Applying GRADE in systematic reviews of complex interventions: Challenges and considerations for a new guidance

---

Ani Movsisyan

Department of Social Policy & Intervention
University of Oxford
Wolfson College

Hilary Term, 2018

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Social Intervention at the University of Oxford

Word count: approx. 81 500

Table of Contents

Acknowledgements

It would have been impossible for me to write this thesis without the contribution and help of many people.

First of all, I would like to thank my supervisor, Prof Paul Montgomery, for providing me with the idea for this project and the opportunity to work as a researcher on the GRADE Guidance for Complex Interventions project (GRADE-CI). I am immensely thankful for his trust, support, and encouragement throughout this work. His guidance over the many years of my work on this thesis was always wise and friendly. I also much appreciate the guidance received from my second supervisor, Dr David Humphreys, with whom I came to work at the end of my project. The psychological support received from him were instrumental in motivating the write-up of my thesis work.

I would like to thank all the participants of the GRADE-CI project, those who participated in the interviews, the Delphi process, and the expert meeting. Many thanks for your time and contribution, I have learned a great deal from interacting with you all.

I am immensely grateful to Dr Sean Grant for his continuous support, mentorship, and friendship throughout this project. I am much looking forward to many more years of collaboration.

I sincerely appreciate Prof Eva Rehfuess for her genuine interest in this project, inspiration on thinking creatively about methods in public health, and commitment to high quality and impactful research.

Finally, I thank my family and friends (in Yerevan and in S17) for their enduring support and faith in me during all these years.

List of Figures

List of Tables

List of Abbreviations

AHRQ	Agency for Healthcare Research and Quality
CERQual	Confidence in the Evidence from Reviews of Qualitative research
CONSORT	Consolidated Standards of Reporting Trials
EBP	Evidence-Based Practice
EtD	Evidence to Decision
GEPHI	Grading of Evidence for Public Health Interventions
GGG	GRADE Guidance Group
GRADE	Grading of Recommendations Assessment, Development and Evaluations
GWG	GRADE Working Group
iCAT_SR	intervention Complexity Assessment Tool for Systematic Reviews
JCE	Journal of Clinical Epidemiology
LMICs	Low- and Middle-Income Countries
MRC	Medical Research Council
NICE	National Institute for Health and Care Excellence
NRS	Nonrandomised Studies
OIS	Optimal Information Size
PICO	Population, Intervention, Comparison, Outcome
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QCA	Qualitative Comparative Analysis
RCT	Randomised controlled trial
RoB	Risk of Bias
ROBINS-I	Risk of Bias in Nonrandomised Studies of Interventions
SoFs	Summary of Findings
USPSF	US Preventive Services Task Force
WHO	World Health Organization

Applying GRADE in systematic reviews of complex interventions: Challenges and considerations for a new guidance

Ani Movsisyan
Wolfson College, University of Oxford
DPhil Social Intervention
Hilary Term, 2018

Abstract

Background: The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach offers a transparent framework for rating the certainty of evidence in systematic reviews. Concerns, however, have been raised that use of GRADE beyond biomedical interventions frequently downgrades the "best evidence possible" for many complex interventions. This DPhil thesis aims to (1) further investigate the challenges of using GRADE in systematic reviews of complex interventions, (2) explore how the GRADE approach can be advanced to address these challenges, and (3) inform the write-up and dissemination of a new GRADE guidance for complex interventions.

Methods: To address the broad aims of this thesis a range of methodological approaches were employed, primarily drawing on the best-practice techniques for developing research reporting guidelines (see Chapter 2). First, a systematic literature review method was used to establish whether an adequate system already exists for rating the certainty of evidence for complex interventions and informing the need for a new guidance (Chapter 3). Further consultation with experts, including semi-structured interviews with review authors and GRADE methodologists, provided a nuanced understanding of the challenges of applying GRADE in reviews of complex interventions and suggestions for advancing the guidance on GRADE (Chapter 4). Agreement around these suggestions was explored in a Delphi-based online expert panel (Chapter 5), and

the content of the new GRADE guidance for complex interventions was discussed in-depth in a three-day expert meeting held in Oxford in May 2017 (Chapter 6).

Results: The systematic literature review identified a few systems attempting to modify GRADE for public health interventions; however, there was little reporting of rigorous procedures in the development and dissemination of these systems. Qualitative interviews captured differences in views on GRADE use between review authors and GRADE methodologists. Specifically, GRADE methodologists found it critical to consider GRADE from the beginning of the review process and exercise judgment in GRADE ratings. Review authors, on the other hand, often thought of GRADE as an "annoying add-on" at the end of the review process and felt challenged by the need to be more interpretative with evidence and sift through many publications on GRADE. Suggestions were made to enhance the GRADE guidance. No significant disagreement was found in the online expert panel on any domain of evidence, and the expert meeting provided further insights into the content of the new GRADE guidance for complex interventions. Participants agreed that the new guidance should specify the meaning of the construct of "certainty of evidence" for complex interventions, consider revisions of the initial categorisation of evidence based on study design, and better assess the coherence of the causal pathway of complex interventions.

Conclusion: This thesis work consolidates up-to-date methodological knowledge on reviewing complex interventions by providing critical examination of the existing approaches and new insights. In transparent reporting of the research phases, it informs development of a new GRADE guidance on rating the certainty of evidence in systematic reviews of complex interventions.

Dissemination of the thesis work

Peer-reviewed journal articles

Published

- **Movsisyan A**, Melendez-Torres GJ, Montgomery P. (2016). A harmonised guidance is needed on how to "properly" frame review questions to make the best use of all available evidence in the assessment of effectiveness of complex interventions. *Journal of Clinical Epidemiology*, 77,139-141.
- **Movsisyan A**, Melendez-Torres GJ, Montgomery P. (2016). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *Journal of Clinical Epidemiology*, 70,191-199.
- **Movsisyan A**, Melendez-Torres GJ, Montgomery P. (2016). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared to those of simple interventions. *Journal of Clinical Epidemiology*, 78, 22-33.
- **Movsisyan A**, Dennis J, Rehfuess E, Grant S, Montgomery P. (2018). Rating the quality of a body of evidence on the effectiveness of health and social interventions: a systematic review and mapping of evidence domains. *Research Synthesis Methods,* In press.

In submission

- **Movsisyan A**, Montgomery P, Grant S. Experiences and practices of using GRADE in systematic reviews of complex interventions: Semi-structured interviews with review authors and methodologists. *Systematic Reviews.*
- **Movsisyan A**, Grant S, Montgomery P. Rating the certainty of evidence in systematic reviews of complex interventions: an expert meeting. *PLOS ONE.*
- Grant S, **Movsisyan A**, Montgomery P. Rating the certainty of evidence in systematic reviews of complex interventions: an online-modified Delphi process. *Implementation Science.*
- Montgomery P, **Movsisyan A**, Grant S. Considerations of complexity in rating the certainty of evidence in systematic reviews: A primer on using the GRADE approach in global health. *BMJ Global Health.*

Academic conferences and workshops

Oral presentations

- Montgomery P, Mayo-Wilson E, Grant S, **Movsisyan A.** (26 September 2016). Registering, reporting and reviewing social intervention trials to inform policy and practice. Panel presentation at the What Works Global Summit. London, UK.
- **Movsisyan A.** (12 March 2016). Applying the GRADE approach to complex interventions: An empirical investigation and development of a GRADE extension. World Congress on Public Health & Nutrition. Madrid, Spain.

Seminars and workshops

- **Movsisyan A**. (20 July 2017). Using the GRADE approach in systematic reviews of complex interventions: the framework, challenges and ways forward. Centre for Evidence-Based Intervention (CEBI) Research Group Meeting, Department of Social Policy and Interventions, University of Oxford. Oxford, UK.
- **Movsisyan A**, Montgomery P, Grant S. (24 May 2017). Developing GRADE guidance for rating the certainty of evidence in systematic reviews of complex interventions. A three-day expert meeting organised by the project on *Developing GRADE guidance for complex interventions* to finalise the content of the new GRADE guidance. Oxford, UK.
- **Movsisyan A**, Montgomery P, Grant S. (26 April 2017). Developing GRADE guidance for rating the certainty of evidence of complex interventions. GRADE Working Group meeting. Rome, Italy.
- Montgomery P, **Movsisyan A**, Grant S. (23 January 2017). Using GRADE in systematic reviews of complex interventions. A two-day workshop organised by the Department of Maternal, Newborn, Child and Adolescent Health, World Health Organization *on retrieval, synthesis and assessment of evidence on complex, multidisciplinary interventions*. Freising, Germany.
- **Movsisyan A**, Grant S. (25 October 2016). Grading the certainty of a body of evidence for complex interventions. A meeting organised at the 24th Cochrane Colloquium. Seoul, South Korea.
- Grant S, **Movsisyan A.** (20 September 2016). Grading the quality of a body of evidence for complex interventions: challenges and the way forward. Webinar delivered as part of the Evidence-based Practice Center (EPC) learning webinar series, Agency for Healthcare Research and Quality.
- **Movsisyan A**, Montgomery P. (16 August 2016). Assessing the quality of the body of evidence underpinning statements on the effectiveness of social interventions. A two-day workshop held at the Department of Maternal, Newborn, Child and Adolescent Health, World Health Organization on *retrieval, synthesis and assessment of evidence on complex, multidisciplinary interventions*. Geneva, Switzerland.
- **Movsisyan A**, Montgomery P. (19 April 2016). Developing a GRADE extension for complex social interventions. NCRM Annual Centre Meeting. Chilworth Manor, Southampton, UK.
- Montgomery P, **Movsisyan A**. (18 February 2016). GRADE extension for complex social interventions. A scoping meeting held at the Department of Maternal, Newborn, Child and Adolescent Health, World Health Organization on *retrieval, synthesis and assessment of evidence on complex, multidisciplinary interventions*. Geneva, Switzerland.
- Montgomery P, **Movsisyan A**. (5 March 2015). Introducing an extension to the GRADE approach for complex social interventions. GRADE Working Group meeting. Amsterdam, Netherlands.

Poster

- Montgomery P, **Movsisyan A**, Grant S. (31 May 2016). Development of a GRADE extension for complex social interventions. Poster presented at the 24th Annual Meeting of the Society for Prevention Research (SPR). San Francisco, CA, USA.

