at the description of the methods used in the included studies. If needed, use the algorithm at the end of this document.
3. Rate the criteria of the chosen category. For example, if the paper is a qualitative study, only rate the five criteria in the qualitative category. The 'Can't tell' response category means that the paper do not report appropriate information to answer 'Yes' or 'No', or that report unclear information related to the criterion. Rating 'Can't tell' could lead to look for companion papers, or contact authors to ask more information or clarification when needed. In Part II of this document, indicators are added for some criteria. The list is not exhaustive and not all indicators are necessary. You should agree among your team which ones are important to consider for your field and apply them uniformly across all included studies from the same category.

How to score?

In the literature, it is discouraged to calculate an overall score from the ratings of each criterion. Instead, it is advised to provide a more detailed presentation of the ratings of each criterion to better inform the quality of the included studies. This may lead to perform a sensitivity analysis (i.e., to consider the quality of studies by contrasting their results). Excluding studies with low methodological quality is usually discouraged.

How to cite this document?

Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, Gagnon M-P, Griffiths F, Nicolau B, O'Cathain A, Rousseau M-C, Vedel I. Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

For dissemination, application, and feedback: Please contact mixed.methods.appraisal.tool@gmail.com

For more information: http://mixedmethodsappraisaltoolpublic.pbworks.com/

Part I: Mixed Methods Appraisal Tool (MMAT), version 2018

Category of study designs	Methodological quality criteria	Responses			
		Yes	No	Can't tell	Comments
Screening questions (for all types)	S1. Are there clear research questions?				
	S2. Do the collected data allow to address the research questions?				
	Further appraisal may not be feasible or appropriate when the answer is 'No' or 'Can't tell' to one or both screening questions.				
1. Qualitative	1.1. Is the qualitative approach appropriate to answer the research question?				
	1.2. Are the qualitative data collection methods adequate to address the research question?				
	1.3. Are the findings adequately derived from the data?				
	1.4. Is the interpretation of results sufficiently substantiated by data?				
	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?				
2. Quantitative randomized controlled trials	2.1. Is randomization appropriately performed?				
	2.2. Are the groups comparable at baseline?				
	2.3. Are there complete outcome data?				
	2.4. Are outcome assessors blinded to the intervention provided?				
	2.5 Did the participants adhere to the assigned intervention?				
3. Quantitative non-randomized	3.1. Are the participants representative of the target population?				
	3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?				
	3.3. Are there complete outcome data?				
	3.4. Are the confounders accounted for in the design and analysis?				
	3.5. During the study period, is the intervention administered (or exposure occurred) as intended?				

4. Quantitative descriptive	4.1. Is the sampling strategy relevant to address the research question?				
	4.2. Is the sample representative of the target population?				
	4.3. Are the measurements appropriate?				
	4.4. Is the risk of nonresponse bias low?				
	4.5. Is the statistical analysis appropriate to answer the research question?				
5. Mixed methods	5.1. Is there an adequate rationale for using a mixed methods design to address the research question?				
	5.2. Are the different components of the study effectively integrated to answer the research question?				
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?				
	5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?				
	5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?				

Part II: Explanations

1. Qualitative studies	Methodological quality criteria
"Qualitative research is an approach for exploring and understanding the meaning individuals or groups ascribe to a social or human problem" (Creswell, 2013b, p. 3). Common qualitative research approaches include (this list if not exhaustive): **Ethnography** The aim of the study is to describe and interpret the shared cultural behaviour of a group of individuals. **Phenomenology** The study focuses on the subjective experiences and interpretations of a phenomenon encountered by individuals. **Narrative research** The study analyzes life experiences of an individual or a group. **Grounded theory** Generation of theory from data in the process of conducting research (data collection occurs first).	1.1. Is the qualitative approach appropriate to answer the research question? Explanations The qualitative approach used in a study (see non-exhaustive list on the left side of this table) should be appropriate for the research question and problem. For example, the use of a grounded theory approach should address the development of a theory and ethnography should study human cultures and societies. This criterion was considered important to add in the MMAT since there is only one category of criteria for qualitative studies (compared to three for quantitative studies).
	1.2. Are the qualitative data collection methods adequate to address the research question? Explanations This criterion is related to data collection method, including data sources (e.g., archives, documents), used to address the research question. To judge this criterion, consider whether the method of data collection (e.g., in depth interviews and/or group interviews, and/or observations) and the form of the data (e.g., tape recording, video material, diary, photo, and/or field notes) are adequate. Also, clear justifications are needed when data collection methods are modified during the study.
	1.3. Are the findings adequately derived from the data? Explanations This criterion is related to the data analysis used. Several data analysis methods have been developed and their use depends on the research question and qualitative approach. For example, open, axial and selective coding is often associated with grounded theory, and within- and cross-case analysis is often seen in case study.

Case study In-depth exploration and/or explanation of issues intrinsic to a particular case. A case can be anything from a decision-making process, to a person, an organization, or a country. **Qualitative description** There is no specific methodology, but a qualitative data collection and analysis, e.g., in-depth interviews or focus groups, and hybrid thematic analysis (inductive and deductive). Key references: Creswell (2013a); Sandelowski (2010); Schwandt (2015)	1.4. Is the interpretation of results sufficiently substantiated by data? Explanations The interpretation of results should be supported by the data collected. For example, the quotes provided to justify the themes should be adequate.
	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation? Explanations There should be clear links between data sources, collection, analysis and interpretation.

2. Quantitative randomized controlled trials	Methodological quality criteria
Randomized controlled clinical trial: A clinical study in which individual participants are allocated to intervention or control groups by randomization (intervention assigned by researchers). Key references: Higgins and Green (2008); Higgins et al. (2016); Oxford Centre for Evidence-based Medicine (2016); Porta et al. (2014)	2.1. Is randomization appropriately performed? Explanations In a randomized controlled trial, the allocation of a participant (or a data collection unit, e.g., a school) into the intervention or control group is based solely on chance. Researchers should describe how the randomization schedule was generated. A simple statement such as 'we randomly allocated' or 'using a randomized design' is insufficient to judge if randomization was appropriately performed. Also, assignment that is predictable such as using odd and even record numbers or dates is not appropriate. At minimum, a simple allocation (or unrestricted allocation) should be performed by following a predetermined plan/sequence. It is usually achieved by referring to a published list of random numbers, or to a list of random assignments generated by a computer. Also, restricted allocation can be performed such as blocked randomization (to ensure particular allocation ratios to the intervention groups), stratified randomization (randomization performed separately within strata), or minimization (to make small groups closely similar with respect to several characteristics). Another important characteristic to judge if randomization was appropriately performed is allocation concealment that protects assignment sequence until allocation. Researchers and participants should be unaware of the assignment sequence up to the point of allocation. Several strategies can be used to ensure allocation concealment such relying on a central randomization by a third party, or the use of sequentially numbered, opaque, sealed envelopes (Higgins et al., 2016).
	2.2. Are the groups comparable at baseline? Explanations Baseline imbalance between groups suggests that there are problems with the randomization. Indicators from baseline imbalance include: "(1) unusually large differences between intervention group sizes; (2) a substantial excess in statistically significant differences in baseline characteristics than would be expected by chance alone; (3) imbalance in key prognostic factors (or baseline measures of outcome variables) that are unlikely to be due to chance; (4) excessive similarity in baseline characteristics that is not compatible with chance; (5) surprising absence of one or more key characteristics that would be expected to be reported" (Higgins et al., 2016, p. 10).

	2.3. Are there complete outcome data? Explanations Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field and apply this uniformly across all the included studies. For instance, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003), and 30% for a follow-up of more than one year (Viswanathan and Berkman, 2012).
	2.4. Are outcome assessors blinded to the intervention provided? Explanations Outcome assessors should be unaware of who is receiving which interventions. The assessors can be the participants if using participant reported outcome (e.g., pain), the intervention provider (e.g., clinical exam), or other persons not involved in the intervention (Higgins et al., 2016).
	2.5 Did the participants adhere to the assigned intervention? Explanations To judge this criterion, consider the proportion of participants who continued with their assigned intervention throughout follow-up. “Lack of adherence includes imperfect compliance, cessation of intervention, crossovers to the comparator intervention and switches to another active intervention.” (Higgins et al., 2016, p. 25).