Introduction to thesis

While assessing risks of bias in each *individual* study included in a systematic review is an important and well-established practice (Higgins & Green, 2011; Juni, Altman, & Egger, 2001), rating the certainty of evidence is a comparatively new practice that indicates the credibility and trustworthiness of a *body of evidence* synthesised across many studies in relation to a specific research question (Gough, 2007; Sackett, 2000). Systems for rating the certainty of evidence typically involve an examination of various characteristics of a body of evidence (for example, risk of bias, heterogeneity of evidence across individual studies) that ultimately results in an overall rating of that body of evidence (for example, “high”, “moderate”, or “low” certainty of evidence). These ratings are then commonly used in guideline development to inform recommendations about implementing interventions in practice (West, King, & Carey, 2002). The certainty of evidence rating can, therefore, have a large influence on how decision-makers use evidence from evidence syntheses.

Established by an international collaboration of professionals, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach provides the most prominent framework for rating the certainty of evidence in systematic reviews in order to inform development of recommendations in clinical and public health guidelines (Guyatt et al., 2008). It is currently used by more than 100 organisations worldwide and provides an outcome-centric approach to rating the certainty of evidence (Balshem et al., 2011). This means that ratings are provided for a body of evidence contributing to each separate outcome in a systematic review. Primarily developed and validated for biomedical interventions, application of GRADE to interventions in areas of

broader health and social policy has, however, been difficult with some representatives of the public health community even considering it inappropriate (Movsisyan, Melendez-Torres, & Montgomery, 2016b; Rehfuess & Akl, 2013). Interventions in these areas of practice are contrasted to "simpler" biomedical interventions and are frequently described as "complex".

There is a growing interest in complex interventions (Craig et al., 2008), and many projects have been launched in the recent years aiming to provide guidance for considering complexity in systematic reviews of interventions (see Chapter 1 for a further discussion on "complexity" in systematic reviews). These include a series of articles published in the *Journal of Clinical Epidemiology* in 2013 (Anderson, Petticrew, et al., 2013), development of the iCAT-SR tool for assessing the complexity of interventions within systematic reviews (Lewin et al., 2017), and another series of articles published in the *Journal of Clinical Epidemiology* in 2017 by the Agency for Healthcare Research and Quality (AHRQ) (Guise, Chang, Butler, Viswanathan, & Tugwell, 2017). The AHRQ series also includes papers describing an extension of the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRIMSA) for complex interventions (Guise, Butler, et al., 2017a; Guise, Butler, et al., 2017b). In light of these efforts calling to incorporate a "complexity perspective" in systematic reviews and practice guidelines (Rutter et al., 2017), this thesis seeks to further explore issues around the GRADE approach in the context of complex interventions, and provide specific considerations for developing a new GRADE guidance tailored to these interventions. This research is expected to inform the write-up of a GRADE guidance for complex interventions.

This thesis includes a total of 7 chapters. Chapter 1 provides a broad overview of the literature, including introduction of the GRADE approach, detailed discussion of the

different approaches to conceptualising complexity, and implications of complexity for systematic reviews. Existing literature on reported challenges of using GRADE in reviews of complex interventions is also summarised, which includes two manuscripts by the DPhil candidate published in the *Journal of Clinical Epidemiology* (Movsisyan, Melendez-Torres, & Montgomery, 2016a; Movsisyan et al., 2016b). Chapter 2 specifies the strategy adopted by this thesis work to explore development of the GRADE guidance for complex interventions. This consists of four phases of research drawing on the best practices for developing research reporting guidelines (Moher, Schulz, Simera, & Altman, 2010). Chapters 3 to 6 describe the results of this thesis project. Specifically, Chapter 3 reports on a systematic literature review summarising the existing systems for rating the certainty of evidence from health and social policy fields; a manuscript adaptation of this chapter has been published in *Research Synthesis Methods* (Movsisyan, Dennis, Rehfuess, Grant, & Montgomery, 2018). Chapter 4 discusses results from an in-depth qualitative investigation into using GRADE in reviews of complex interventions by way of comparing experiences and practices of review authors and GRADE methodologists; a manuscript adaptation of this chapter has been submitted for publication in *Systematic Reviews*. Chapter 5 describes on an online-modified Delphi panel to explore areas of agreement and disagreement among stakeholders about domains of evidence to consider for the new GRADE guidance; a manuscript adaptation of this chapter has been submitted for publication in *Implementation Science*. Chapter 6 summarises discussions from a formal face-to-face expert meeting aiming to finalise the content of the new GRADE guidance; a manuscript adaptation of this chapter has been submitted for publication in *PLOS ONE.* Finally, Chapter 7 discusses the contribution of this thesis work and implications for writing and disseminating an official GRADE guidance for complex

interventions. It also outlines the strengths and limitations of the work and provides recommendations for future research. Drafts from this chapter have been included in a manuscript in submission for publication in *BMJ Global Health* as part of the World Health Organization (WHO) series of papers on strengthening the methods for retrieval, synthesis and assessment of complex health interventions. Collectively, these chapters contribute to four phases of research into development of the GRADE guidance for complex interventions. Figure I outlines the logic model of this thesis, and how the thesis chapters contribute to the different phases of the research (Moher et al., 2010). It also outlines the papers from this thesis work, which are already published or in submission.

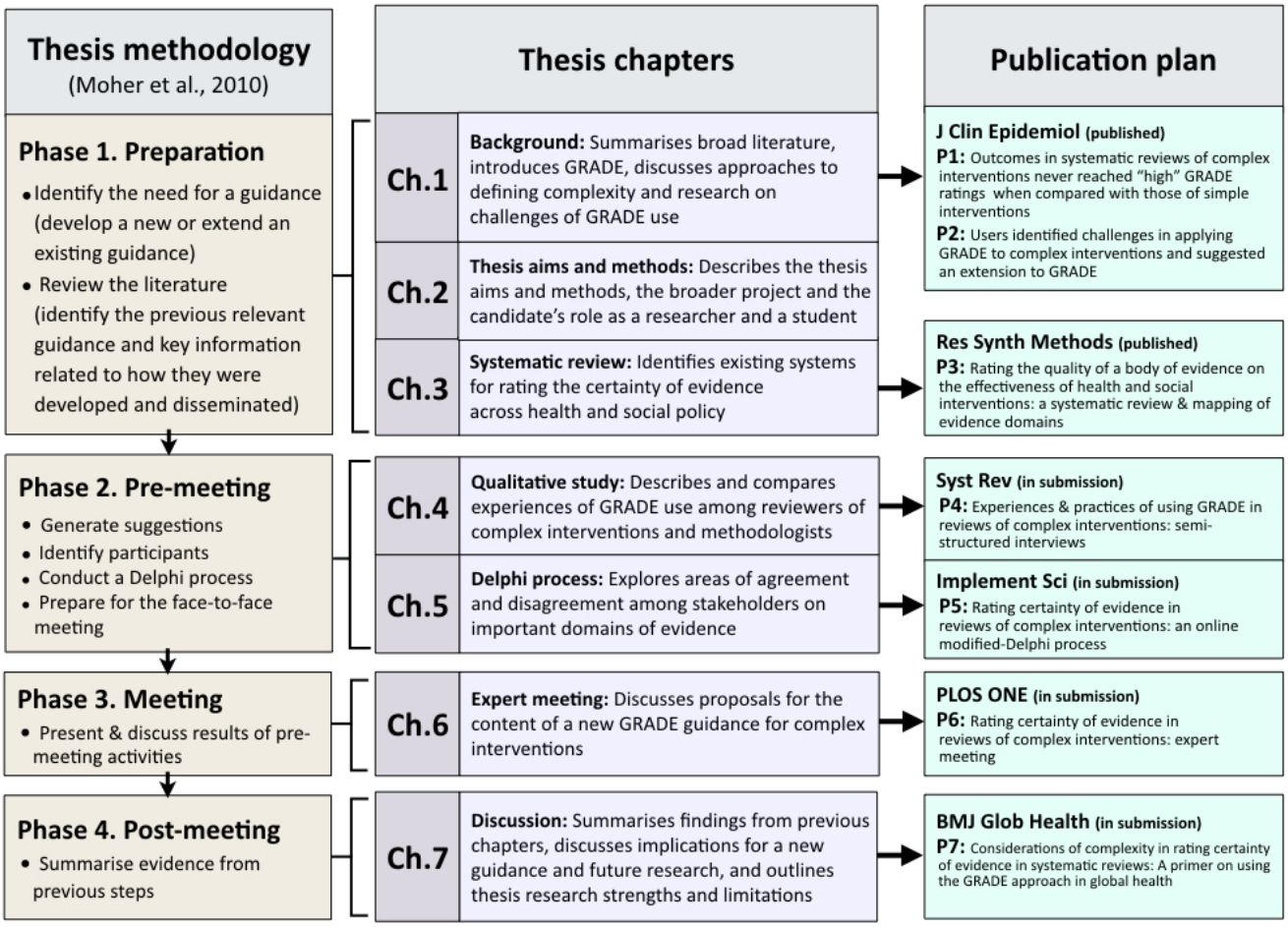


Figure I. Thesis logic model outlining how the thesis chapters link with the overall thesis methodology (Moher et al., 2010) and papers published or in submission

Key definitions for thesis

Interventions

Interventions can be generally defined as sets of activities, professional behaviours, practices, or ways of organisation implemented with the intention to bring changes (i.e., outcomes of action) by modifying the malleable determinants (i.e., mechanisms of action) (Craig et al., 2008; Fraser, 2009; Tanner-Smith & Grant, 2018). Following this broad definition, interventions can be classified in a variety of ways, such as based on the outcomes, mechanisms and/or levels of action. For example, public health interventions aim to bring about changes in health-related outcomes among large population groups (Rehfuess & Akl, 2013). Depending on the levels of action, further types of public health interventions can be distinguished. For example, public health interventions targeting the upstream determinants of health, such as macro environmental factors are often referred to as population health and policy (PHP) interventions (Armstrong et al., 2015), as they require community or policy-level actions. Examples of these interventions include restriction or banning of alcohol advertising (Petticrew et al., 2017) and built environment interventions for physical activity (Tully et al., 2013). On the other hand, interventions targeting downstream determinants of health, such as individual characteristics and behaviours require only individual- or group-level actions (Michie, van Stralen, & West, 2011; Rychetnik, Hawe, Waters, Barratt, & Frommer, 2004; Tanner-Smith & Grant, 2018). Examples of these interventions include family and parenting interventions for mental health (Woolfenden, Williams, & Peat, 2001) and brief alcohol interventions for adolescents and young adults (Tanner-Smith & Lipsey, 2015).