<table>
<tr><th>3. Quantitative non-randomized studies</th><th>Methodological quality criteria</th></tr>
<tr><td rowspan="3">Non-randomized studies are defined as any quantitative studies estimating the effectiveness of an intervention or studying other exposures that do not use randomization to allocate units to comparison groups (Higgins and Green, 2008).

Common designs include (this list if not exhaustive):

Non-randomized controlled trials
The intervention is assigned by researchers, but there is no randomization, e.g., a pseudo-randomization. A non-random method of allocation is not reliable in producing alone similar groups.

Cohort study
Subsets of a defined population are assessed as exposed, not exposed, or exposed at different degrees to factors of interest. Participants are followed over time to determine if an outcome occurs (prospective longitudinal).

Case-control study
Cases, e.g., patients, associated with a certain outcome are selected, alongside a corresponding group of controls. Data is collected on whether cases and controls were exposed to the factor under study (retrospective).</td><td>3.1. Are the participants representative of the target population?

Explanations
Indicators of representativeness include: clear description of the target population and of the sample (inclusion and exclusion criteria), reasons why certain eligible individuals chose not to participate, and any attempts to achieve a sample of participants that represents the target population.</td></tr>
<tr><td>3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?

Explanations
Indicators of appropriate measurements include: the variables are clearly defined and accurately measured; the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure; validated and reliability tested measures of the intervention/exposure and outcome of interest are used, or variables are measured using ‘gold standard’.</td></tr>
<tr><td>3.3. Are there complete outcome data?

Explanations
Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field (and based on the targeted journal) and apply this uniformly across all the included studies. For example, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003), and 30% for follow-up of more than one year (Viswanathan and Berkman, 2012).</td></tr>
</table>

<table>
<tr><td rowspan="2">Cross-sectional analytic study
At one particular time, the relationship between health-related characteristics (outcome) and other factors (intervention/exposure) is examined. E.g., the frequency of outcomes is compared in different population subgroups according to the presence/absence (or level) of the intervention/exposure.

Key references for non-randomized studies: Higgins and Green (2008); Porta et al. (2014); Sterne et al. (2016); Wells et al. (2000)</td><td>3.4. Are the confounders accounted for in the design and analysis?

Explanations
Confounders are factors that predict both the outcome of interest and the intervention received/exposure at baseline. They can distort the interpretation of findings and need to be considered in the design and analysis of a non-randomized study. Confounding bias is low if there is no confounding expected, or appropriate methods to control for confounders are used (such as stratification, regression, matching, standardization, and inverse probability weighting).</td></tr>
<tr><td>3.5 During the study period, is the intervention administered (or exposure occurred) as intended?

Explanations
For intervention studies, consider whether the participants were treated in a way that is consistent with the planned intervention. Since the intervention is assigned by researchers, consider whether there was a presence of contamination (e.g., the control group may be indirectly exposed to the intervention) or whether unplanned co-interventions were present in one group (Sterne et al., 2016).

For observational studies, consider whether changes occurred in the exposure status among the participants. If yes, check if these changes are likely to influence the outcome of interest, were adjusted for, or whether unplanned co-exposures were present in one group (Morgan et al., 2017).</td></tr>
</table>

<table>
<tr><th>4. Quantitative descriptive studies</th><th>Methodological quality criteria</th></tr>
<tr><td rowspan="3">Quantitative descriptive studies are “concerned with and designed only to describe the existing distribution of variables without much regard to causal relationships or other hypotheses” (Porta et al., 2014, p. 72). They are used to monitoring the population, planning, and generating hypothesis (Grimes and Schulz, 2002).

Common designs include the following single-group studies (this list if not exhaustive):

Incidence or prevalence study without comparison group
In a defined population at one particular time, what is happening in a population, e.g., frequencies of factors (importance of problems), is described (portrayed).

Survey
“Research method by which information is gathered by asking people questions on a specific topic and the data collection procedure is standardized and well defined.” (Bennett et al., 2011, p. 3).

Case series
A collection of individuals with similar characteristics are used to describe an outcome.

Case report
An individual or a group with a unique/unusual outcome is described in detail.</td><td>4.1. Is the sampling strategy relevant to address the research question?

Explanations
Sampling strategy refers to the way the sample was selected. There are two main categories of sampling strategies: probability sampling (involve random selection) and non-probability sampling. Depending on the research question, probability sampling might be preferable. Non-probability sampling does not provide equal chance of being selected. To judge this criterion, consider whether the source of sample is relevant to the target population; a clear justification of the sample frame used is provided; or the sampling procedure is adequate.</td></tr>
<tr><td>4.2. Is the sample representative of the target population?

Explanations
There should be a match between respondents and the target population. Indicators of representativeness include: clear description of the target population and of the sample (such as respective sizes and inclusion and exclusion criteria), reasons why certain eligible individuals chose not to participate, and any attempts to achieve a sample of participants that represents the target population.</td></tr>
<tr><td>4.3. Are the measurements appropriate?

Explanations
Indicators of appropriate measurements include: the variables are clearly defined and accurately measured, the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure; validated and reliability tested measures of the outcome of interest are used, variables are measured using ‘gold standard’, or questionnaires are pre-tested prior to data collection.</td></tr>
</table>

Key references: Critical Appraisal Skills Programme (2017); Draugalis et al. (2008)	4.4. Is the risk of nonresponse bias low? Explanations Nonresponse bias consists of "an error of nonobservation reflecting an unsuccessful attempt to obtain the desired information from an eligible unit." (Federal Committee on Statistical Methodology, 2001, p. 6). To judge this criterion, consider whether the respondents and non-respondents are different on the variable of interest. This information might not always be reported in a paper. Some indicators of low nonresponse bias can be considered such as a low nonresponse rate, reasons for nonresponse (e.g., noncontacts vs. refusals), and statistical compensation for nonresponse (e.g., imputation). The nonresponse bias is might not be pertinent for case series and case report. This criterion could be adapted. For instance, complete data on the cases might be important to consider in these designs.
	4.5. Is the statistical analysis appropriate to answer the research question? Explanations The statistical analyses used should be clearly stated and justified in order to judge if they are appropriate for the design and research question, and if any problems with data analysis limited the interpretation of the results.

5. Mixed methods studies	Methodological quality criteria
Mixed methods (MM) research involves combining qualitative (QUAL) and quantitative (QUAN) methods. In this tool, to be considered MM, studies have to meet the following criteria (Creswell and Plano Clark, 2017): (a) at least one QUAL method and one QUAN method are combined; (b) each method is used rigorously in accordance to the generally accepted criteria in the area (or tradition) of research invoked; and (c) the combination of the methods is carried out at the minimum through a MM design (defined *a priori*, or emerging) and the integration of the QUAL and QUAN phases, results, and data. Common designs include (this list if not exhaustive): **Convergent design** The QUAL and QUAN components are usually (but not necessarily) concomitant. The purpose is to examine the same phenomenon by interpreting QUAL and QUAN results (bringing data analysis together at the interpretation stage), or by integrating QUAL and QUAN datasets (e.g., data on same cases), or by transforming data (e.g., quantization of qualitative data). **Sequential explanatory design** Results of the phase 1 - QUAN component inform the phase 2 - QUAL component. The	5.1. Is there an adequate rationale for using a mixed methods design to address the research question? Explanations The reasons for conducting a mixed methods study should be clearly explained. Several reasons can be invoked such as to enhance or build upon qualitative findings with quantitative results and vice versa; to provide a comprehensive and complete understanding of a phenomenon or to develop and test instruments (Bryman, 2006).
	5.2. Are the different components of the study effectively integrated to answer the research question? Explanations Integration is a core component of mixed methods research and is defined as the "explicit interrelating of the quantitative and qualitative component in a mixed methods study" (Plano Clark and Ivankova, 2015, p. 40). Look for information on how qualitative and quantitative phases, results, and data were integrated (Pluye et al., 2018). For instance, how data gathered by both research methods was brought together to form a complete picture (e.g., joint displays) and when integration occurred (e.g., during the data collection-analysis or/and during the interpretation of qualitative and quantitative results).
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted? Explanations This criterion is related to meta-inference, which is defined as the overall interpretations derived from integrating qualitative and quantitative findings (Teddlie and Tashakkori, 2009). Meta-inference occurs during the interpretation of the findings from the integration of the qualitative and quantitative components, and shows the added value of conducting a mixed methods study rather than having two separate studies.