Social interventions

Interventions have also been classified based on the mechanisms of action, i.e., the processes that they intend to modify to affect certain outcomes. Following this logic, Grant et al. (2014) distinguish between "biomedical" and "social" interventions. While biomedical interventions operate through biological mechanisms, such as physiological processes, social interventions work *through "malleable mental processes and social phenomena, such as cognitions, emotions, behaviours, interpersonal relationships, and salient physical and social environments"* (Grant et al., 2014). Social interventions are studied across a range of practice domains, including criminology, public health, psychology, education, and social work and social welfare (see Table I). Another key distinction of social interventions highlighted by the authors is the agency of those engaging with the intervention (Grant, 2014; May, 2013; Montgomery et al., 2013). As participants in social interventions are active agents in a social system, they may respond differently to the intervention activities (Hawe, Shiell, & Riley, 2009). As a result, social interventions are often described as "complex" (Craig et al., 2008), and thus warrant questions beyond potentially arbitrary estimates of effect, such as "what happens" and through what mechanism(s) interventions work (or not work) when delivered in a social system (Petticrew, 2015).

Complex interventions

While approaches to conceptualising complex interventions are still evolving, most definitions of complex interventions in intervention research draw heavily on the dimensions of complexity initially described in the UK Medical Research Council (MRC) guidance for developing and evaluating complex interventions (Craig et al., 2008). More

recent definitions of complex interventions, however, extend the MRC dimensions to also incorporate concepts from complex systems thinking, such as dynamic and non-linear causal pathways between interventions and outcomes and context-specificity of the effects (Lewin et al., 2017; Petticrew, Anderson, et al., 2013). As shown in Table II, complex interventions can be described based on (1) the sources of the intervention itself (for example, interventions having multiple components), and (2) the sources of the interventions' causal pathway (for example, dynamic and non-linear relationships between interventions and outcomes). A more detailed discussion of different perspectives to conceptualising complexity in intervention research and the rationale for the approach to defining complex interventions supported in this thesis is presented in Chapter 1.

Table I. Examples of practice domains studying social interventions

	Intervention	**Mechanisms of Action**	**Outcomes**
Public health	Built environment interventions	Structural, self-efficacy, social support, motivation	Physical activity, mental health, quality of life
Education	Technical & vocational education	Motivation, knowledge skills, behaviours	Employability
Crime & Justice	Restorative justice conferencing	Emotions and behaviours	Repeat offending, victim satisfaction
Psychology	Parenting interventions	Cognitions, relationships. Emotions, behaviours	Conduct behaviour, mental health
Social work & welfare	Kinship care	Social environment, positive bonds	Mental health, educational attainment
International development	Saving promotion interventions in LMICs	Social environment, norms, attitudes, self-efficacy, self-control, behaviours	Poverty, educational attainment, health

Table II. Sources of complexity in systematic reviews, adapted from Petticrew et al. (2013)

1. Characteristics of the intervention itself
Number of components
Number of groups or organisational levels targeted
Degree of flexibility or tailoring permitted
Self-organisation, adaptivity and changes over time
2. Characteristics of the intervention's causal pathway
Number and variability of outcomes
Non-linear relationships and phase changes
Number of mediators and moderators of effect
Positive/negative feedback loops
Synergies/dysynergies between components
Interaction with context

Certainty of evidence

In this thesis, certainty of evidence is broadly used to describe the credibility and trustworthiness of a body of evidence, derived from evidence synthesis that integrates results across individual studies. This body of evidence represents a *totality* of evidence across studies in relation to a specific research question (e.g., in GRADE, certainty of evidence denotes confidence in the estimates of effect for a specific outcome in a systematic review). Evidence rating systems are used to rate the certainty of evidence (Guyatt et al., 2008; West et al., 2002). By comparison, quality or critical assessment tools are designed to be used for assessing the trustworthiness of different types of *individual* studies, such as randomised controlled trials, systematic reviews, diagnostic studies, qualitative studies. Examples of these tools include Critical Appraisal Skills Programme (CASP) checklists for individual study designs ("Critical Appraisal Skills Programme", 2017).

Practice guidelines

Practice guidelines are commonly defined as documents containing recommendations on interventions, whether these are clinical, public health or policy, informed by a systematic review of evidence, including an assessment of the benefits and harms of alternative interventions (IoM, 2011; NICE, 2014; WHO, 2014). According to the World Health Organization (WHO) handbook for guideline development, a recommendation provides information about what policy-makers, health-care providers or patients should do, implies a choice between different interventions and has implications for the use of resources (WHO, 2014).

The GRADE approach

The most robust and widely applied evidence rating system is the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. It provides a structured framework and describes specific domains to rate the certainty of evidence in systematic reviews and practice guidelines (Guyatt, Oxman, Schunemann, Tugwell, & Knottnerus, 2011). These domains have been developed by the GRADE Working Group through expert consensus. Over the last fifteen years, the GRADE approach has been widely adopted by systematic reviewers and guideline developers in healthcare, including over 100 organisations worldwide. For instance, WHO uses the GRADE approach to inform global recommendations on multi-disciplinary interventions and public health policies (WHO, 2014); the Cochrane Collaboration has also mandated use of GRADE in Cochrane intervention reviews (Higgins, Lasserson, Chandler, Tovey, & Churchill, 2016). Further details about this approach and the activities of the GRADE Working Group are discussed in Chapter 1.

References

Anderson, L. M., Petticrew, M., Chandler, J., Grimshaw, J., Tugwell, P., O'Neill, J., . . . Shemilt, I. (2013). Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol, 66*(11), 1205-1208.

Armstrong, R., Campbell, M., Craig, P., Hoffmann, T., Katikireddi, S. V., & Waters, E. (2015). Reporting guidelines for population health and policy interventions: TIDieR-PHP. *Lancet, 386*(S19).

Barlow, J., & Coren, E. (2018). The effectiveness of parenting programs: A review of Campbell Reviews. Res Soc Work Pract, *28*(1), 99-102.

Balshem, H., Helfand, M., Schunemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., . . . Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol, 64*(4), 401-406.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, P. (2008). Developing and evaluating complex interventions: new guidance. Retrieved 8 Feb, 2018 from https://www.mrc.ac.uk/documents/pdf/developing-and-evaluating-complex-interventions/

Critical Appraisal Skills Programme (CASP) (2017). Retrieved 5 Feb, 2018 from http://www.casp-uk.net/

Fraser, M., W. (2009). *Intervention research: developing social programs*. New York, Oxford: Oxford University Press.

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. Research Papers in Education, 22(2), 213-228.

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. (2017). Retrieved 5 Feb, 2018 from http://gradeworkinggroup.org/

Grant, S. (2014). *Development of a CONSORT extension for social and psychological interventions.* DPhil thesis. Department of Social Policy and Intervention, Oxford: University of Oxford.

Guise, J. M., Butler, M., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017a). AHRQ series on complex intervention systematic reviews-paper 7: PRISMA-CI elaboration and explanation. *J Clin Epidemiol, 90*, 51-58.

Guise, J. M., Butler, M. E., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017b). AHRQ series on complex intervention systematic

reviews-paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol, 90*(43-50).

Guise, J. M., Chang, C., Butler, M., Viswanathan, M., & Tugwell, P. (2017). AHRQ series on complex intervention systematic reviews-paper 1: an introduction to a series of articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol, 90*, 6-10.

Guyatt, G. H., Oxman, A., D., Schunemann, H. J., Tugwell, P., & Knottnerus, A. (2011). GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol, 64*(4), 380-382.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., . . . GRADE Working Group (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*(7650), 924-926.

Hawe, P., Shiell, A., & Riley, T. (2009). Theorising interventions as events in systems. *Am J Community Psychol, 43*(3-4), 267-276.

Higgins, J., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. Retrieved 20 Dec, 2017 from http://www.handbook.cochrane.org/

Higgins, J., Lasserson, T., Chandler, J., Tovey, D., & Churchill, R. (2016). *Methodological expectations of Cochrane interventions reviews*. Cochrane: London.

IoM. (2011). Clinical practice guidelines we can trust. Institute of Medicine (US) Committee on Standards for Developing Trustwothy Clinical Practice Guidelines. Retrieved 8 Feb, 2018 from https://www.ncbi.nlm.nih.gov/books/NBK209539/

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

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., . . . Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

May, C. (2013). Towards a general theory of implementation. *Implement Sci, 8*, 18.

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci, 6*, 42.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2).

Möhler, R., Kopke, S., & Meyer, G. (2015). Criteria for Reporting the Development and Evaluation of Complex Interventions in healthcare: revised guideline (CReDECI 2). *Trials, 16*, 204.

Montgomery, P., Grant, S., Hopewell, S., Macdonald, G., Moher, D., Michie, S., & Mayo-Wilson, E. (2013). Protocol for CONSORT-SPI: an extension for social and psychological interventions. *Implement Sci, 8*, 99.

Movsisyan, A., Dennis, J., Rehfuess, E., Grant, S., & Montgomery, P. (2018). Rating the quality of a body of evidence on the effectiveness of health and social interventions: A systematic review and mapping of evidence domains. *Res Synth Methods.* Accepted Article, In Press: doi: 10.1002/jrsm.1290.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016a). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol, 78*, 22-33.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016b). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

NICE. (2014). Developing NICE guidelines: the manual. National Institute for Health and Care Excellence. Retrieved 7 Feb, 2018 from https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview

Petticrew, M. (2015). Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev, 4*(36).

Petticrew, M., Anderson, L., Elder, R., Grimshaw, J., Hopkins, D., Hahn, R., . . . Welch, V. (2013). Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol, 66*(11), 1209-1214.

Petticrew, M., Shemilt, I., Lorenc, T., Marteau, T. M., Melendez-Torres, G. J., O'Mara-Eves, A., . . . Thomas, J. (2017). Alcohol advertising and public health: systems perspectives versus narrow perspectives. *J Epidemiol Community Health, 71*(3), 308-312.

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T., . . . White, M. (2017). The need for a complex systems model of evidence for public health. *Lancet*, 9;390(10112), 2602-2604.

Rychetnik, L., Hawe, P., Waters, E., Barratt, A., & Frommer, M. (2004). A glossary for evidence based public health. *J Epidemiol Community Health, 58*(7), 538-545.

Sackett, D., L., Straus, S., E., Richardson, M., Rosenberg, W., S., & Haynes, R., B. (2000). *Evidence-Based Medicine: how to practice and teach EBM* (2nd ed.). London: Churchill Livingstone.

Tanner-Smith, E. E., & Grant, S. (2018). Meta-analysis of complex interventions. *Annu Rev Public Health, 39*, 16.11-16.17.

Tanner-Smith, E. E., & Lipsey, M. W. (2015). Brief alcohol interventions for adolescents and young adults: a systematic review and meta-analysis. *J Subst Abuse Treat, 51*, 1-18.

Tully, M. A., Kee, F., Foster, C., Cardwell, C. R., Weightman, A. L., & Cupples, M. E. (2013). Built environment interventions for increasing physical activity in adults and children. *Cochrane Database Syst Rev, (1),* CD010330.

West, S., King, V., & Carey, T. e. a. (2002). *Systems to rate the strength of scientific evidence.* Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). Rockville, MD: Agency for Healthcare Research and Quality.

WHO. (2014). *World Health Organization Handbook for Guideline Development* (2nd ed.). Geneva, Switzerland: WHO Press.

Woolfenden, S. R., Williams, K., & Peat, J. (2001). Family and parenting interventions in children and adolescents with conduct disorder and delinquency aged 10-17. *Cochrane Database Syst Rev,* (2), CD003015.

Chapter 1. Background

Two papers including drafts from this chapter have been published in the *Journal of Clinical Epidemiology*

Chapter overview

This chapter provides an overview of literature and discusses the problem leading to the thesis aims. First, key principles of the evidence-based practice (EBP) model are described, including the role and utility of systematic reviews in intervention research, and the GRADE approach is introduced as a system for rating the certainty of evidence largely embedded in EBP. Further, using a historical perspective, it discusses existing approaches to conceptualising complexity in intervention research and outlines the approach to defining complexity supported in this thesis. The chapter then provides an in-depth discussion of the implications of complexity for each phase in a review process, from formulating review questions to synthesising and assessing evidence. Finally, the reported challenges of using the GRADE approach in systematic reviews of complex interventions are summarised, and the case is made for a new tailored GRADE guidance, which addresses sources of complexity in systematic reviewing. The presented literature includes two manuscripts by the DPhil candidate on the use of GRADE in systematic reviews of complex interventions published in the *Journal of Clinical Epidemiology* at the outset of the thesis project.

Evidence-based practice (EBP) model

Principles of EBP

The evidence-based practice (EBP) is defined as a model of decision making based on the conscientious, explicit and judicious use of the best available evidence integrated with client needs, values, preferences and professional expertise (Gambrill, 2006; Haynes, Devereaux, & Guyatt, 2002; Sackett, 2000). The EBP model has been visually depicted by Haynes et al. (2002) (see Figure 1.1) largely expanding from the evidence-based medicine. In early 1990s, the evidence-based medicine (EBM) has been proposed as a "new paradigm" for the practice of medicine emphasising the importance of incorporating the best research findings into the management of patient care, therefore, introducing more scientific rigour into clinical decision-making (Guyatt et al., 1995; Howick, 2011; Sur & Dahm, 2011).

Sackett et al. (2000) describe five steps in EBP:

> *"Step 1—formulate an answerable question;*
>
> *Step 2—track down the best evidence with which to answer that question;*
>
> *Step 3—critically appraise that evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in practice);*
>
> *Step 4—integrate the critical appraisal with professional expertise, client preferences, values and circumstances;*
>
> *Step 5—evaluate the effectiveness and efficiency in executing Steps 1 to 4 and improve them for next time" (pp. 3-4).*

This linear and standardised process promotes an approach for practice that emphasises the use of high-quality research evidence to inform decisions regarding the effective interventions. In the meantime, Step 3, that is, the critical appraisal of evidence, reflects another key principle of the EBP paradigm. This involves use of systematic and

transparent methods to examine the quality of research evidence, its trustworthiness and validity (Wortman, 1994). Finally, placed in the core of the model, is the professional expertise as a key ability of eliciting, appraising and integrating different sources of evidence. In this way, the EBP model is contingent upon the availability of rigorous research evidence, which can be critically appraised and integrated with other considerations in the model to draw the best practice decisions. While developed and first used in the context of clinical medicine to manage individual patients (Satterfield et al., 2009), the EBP model has since been advocated in other fields of practice, such as public health, highlighting the importance of the valid and reliable research evidence in decisions on population health and society at large.

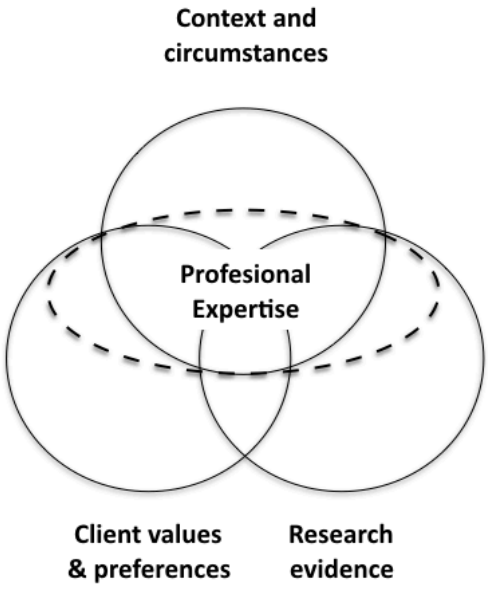


Figure 1.1. Evidence-based practice model, adapted from Haynes et al. (2002)

Standards of evidence in EBP

To emphasise the importance of high quality evidence in decision making, much attention in the EBP model has been given to advancing the methods for rigorous

conduct and appraisal of research. In the EBP tradition, this is well illustrated by employment of hierarchies of evidence to guide practitioners in the identification of the "best evidence" (Steps 2 and 3). The traditional hierarchy of evidence focusing on the research of interventions classifies different study designs according to their ability to minimise systematic error, i.e., bias when estimating the effects of interventions (see Figure 1.2) (Flay et al., 2005; Guyatt et al., 1995; Sackett, 2000; Shekelle, Woolf, Eccles, & Grimshaw, 1999).

Figure 1.2. Hierarchy of evidence, adapted from Guyatt et al. (1995)

Randomised controlled trials (RCTs) are regarded as the "gold standard" of comparative studies and are prioritised in the hierarchy, as through random allocation they enable comparable treatment groups, therefore, creating the basis for causal

attribution of any observed effect to the intervention (Field & Hole, 2003). In the meantime, systematic reviews synthesising evidence from multiple primary studies are placed at the top level of the hierarchy. Commensurate with what philosophers call the "principle of total evidence" (Carnap, 1947), that is, a need to include all relevant information for making a valid inductive argument, systematic reviews are deemed a superior source of evidence in the EBP model, as they combine evidence across multiple individual studies.

Systematic reviews are broadly referred to as scientific investigations that reconcile the evidence on a particular problem by way of synthesising and critically appraising the results of primary studies with the application of systematic, transparent, and reproducible methods to minimise bias (Cook, Mulrow, & Haynes, 1997; Gough, Oliver, & Thomas, 2012; Jadad, Cook, & Browman, 1997; Petticrew & Roberts, 2006). The distinctive steps in the conduct of systematic reviews, which set them apart from other types of narrative reviews, largely follow those of the EBP model (Higgins & Green, 2011). These include, (1) formulation of a specific research question based on the PICO approach (Participants, Interventions, Comparisons & Outcomes), (2) adoption of an explicit and reproducible search strategy, (3) use of comprehensive sources with the attempt to identify all studies that would meet the eligibility criteria, (4) rigorous critical appraisal of the methodological quality of the included studies, such as risk of bias assessment, and (5) synthesis of the results and /or meta-analyses, which involves synthesis of quantitative findings with the employment of statistical techniques (Lipsey & Wilson, 2001). This level of scientific rigour and transparency is anticipated to enhance the quality of the review evidence with better control over systematic error that

potentially threatens any retrospective type of research project (Petticrew & Roberts, 2006).

While the evidence hierarchy has served as an important heuristic for researchers and practitioners to help judge the quality of research evidence over years, and is still widely used, it has also been criticised for failing to consider the appropriateness of the different research designs for answering various research questions (Parkhurst & Abeysinghe, 2014). This concern has been particularly voiced in social disciplines, such as public health and social policy, where interventions are often described as "complex", and evaluations need to ask questions beyond intervention effectiveness, or, "does it work" (Moore et al., 2015; Petticrew, 2015; Rychetnik, Frommer, Hawe, & Shiell, 2002). In this light, some researchers have suggested to use typologies of evidence to help decide on the most appropriate design for their specific research question (Gray, 1997). One such typology has been described by Petticrew & Roberts (2003) (see Table 1.1). In this approach, research designs are rated for their appropriateness to answer a specific research question. For example, following the traditional evidence hierarchy, the top row in Table 1.1 rates the appropriateness of research designs for assessing intervention effectiveness; other rows of the typology list research designs that would be appropriate for answering additional research questions relevant for evaluation of complex interventions, such as intervention acceptability and salience (Petticrew & Roberts, 2003). This approach speaks to the recent calls made by Petticrew et al. to broaden the EBP model to ask "what happens" versus "what works" questions (Petticrew, 2015; Petticrew et al., 2017).

Table 1.1. A typology of evidence based on the research question, adapted from Petticrew & Roberts (2003) and Gray (1997)

Research question	Qualitative research	Survey	Case-controls	Cohort studies	RCTs	Quasi-experiments	Non-experiments	Systematic reviews
Effectiveness Does it work?				+	++	+		+++
Process of delivery How does it work?	++	+					+	+++
Salience Does it matter?	++	++						+++
Safety Will it do more good than harm?	+		+	+	++	+	+	+++
Acceptability Is there a will by the targeted audience to take up the intervention?	++	+			+	+	+	+++
Cost-effectiveness Is it worth buying this intervention?					++			+++
Appropriateness Is this the right intervention for this audience?	++	++						++
Satisfaction Are stakeholders satisfied with the service?	++	++	+	+				+

Notes: +++: "highly appropriate design"; ++: "moderately appropriate design"; +: "lowly appropriate design".