<table>
<tr>
<td rowspan="2">purpose is to explain QUAN results using QUAL findings. E.g., the QUAN results guide the selection of QUAL data sources and data collection, and the QUAL findings contribute to the interpretation of QUAN results.

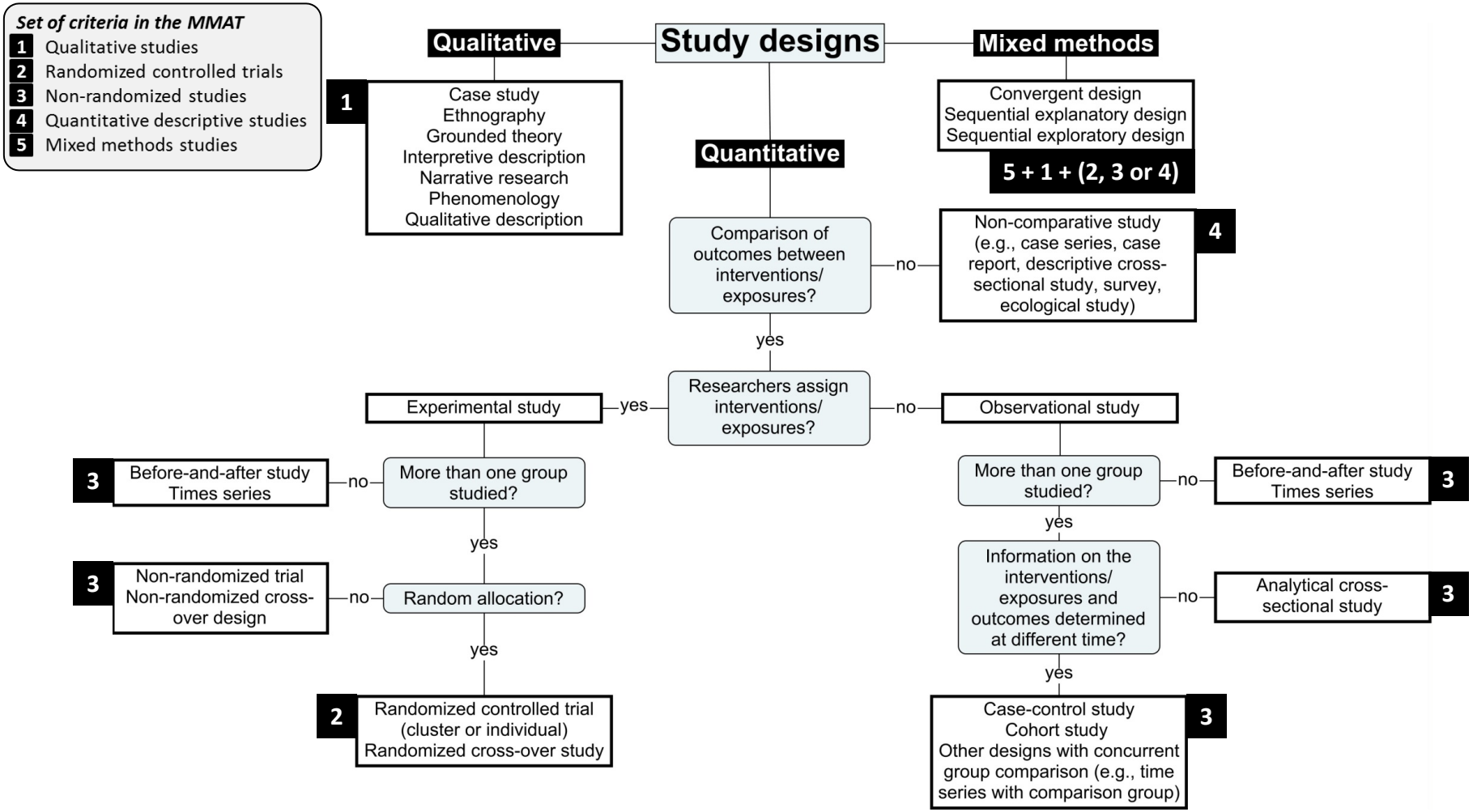
Sequential exploratory design
Results of the phase 1 - QUAL component inform the phase 2 - QUAN component. The purpose is to explore, develop and test an instrument (or taxonomy), or a conceptual framework (or theoretical model). E.g., the QUAL findings inform the QUAN data collection, and the QUAN results allow a statistical generalization of the QUAL findings.