In a similar vein, in addition to answering the "what works" question, multiple functions have been described for systematic reviews, such as stocktaking, mining, and reality checking (Petticrew & Roberts, 2008). The stocktaking function of systematic reviews is sometimes referred to as the function of "mapping" the evidence base

(Petticrew & Roberts, 2008). It allows researchers to identify and possibly shift their current priorities by knowing with some degree of accuracy what has been done previously and where it can be located. Systematic reviews are seen to additionally act as drivers of future research by identifying gaps and highlighting areas of future work. In this view, they can be used as important sources of methodological and other information when planning new studies, and, thereby, help to mitigate research waste (Moher et al., 2016). This methodological and theoretical mining of existing research may inform development of new interventions (Munro, Lewin, Swart, & Volmink, 2007). Finally, the reality checking function of systematic reviews is also important, since systematic reviews provide a parallel commentary on current practices and may challenge the evidence base underlying specific decisions (Petticrew & Roberts, 2008).

The GRADE approach

Since the inception of the EBP model in 1990s, many projects have been launched, primarily in the field of healthcare, and, specifically, in clinical medicine, to advance the methods of systematic reviewing and critical appraisal of evidence. Because critical appraisal of evidence can be a rather subjective process, where researchers' *"predispositions toward a review's results may influence the judgements about the methodological quality of a piece of research"* (Cooper, 1984), structured approaches to critical appraisal have been suggested, such as quality appraisal tools (Jadad et al., 1996; Katrak, Bialocerkowski, Massy-Westropp, Kumar, & Grimmer, 2004; Voss & Rehfuess, 2013) and evidence rating systems (Guyatt et al., 1995; West et al., 2002) as a way to reduce inconsistencies in how evidence is assessed in systematic reviews and practice

guidelines. The leader among these, is the approach developed by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working Group.

The GRADE Working Group started its activities in 2000 as an informal collaboration of healthcare professionals, methodologists, systematic reviewers, and practice guideline developers. The primary goal of this group was to develop a systematic and structured approach for assessing evidence and making practice recommendations based on the synthesised and appraised evidence. The group aimed for this approach to be used universally to facilitate the evidence-based practice and decision-making in healthcare worldwide.

Embedded within the EBP paradigm, the GRADE approach can be broadly conceptualised as a process involving three main phases (see Figure 1.3) (Guyatt, Oxman, Akl, et al., 2011): the first phase contains a set of steps common to systematic reviews, such as formulation of a specific review question following the PICO approach, identification, appraisal and synthesis of evidence; the second phase involves application of the GRADE domains of evidence for rating the certainty of evidence produced by systematic reviews; and finally, the third phase is specific for guideline development , and involves steps for making practice recommendations. In this phase different factors of relevance to a health decision, such as issues around intervention acceptability, feasibility and equity are systematically considered in addition to the evidence on intervention effectiveness. While the latter phase is currently referred to as *Evidence to Decision (EtD) frameworks* (Alonso-Coello et al., 2016) and has been advanced by a separate group of researchers within the GRADE Working Group, in the context of systematic reviewing, “GRADE” is commonly used to denote the approach for rating the certainty in the estimates of effect of an intervention (i.e., Phase 2). Accordingly, in this

thesis, "GRADE" will be used hereafter, to refer to the certainty of evidence rating phase of the broader GRADE process (see Figure 1.3). This structure of the GRADE approach, which separates the certainty of evidence judgments in systematic reviews from those related to the strength of recommendations in practice guidelines (Guyatt et al., 2008) allows for an expanded model of decision making, meanwhile aiming to integrate all factors of relevance to a health decision (Cartwright & Hardie, 2012).

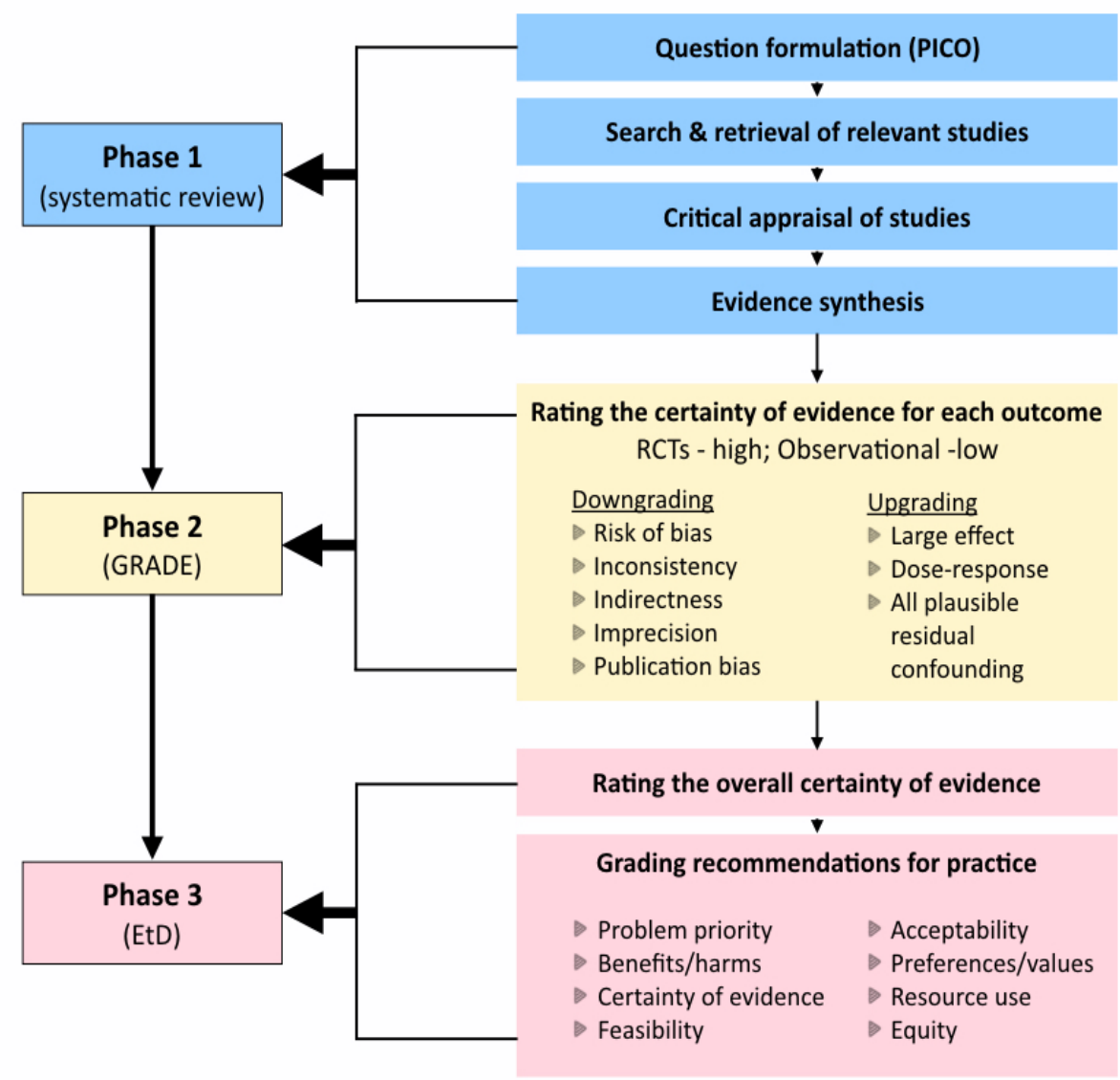


Figure 1.3. The GRADE process for systematic reviewing and developing practice recommendations, adapted from Guyatt et al. (2011) and Alonso-Coello et al. (2016)

Since its introduction, the GRADE approach has had significant influence in healthcare; it is currently endorsed by more than 100 organisations worldwide, including the leading organisations in systematic reviewing and guideline development in clinical medicine and public health, such as the Cochrane Collaboration, the World Health Organization (WHO), UpToDate, and the National Institute for Health and Care Excellence (NICE) Clinical Guidelines Programme ("GRADE Working Group," 2017).

GRADE for rating the certainty of evidence in systematic reviews

As shown in Figure 1.3, the GRADE approach has been designed for systematic reviews asking a structured question on the comparative effectiveness of an intervention, as formulated by the PICO elements. In this regard, GRADE offers an "outcome centric" method for rating the certainty of evidence generated from systematic reviews, i.e., it rates the certainty of evidence for each outcome of a systematic review separately (Guyatt et al., 2008). After a clear specification of relevant elements of PICO, GRADE proceeds to the classification of important outcomes: ideally, all important outcomes need to be identified and rated (Guyatt, Oxman, Kunz, Atkins, et al., 2011). If a systematic review is conducted with a purpose to inform a particular guideline, the GRADE approach suggests to classify the outcomes into the categories of *critical*, outcomes that are *important but not critical* and those of *limited importance* (Guyatt, Oxman, Kunz, Atkins, et al., 2011). The first two categories of outcomes are later used in the EtD frameworks to inform judgments about the strength of recommendations when developing practice guidelines.

In the context of systematic reviewing, the GRADE guidelines define "certainty of evidence" ("quality" has also been used in the previous GRADE publications) as the

extent of confidence that the estimate of effect for a specific outcome is correct (Balshem et al., 2011). Recently, however, this definition has been clarified by the GRADE Working Group as the extent to which reviewers can be confident that *"the true effect for a specific outcome lies on one side of a specified threshold or within a chosen range"* (Hultcrantz et al., 2017). The revised guidance suggests three types of certainty of evidence ratings: noncontextualised, partially contextualised and fully contextualised, which mostly applies to guideline development (see Table 1.2 below and Chapter 4 for more details). In this new conceptualisation, the certainty of evidence ratings are explicitly acknowledged to be contingent upon a priori defined thresholds of what may be considered as meaningful effects in different contexts. These thresholds and the resultant ratings may, therefore, vary depending on the context and purpose of the review (Hultcrantz et al., 2017).

The key notion behind the GRADE ratings is that the observed estimates of effect cannot be sufficient by themselves to make causal claims about the intervention. It is imperative to know how these estimates have been generated to make a judgment about the trustworthiness of the obtained results (Guyatt et al., 2008). In scientific literature, this is often referred to as the *internal validity* of evidence, that is to say, the extent to which the results of an investigation are free from bias and can be truly attributed to the intervention under evaluation (Deeks et al., 2003). By comparison, the *external validity* reflects the extent to which an investigation's findings are generalisable to a target population beyond the sample included in the study (Deeks et al., 2003). Multiple examples in clinical medicine have shown how the lack of the validity considerations has resulted in misleading recommendations, such as those encouraging postmenopausal women to use hormone replacement therapy to decrease the

cardiovascular risk, when in fact, high certainty evidence has shown that it fails to do so and may even increase the risk (Guyatt et al., 2008). GRADE, therefore, suggests structured domains of evidence to assess the confidence in the obtained estimates of effect and, thereby, to provide support for causal claims made in systematic reviews. These domains are largely based on the "viewpoints" discussed by Sir Austin Bradford Hill to identify causal relations (Hill, 1965; Schunemann, Hill, Guyatt, Akl, & Ahmed, 2011) (see Box 1.1).

Table 1.2. Approaches to defining certainty of evidence in GRADE, adapted from (Hultcrantz et al. (2017)

Setting	Degree of contextualisation	Threshold or range	How to set	What certainty rating represents
Systematic reviews	Non-contexualised	Range: 95% CI	Using existing limits of the 95% CI	Certainty that the effect lies within the confidence interval
		OR≠1; RR≠1; HR≠1; RD≠0	Using the threshold of null effect	Certainty that the effect of one treatment differs from another
Systematic reviews	Partially contextualised	Specified magnitude of effect	E.g., small effect is the effect small enough to not use the intervention if adverse effects/costs are appreciable	Certainty in a specified magnitude of effect for one outcome (e.g., trivial, small, moderate or large)
Practice guidelines	Fully contextualised	Threshold determined with consideration of all critical outcomes	Considering the range of effects on all critical outcomes, and the values & preferences	Confidence that the direction of the net effect will not differ from one end of the certainty range to the other

Notes: CI = confidence interval; HR = hazard ratio; OR = odds ratio; RD = risk difference; RR = relative risk

Box 1.1. Bradford Hill's "viewpoints" on causation (Hill, 1965)

1. Strength of association: magnitude of association between exposure and outcome
2. Consistency: consistent results from different populations, circumstances and times
3. Specificity: specific outcomes related to specific exposures
4. Temporality: exposure temporality preceding outcome
5. Biological gradient: dose-response relationship between exposure and outcome
6. Biological plausibility: existence of models explaining the association of interest
7. Coherence: association is supported by the natural history of the disease
8. Experiment: experimental manipulation of exposure
9: Analogy: existing similar associations supporting the causal relationship

In the GRADE approach, the certainty of evidence is ultimately categorised into one of four ratings—high, moderate, low, or very low (Balshem et al., 2011). The process of rating the certainty of evidence for each outcome starts with a consideration of the designs of the included studies: if the body of evidence contributing to an outcome consists of randomised controlled trials (RCTs), the certainty of evidence is initially given a rating of "high", while a body of evidence consisting of nonrandomised studies (NRSs) is initially given a certainty of evidence rating of "low". The body of evidence is then assessed by considering eight further domains (see Table 1.3). Assessments within five domains—risk of bias, indirectness, inconsistency, imprecision and publication bias—are used to downgrade the initial certainty of evidence rating by one or two levels. For a body of evidence consisting of NRSs and initially rated as a "low" certainty, assessments within the three remaining domains—magnitude of the effect, dose-response relationship in the effect and counteracting plausible residual bias or confounding—may be used to upgrade the rating by one or two levels (Balshem et al., 2011). While these domains aim to provide a systematic and structured process for rating the certainty of evidence, it is important to note that the GRADE Working Group supports judicious use

of these domains as general guiding principles, rather than as prescribed criteria in a checklist (Balshem et al., 2011; Guyatt, Oxman, Akl, et al., 2011). Summary of Findings (SoFs) tables and Evidence Profiles are used to summarise the body of evidence contributing to and provide certainty of evidence ratings for each main outcome in a systematic review.

Table 1.3. Domains for rating the certainty of evidence, adapted from Guyatt et al. (2011)

Study design		Certainty of evidence	Lower if	Higher if
Randomised trial	→	**High**	**Risk of bias** - 1 Serious - 2 Very serious	**Large effect** + 1 Large + 2 Very large
		Moderate	**Inconsistency** - 1 Serious - 2 Very serious	**Dose response** +1 Evidence of a gradient
Nonrandomised study	→	**Low**	**Indirectness** - 1 Serious - 2 Very serious	**All plausible confounding** +1 Would reduce a demonstrated effect or +1 Would suggest a spurious effect when results show no effect
		Very low	**Imprecision** - 1 Serious - 2 Very serious **Publication bias** - 1 Serious - 2 Very serious	

Study design: As discussed above, evidence hierarchies have played an important role within the EBP model as a heuristic tool to distinguish between high and low quality research evidence on intervention effects. In line with the viewpoints of Bradford Hill, randomised controlled trials (RCTs) are prioritised in the EBP model, as they are able to provide grounds for causal inference through randomisation, thereby, formation of a comparable "counterfactual" and experimental manipulation of the intervention. The

latter also addresses the temporality of the relationship between the outcome and the intervention (see Box 1). While the GRADE approach endorses this hierarchy of evidence through initial categorisation of evidence based on study design, it allows flexibility in employing other considerations, such as risk of bias to challenge and possibly modify the initial certainty of evidence rating. This represents a more refined approach as compared to the preceding evidence rating systems, wherein judgments of the certainty of evidence are reduced to the study design (Ebell et al., 2004).

Risk of bias: GRADE defines the flaws in the design of RCTs or NRSs (i.e., the internal validity) as risk of bias or study limitations (Guyatt, Oxman, Vist, et al., 2011). While the GRADE guidelines outline the general domains of bias for both RCTs and NRSs, they predominantly follow those described in the Cochrane Risk of Bias (RoB) tools, such as allocation concealment, blinding of intervention providers and participants, incomplete accounting of participants and outcome events and selective outcome reporting (Higgins & Green, 2011).

Inconsistency: GRADE identifies inconsistency in evidence if the effect estimates differ widely across studies, resulting in high levels of heterogeneity. If there are no plausible explanations for the observed heterogeneity based on the characteristics of the population, intervention, outcome or study methods, the certainty of evidence is lowered (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al., 2011). GRADE guidelines describe four criteria for downgrading the certainty of evidence for inconsistency: point estimates vary widely across studies; confidence intervals show minimal or no overlap; the statistical test of heterogeneity shows low **P**-values, and the $\mathbf{I^2}$ is large. If plausible explanations, however, are derived from subgroup analyses or meta-regressions, review authors are advised to provide different estimates

across participant groups, interventions or outcomes (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al., 2011).

Imprecision: According to GRADE, imprecision in data can result from an inadequate sample size or few events (Guyatt, Oxman, Kunz, Brozek, et al., 2011). GRADE suggests a method for imprecision estimation that combines data on the width of the 95% confidence interval, inclusion/exclusion of the point of no effect in the 95% CI and comparison of the number of participants in the review, with the threshold of the "optimal information size" (OIS). The latter represents the number of participants for a single adequately powered trial (Guyatt, Oxman, Kunz, Brozek, et al., 2011).

Indirectness: Indirectness occurs due to lack of evidence on the important measures of interest, resulting in the use of surrogates and approximations (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Falck-Ytter, et al., 2011). GRADE differentiates between four sources of indirectness in systematic reviews: outcomes (when surrogate outcomes are the measures), comparisons (no head-to-head comparisons between the alternative interventions), populations and interventions. If the first two are particularly relevant to systematic reviews, the last two are primarily used in the context of guideline development (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Falck-Ytter, et al., 2011).

Publication bias: The GRADE approach highlights the need for investigation of publication bias in systematic reviews, which occurs as a result of inadequate reporting of studies, particularly those demonstrating no effect (Guyatt, Oxman, Montori, et al., 2011). In general, GRADE guidelines suggest rating down the certainty of evidence for the likelihood of publication bias, when the evidence consists of a number of small studies,

particularly, when they are industry sponsored. It also encourages using funnel plots and statistical tests of asymmetry (Guyatt, Oxman, Montori, et al., 2011).

Upgrading evidence: Although GRADE initially downgrades the certainty of evidence from NRSs, which in comparison to RCTs, are considered to be at higher risk of bias in estimating intervention effects, under several circumstances the ratings can be enhanced to the levels of moderate or high, provided that the evidence has not be downgrade for any of the five domains described above (Guyatt, Oxman, Sultan, et al., 2011). Those circumstances include: NRSs yield large or very large effect estimates, NRSs provide evidence with a dose-response relationship and presence of residual confounding that reduces the effect, or suggests a spurious effect when no effect is observed (Guyatt, Oxman, Sultan, et al., 2011).

Complex interventions: evolving perspectives

As the EBP model spread to other disciplines beyond biomedicine, such as areas of public health and social policy, so have reports of challenges and controversies as to whether this model is indeed suitable for social disciplines, where interventions are often described as "complex" (Craig et al., 2008; Petticrew, Anderson, et al., 2013; Rutter et al., 2017). Part of the scepticism is related to the rigid application of the EBP evidence hierarchy to policy decisions, which often require integration of a broad range of evidence, and consideration of multiple perspectives, values, and competing interests (Cairney & Oliver, 2017; Parkhurst & Abeysinghe, 2014; Sanderson, 2006; Smith, 2013). In this view, the EBP model has been contested for its biomedical roots, and, therefore, narrow rationale and instrumental approach towards evidence generation and use in policy-making. As argued by many researchers, interventions in public health and social policy are radically different from "biomedical" interventions, in that they commonly operate by psycho-social processes, such as cognitions, emotions, behaviours, norms, interpersonal relationships, environmental and system-level changes and use pathways to impact, which are complex and context-dependent (Craig et al., 2008; Grant, 2014). Variation in the design, implementation and effects of these interventions pose challenges to the use of conventional methods of EBP, such as RCTs and meta-analyses, which have predominantly been designed and employed for evaluating biomedical interventions with relatively homogeneous causal pathways and effects (Melendez-Torres, Bonell, & Thomas, 2015; Petticrew, 2015; Rutter et al., 2017; Tanner-Smith & Grant, 2018; Victora, Habicht, & Bryce, 2004).