Key references: Creswell et al. (2011); Creswell and Plano Clark, (2017); O'Cathain (2010)</td>
<td>5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?

Explanations
When integrating the findings from the qualitative and quantitative components, divergences and inconsistencies (also called conflicts, contradictions, discordances, discrepancies, and dissonances) can be found. It is not sufficient to only report the divergences; they need to be explained. Different strategies to address the divergences have been suggested such as reconciliation, initiation, bracketing and exclusion (Pluye et al., 2009b). Rate this criterion ‘Yes’ if there is no divergence.</td>
</tr>
<tr>
<td>5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

Explanations
The quality of the qualitative and quantitative components should be individually appraised to ensure that no important threats to trustworthiness are present. To appraise 5.5, use criteria for the qualitative component (1.1 to 1.5), and the appropriate criteria for the quantitative component (2.1 to 2.5, or 3.1 to 3.5, or 4.1 to 4.5). The quality of both components should be high for the mixed methods study to be considered of good quality. The premise is that the overall quality of a mixed methods study cannot exceed the quality of its weakest component. For example, if the quantitative component is rated high quality and the qualitative component is rated low quality, the overall rating for this criterion will be of low quality.</td>
</tr>
</table>

Algorithm for selecting the study categories to rate in the MMAT*



*Adapted from National Institute for Health Care Excellence. (2012). *Methods for the development of nice public health guidance*. London: National Institute for Health and Care Excellence; and Scottish Intercollegiate Guidelines Network. (2017). *Algorithm for classifying study design for questions of effectiveness*. Retrieved December 1, 2017, from http://www.sign.ac.uk/assets/study_design.pdf.

References

Abbott, A. (1998). The causal devolution. *Sociological Methods & Research, 27*(2), 148-181.

Bennett, C., Khangura, S., Brehaut, J. C., Graham, I. D., Moher, D., Potter, B. K., et al. (2011). Reporting guidelines for survey research: An analysis of published guidance and reporting practices. *PLoS Medicine, 8*(8), e1001069.

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research, 6*(1), 97-113.

Creswell, J. W. (2013a). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Creswell, J. W. (2013b). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE Publications.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., Smith, K. C. (2011). *Best practices for mixed methods research in the health sciences*. Bethesda, MD: Office of Behavioral and Social Sciences Research, National Institutes of Health. http://obssr.od.nih.gov/mixed_methods_research.

Creswell, J. W., & Plano Clark, V. (2017). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Critical Appraisal Skills Programme. (2017). CASP checklists. Retrieved December 1, 2017, from http://www.casp-uk.net/casp-tools-checklists.

de Vet, H. C., de Bie, R. A., van der Heijden, G. J., Verhagen, A. P., Sijpkes, P., & Knipschild, P. G. (1997). Systematic reviews on the basis of methodological criteria. *Physiotherapy, 83*(6), 284-289.

Draugalis, J. R., Coons, S. J., & Plaza, C. M. (2008). Best practices for survey research reports: A synopsis for authors and reviewers. *American Journal of Pharmaceutical Education, 72*(1), Article 11.

Federal Committee on Statistical Methodology. (2001). *Measuring and reporting sources of error in surveys*. Washington DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Grimes, D. A., & Schulz, K. F. (2002). Descriptive studies: What they can and cannot do. *The Lancet, 359*(9301), 145-149.