Interest in the evaluation of complex interventions, as a distinctive perspective and category of interventions, has increased in the last decade. As summarised in Table 1.4, considerable number of initiatives have been launched since 2008 aiming to provide methodological guidance for evaluating and reviewing evidence of complex interventions, and several of these projects are still ongoing. While all of these initiatives use the same labelling of "complex interventions", they often differ in how they define complexity and the extent to which they engage with the concepts of complexity theory, that is, the scientific study of complex systems (Miller, 2007). A broad view has been taken in this section of the chapter of how conceptualisation of complex interventions has evolved in intervention research in the last decade mainly following the publication of the UK Medical Research Council (MRC) guidance for developing and evaluating complex interventions (Craig et al., 2008). The perspective to complex interventions supported in this thesis work is also outlined in view of the thesis aims and focus.

Table 1.4. Key initiatives providing guidance for assessing complex interventions

Initiative	Purpose	Definition of complexity	Key sources used	Disciplines targeted	Examples used
Medical Research Council (MRC) guidance (Campbell et al., 2000) (Craig et al., 2008) (Moore et al., 2015)	Guidance for developing and evaluation complex interventions	Five dimensions of complexity (see Table 1.5)	Hawe et al. (2004) Rychetnik et al. (2002)	Public health & social policy areas: education, transport, housing	Behaviour change programmes for people at risk of Type 2 diabetes
Journal of Clinical Epidemiology (JCE) series on considering complexity in systematic reviews of interventions (Anderson, Petticrew, et al., 2013)	Guidance for considering complexity in systematic reviews	As a set of properties determining the causal pathways between the intervention & the outcomes	Craig et al. (2008) Galea et al. (2010) Hawe et al. (2009)	Public health; social care	Slum upgrading; obesity prevention; day care for people with mental disorders
CONSORT for Social and Psychological Interventions (SPI) (Grant, 2014)	Guidance for reporting of complex intervention RCTs	Emphasis on malleable social and psychological mechanisms	Craig et al. (2008) Hawe et al (2004)	Public health; psychology; social work; criminology; education	Multi-systemic therapy for juveniles
CReDECI2 (Mohler, Kopke, & Meyer, 2015)	Guidance for reporting the development & evaluation of complex interventions	Three dimensions: - multiple components - different levels of target - flexibility/tailoring of the intervention	Craig et al. (2008) Michie et al., (2009)	Nursing; health psychology; education	Multifaceted interventions for chronic disease management; fall prevention in elderly
INTEGRATE-HTA (Gerhardus, 2016)	Guidance on the integrated assessment of complex health technologies	Five characteristics of complexity: - multiple perspectives - indeterminate phenomena - uncertain causality	Craig et al. (2008) Shiell et al. (2008)	Healthcare; public health	Palliative care

		- unpredictable outcomes - historicity, time and path dependence			
iCAT-SR (Lewin et al., 2017)	A tool for assessing the level of intervention complexity in systematic reviews	Ten dimensions of complexity (see Table 1.6)	Craig et al. (2008) JCE series (2013) Rogers (2008)	Healthcare; public health	Lay health worker interventions in primary and community care
AHRQ series on complex intervention systematic reviews: PRISMA for complex interventions (Guise, Chang, et al., 2017)	Guidance for the conduct and reporting of systematic reviews of complex interventions	Five characteristics / sources of complexity: - intervention - pathway - population - implementation - context	Craig et al. (2008) JCE series (2013) Kuhne et al. (2015)	Healthcare; public health	Preventive interventions for ulcers; slum upgrading; surgery; daily intake of aspirin for 5 years
Retrieval, synthesis and assessment of complex health interventions Ongoing (WHO, 2018)	Guidance for retrieval, synthesis and assessment of complex health interventions	Aspects of complexity (systems perspective): - multi-component - multiple outcomes - context-dependency - system adaptivity - emergent properties - non-linearity / phase changes - feedback loops	AHRQ series (2017) Craig et al. (2008) Diez Roux (2011) Hawe et al. (2009) JCE series (2013) Rutter et al. (2017)	Healthcare; public health; social policy areas	Health promotion interventions (e.g., sexual health education); public health legislation (e.g., Smokefree legislation); organisational (e.g., stroke units)

Notes: AHRQ = Agency for Healthcare Research and Quality; CONSORT = Consolidated Standards of Reporting Trials; CReDECI = Criteria for Reporting Development and Evaluation of Complex Interventions; iCAT-SR = intervention Complexity Assessment Tool for Systematic Reviews; INTEGRATE-HTA = Integrated Health Technology Assessment for Evaluating Complex Health Technologies; JCE = Journal of Clinical Epidemiology; RCTs = Randomised Controlled Trials; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Perspective 1: Complex interventions

The MRC guidance, initially published in 2000 and later revised in 2008, has provided the most influential framework for describing complex interventions (Craig et al., 2008; Datta & Petticrew, 2013). It uses the phrase "complex interventions" to refer to health service, public health and social policy interventions, including behavioural, educational, psychological, occupational and organisational interventions. Five dimensions of complexity are described in the MRC guidance (see Table 1.5): (1) the number of intervention components, (2) the number of behaviours required by those delivering or receiving the intervention, (3) the number of groups or organisational levels targeted by the intervention, (4) the variability of outcomes, and (5) the permitted level of flexibility or tailoring of the intervention.

Table 1.5. Defining complexity based on intervention characteristics (Craig et al. (2008)

Dimensions of complexity
1. Number of and interactions between intervention components
2. Number & difficulty of behaviours required by those delivering or receiving the intervention
3. Number of groups or organisational levels targeted by the intervention
4. Number and variability of outcomes
5. Degree of flexibility or tailoring of the intervention permitted

From this perspective, the "Positive Parenting Program" (or Triple P) can be considered an example of a complex intervention owing to the number of interacting intervention components, the number of behaviours required by those delivering and receiving the interventions, the number of outcomes targeted, and the degree of flexibility of the implementation that is permitted (Sanders, Turner, & Markie-Dadds, 2002). Designed as a parenting and family support multi-level programme, Triple P aims to prevent problems in the family, school and community, as well as to treat behavioural and emotional problems in children and teenagers. It includes components of increasing

intensity delivered to parents of children up to 16 years and includes 5 levels of delivery. Within each level, there is a choice of delivery methods, which highlights the flexibility of the intervention to meet the needs of individual communities (Sanders et al., 2002).

The important presupposition of this perspective, and all other conceptualisations of complex interventions based on the MRC guidance, is that there is a definable "intervention", which can be specified as a set of professional behaviours, practices, or ways of organising a service, and which aims to be bring about identifiable outcomes. This set of behaviours and practices are linked by an explicit or implicit theory on how they operate to produce the outcomes (Petticrew, 2011). Since its publication, the MRC guidance has served as an important heuristic for operationalising complexity in the design and evaluation of interventions and is still one of the most frequently used references across methodological work on considering complexity in intervention research (see Table 1.4). For example, a newly developed Complexity Assessment Tool for Systematic Reviews (iCAT_SR; see Table 1.6) builds on the complexity dimensions from the MRC guidance to describe levels of intervention complexity in systematic reviews (Lewin et al., 2017). Despite this, arguments have also been raised that the definition of complexity described in the MRC guidance does not take the understanding of complexity far enough by focusing on interventions as "packages" of activities and locating sources of complexity only within the constituent parts of interventions themselves (Hawe et al., 2009). This criticism has mainly been raised by the proponents of the complex systems perspective; according to them, the MRC guidance suggests a view on complexity that should be framed as *"complicated"* rather than *"complex"* (Glouberman & Zimmerman, 2002; Shiell et al., 2008).

Table 1.6. The iCAT_SR dimensions of complexity, adapted from Lewin et al. (2017)

Core dimension of complexity	Assessment levels
1. Active components included in the intervention, in relation to the comparison	• More than 1 component and delivered as a bundle • More than 1 component • One component • Varies
2. Behaviours or actions of intervention recipients or participants to which the interventions is directed	• Multi-target • Dual target • Single target • Varies
3. Organisational levels and categories targeted by the intervention	• Multi-target • Dual target • Single target
4. The degree of tailoring intended or flexibility permitted across sites or individuals implementing the intervention	• High tailored/flexible • Moderately tailored/flexible • Inflexible • Varies
5. The level of skill required by those delivering the intervention in order to meet the intervention objectives	• High level skills • Intermediate level skills • Basic skills • Varies
6. The level of skill required for the targeted behaviour when entering the included studies by those receiving the interventions	• High level skills • Intermediate level skills • Basic skills • Varies
Optional dimension of complexity	**Assessment levels**
7. The degree of interaction between intervention components, including the independence/interdependence of intervention components	• High level interaction • Moderate level interaction • Independent • Varies • Unclear or unable to assess
8. The degree to which the effects of the intervention are dependent on the context or setting in which it is implemented	• Highly context dependent • Moderately context dependent • Independent of context • Varies • Unclear or unable to assess
9. The degree to which the effects of the intervention are changed by recipient or provider factors	• Highly dependent on individual-level factors • Moderately dependent on individual-level factors • Largely dependent on individual-level factors • Varies • Unclear or unable to assess
10. The nature of the causal pathway between the intervention and the outcome it is intended to effect	• Pathway variable, long • Pathway linear, long • Pathway linear, short • Varies • Unclear or unable to assess

Perspective 2: Complex interventions and their causal pathways

A distinction between *complicated* and *complex* aspects of interventions has been well articulated by Rogers (2008) drawing on examples of programme theories to evaluate these aspects. Rogers defines programme theory, also referred to in the literature as theory of change (Weiss, 1998), programme logic (Funnell, 1997), intervention logic (Nagarajan & Vanheukelen, 1997), and casual pathway (Douthwaite, Kuby, Fliert, & Schultz, 2003), as a way of developing a causal model linking intervention inputs and activities to a chain of intended or observed outcomes, and then using this model to guide the assessment of an intervention (Rogers, 2008). Rogers describes three aspects of interventions, which make their evaluation *complicated*. These include interventions implemented through multiple agencies, interventions with multiple simultaneous causal strands, and those with alternative causal strands. In order for an intervention to be denoted as *complex*, it is important that the intervention programme theory additionally describes a non-linear relationship, emergent outcomes and a degree of agency and interaction between those delivering and those receiving the intervention (Rogers, 2008). Non-linearity and emergence are key concepts from complexity science and have their roots in non-linear, dynamic dissipative systems theory (Kellert, 1993). In non-linear thermodynamics, dissipative systems are characterised as open systems capable of assimilating large amounts of energy from their environments and converting that energy into structural complexity (Prigogine, 1996). These processes of energy conversion can lead to rapid changes in the systems, which transform them into a qualitatively new state. The prediction of behaviours of these systems is, therefore, not straightforward. In terms of intervention evaluation, non-linearity would thus suggest recursive rather than unidirectional relationships in the intervention to outcome

pathway, where changes in behaviours may create conditions for behaviours to change further (e.g., healthy food availability promoting a healthier diet, which in turn, creates greater demand for healthy foods) (Patton, 1997). There may also be "tipping points" in the pathway of an intervention, where disproportionally large intervention effects appear suddenly followed by a long period of no effect, or as a result of achieving certain critical levels (Rogers, 2008). The notion of emergence is another challenging aspect of complex systems as it suggests that the specific outcomes produced by an intervention emerge during the implementation of the intervention. Emergent outcomes, therefore, may require an emergent logic model, that is, a flexible programme theory expected to evolve with the implementation of the intervention (Rogers, 2008).