Higgins, J. P., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley Online Library.

Higgins, J. P. T., Sterne, J. A. C., Savović, J., Page, M. J., Hróbjartsson, A., Boutron, I., et al. (2016). A revised tool for assessing risk of bias in randomized trials. In Chandler, J., McKenzie, J., Boutron, I. & Welch, V. (Eds.), *Cochrane Methods. Cochrane Database of Systematic Reviews*, Issue 10 (Suppl 1).

Hong, Q. N. (2018). *Revision of the Mixed Methods Appraisal Tool (MMAT): A mixed methods study* (Doctoral dissertation). Department of Family Medicine, McGill University, Montréal.

MacLehose, R. R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., & Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment, 4*(34), 1-154.

Morgan, R., Sterne, J., Higgins, J., Thayer, K., Schunemann, H., Rooney, A., et al. (2017). *A new instrument to assess Risk of Bias in Non-randomised Studies of Exposures (ROBINS-E): Application to studies of environmental exposure.* Abstracts of the Global Evidence Summit, Cape Town, South Africa. Cochrane Database of Systematic Reviews 2017, Issue 9 (Suppl 1). https://doi.org/10.1002/14651858.CD201702.

O'Cathain, A. (2010). Assessing the quality of mixed methods research: Towards a comprehensive framework. In Tashakkori, A. & Teddlie, C. (Eds.), *Handbook of Mixed methods in social and behavioral research* (pp. 531-555). Thousand Oaks, CA: SAGE Publications.

Oxford Centre for Evidence-based Medicine. (2016). *Levels of evidence*. Retrieved February 19, 2018, from https://www.cebm.net/2016/05/ocebm-levels-of-evidence/.

Pace, R., Pluye, P., Bartlett, G., Macaulay, A. C., Salsberg, J., Jagosh, J., et al. (2012). Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. *International Journal of Nursing Studies, 49*(1), 47-53.

Plano Clark, V. L., & Ivankova, N. V. (2015). *Mixed methods research: A guide to the field*. Thousand Oaks, CA: SAGE Publications.

Pluye, P., Gagnon, M. P., Griffiths, F., Johnson-Lafleur, J. (2009a). 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.

Pluye, P., Grad, R. M., Levine, A., & Nicolau, B. (2009b). Understanding divergence of quantitative and qualitative data (or results) in mixed methods studies. *International Journal of Multiple Research Approaches, 3*(1), 58-72.

Pluye, P., Garcia Bengoechea, E., Granikov, V., Kaur, N., & Tang, D. L. (2018). A world of possibilities in mixed methods: Review of the combinations of strategies used to integrate the phases, results, and qualitative and quantitative data. *International Journal of Multiple Research Approaches*, 10(1), 41-56.

Porta, M. S., Greenland, S., Hernán, M., dos Santos Silva, I., Last, J. M. (2014). *A dictionary of epidemiology*. New York: Oxford University Press.

Sandelowski, M. (2010). What's in a name? Qualitative description revisited. *Research in Nursing and Health, 33*(1), 77-84.

Schwandt, T. A. (2015). *The SAGE dictionary of qualitative inquiry*. Thousand Oaks, CA: SAGE Publications.

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.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal, 355(*i4919).

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences.* Thousand Oaks, CA: SAGE Publications.

Thomas, B. H., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: Providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176-184.

Van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., & Editorial Board of the Cochrane Collaboration Back Review Group. (2003). Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine, 28*(12), 1290-1299.

Viswanathan, M., & Berkman, N. D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology, 65*(2), 163-178.

Wells, G., Shea, B., O'connell, D., Peterson, J., Welch, V., Losos, M., et al. (2000). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Retrieved April 16, 2016, from http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf.

Zaza, S., Wright-De Agüero, L. K., Briss, P. A., Truman, B. I., & Hopkins, D. P. (2000). Data collection instrument and procedure for systematic reviews in the guide to community preventive services. *American Journal of Preventive Medicine, 188*(Suppl 1), 44-74.