While the distinction between complicated and complex aspects of an intervention can serve as a useful framework to guide intervention evaluation (Glouberman & Zimmerman, 2002), it is important to note that clear boundaries between complicated and complex interventions are hard to establish. In fact, many researchers of complex interventions argue for a pragmatic approach that looks at a spectrum of complexity in the aspects of the intervention evaluation, rather than classifying the interventions themselves into the categories of "simple", "complicated" and "complex" (Craig et al., 2008; Lewin et al., 2017; Petticrew, 2011; Rogers, 2008). Figure 1.4 presents an example of a logic model for broadly defined community capacity building programmes, including complicated, as well as complex aspects. While highlighting the possible causal pathways and feedback loops, it describes a generic model of capacity building programmes that can be used across a wide range of initiatives, including community leadership development, family strengthening and volunteer training. Because of the multiple and alternative causal pathways, it does not

provide a priori defined outcomes, but instead, supports consideration of emergent outcomes as the implementation unfolds capitalising on opportunities in each specific context.

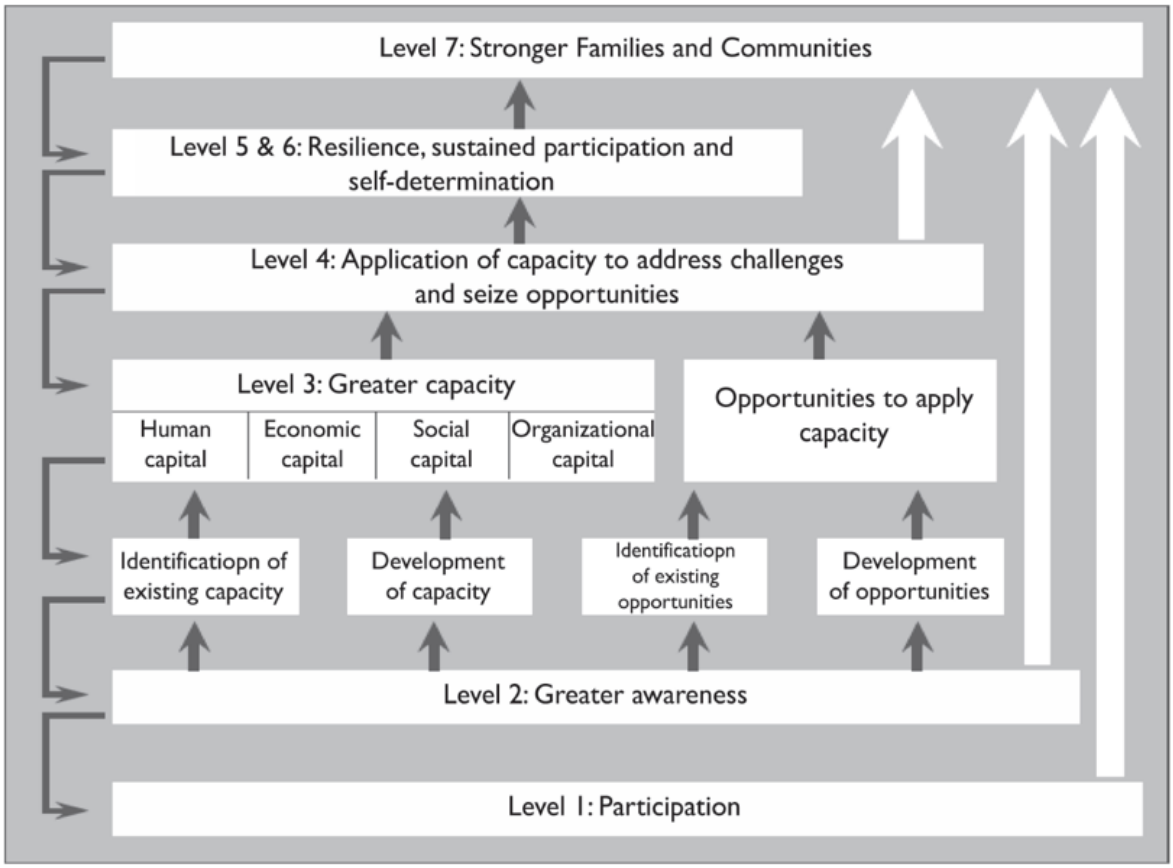


Figure 1.4. Logic model for a community capacity building programme with complicated and complex aspects (CIRCLE, 2006; Rogers, 2008)

The emphasis on the characteristics of interventions' causal pathway in addition to the characteristics of the interventions themselves can be viewed as an extended perspective on describing complex interventions. Drawing on the dimensions of the MRC guidance, this perspective attempts to engage more with the concepts from the complex systems approach, such as non-linearity of relationships and emergent properties. The series of methodological articles published in 2013 in the *Journal of Clinical Epidemiology*

on considering complexity in systematic reviews played an important role in advancing this perspective (Anderson, Petticrew, et al., 2013). In this series, complexity is characterised as *"a set of properties of the causal relationship between an intervention and its outcomes, that is, an array of potential influences on the direction and magnitude of intervention effect derived from variant properties of the systems in which interventions are introduced"* (Anderson, Petticrew, et al., 2013). The effects of an intervention are, therefore, likely to be contingent on different characteristics of (1) the intervention itself and (2) the intervention's causal pathway (see Table 1.7) (Petticrew, Anderson, et al., 2013). While this perspective on complex interventions still largely follows the MRC guidance, discussed above, in conceptualising interventions as discrete and bounded activities, by focusing on the characteristics of the causal pathway, it, however, allows for a description of an intervention that is more dynamic: the relationship between an intervention and its outcomes is seen to be modifiable by the variant characteristics of the context, the implementation process and the system (Anderson, Petticrew, et al., 2013; Noyes et al., 2013).

Realist evaluation can also be viewed to follow this perspective through its emphasis on the role of Context, Mechanisms and Outcomes (CMO), frequently referred to as CMO pattern configurations, for explaining and understanding interventions (Pawson, 2006). A CMO configuration in realist evaluation and realist synthesis is a proposition that aims to identify what it is about the intervention that works, for whom and in what circumstances. Realist analysis, therefore, aims to provide not only evidence of intervention effectiveness, but also an explanation of why the outcomes occurred as they did to help develop and improve the content and targeting of future interventions (Linsley, Howard, & Owen, 2015; Wong, Greenhalgh, Westhorp, Buckingham, & Pawson,

2013). Similar definition and perspectives to operationalising complexity, however, following a more conventional approach to systematic reviewing, have been adopted in the newly developed reporting guidance for systematic reviews and meta-analyses of complex interventions (PRISMA extension for complex interventions) (Guise, Butler, et al., 2017b). The PRISMA-CI guidance outlines five key sources of complexity in systematic reviews: intervention complexity, pathway complexity, population complexity, implementation complexity, and contextual complexity.

Table 1.7. Sources of complexity in systematic reviews, adapted from Petticrew et al. (2013)

1. Characteristics of the intervention itself
Number of components
Number of groups or organisational levels targeted
Degree of flexibility or tailoring permitted
Self-organisation, adaptivity and changes over time
2. Characteristics of the intervention's causal pathway
Number and variability of outcomes
Non-linear relationships and phase changes
Number of mediators and moderators of effect
Positive/negative feedback loops
Synergies/dysynergies between components
Interaction with context

Perspective 3: Complexity theory and systems thinking

In parallel to the perspectives discussed above, a third perspective on conceptualising complexity in intervention research should be differentiated, namely the complex systems perspective, which is often referred to as "systems thinking" (Meadows, 2008; Petticrew et al., 2017; Rutter et al., 2017). Compared to other domains of practice, such as economics, climate change and urban science, the systems

perspective has had less prominence in health sciences (Leischow, Best, & Trochim, 2008); however, interest in it has been growing in the recent decade. A distinctive characteristic of this perspective is that it promotes a shift in thinking of change-processes in human populations by drawing heavily on social networks and complex adaptive systems theory (Shiell et al., 2008). Contrary to the conventional thinking about interventions as "packages" of activities, systems perspective suggests to view interventions as "events in systems" putting more emphasis on the need to explore the dynamic properties of the contexts into which the intervention is introduced (Hawe et al., 2009). This perspective, thereby, aims to move away from conceptualising interventions as discrete set of activities, to viewing interventions as embedded in wider systems (Hawe et al., 2009; Shiell et al., 2008). From this perspective, interventions can be defined as *"time-limited series of events, new activity settings and technologies that have the potential to transform the system, because of their interaction with context and the capability created from this interaction"* (Hawe et al., 2009). Hawe made an important contribution to advancing systems thinking in public health research. In the paper on "theorising interventions as events in systems" (2009), Hawe suggests to think of schools, communities and worksites as complex ecological systems, which can be theorised on three dimensions: (1) their constituent activity settings (e.g., classrooms); (2) the social networks that connect the people and the settings; and (3) time (Hawe et al., 2009). Intervention effects in these systems are seen to result from dynamic processes, such as interactions among different actors, rather than from direct uptake of the intervention components by people or units in the system. By way of illustration, in order to understand the impact of health promotion interventions in high schools, Hawe & Ghali (2008) suggest conducting a social network analysis to explore important people

to get “on side with an intervention”, as well as to understand how the intervention itself might change the social structure in the school/classroom setting. In a similar vein, research has shown how banning smoking in public places in the UK resulted not only in changes in health outcomes, but also brought about changes in the wider social system, including patterns of drinking and socialising in the community (Gruer, Tursan d'Espaignet, Haw, Fernandez, & Mackay, 2012; Petticrew et al., 2017). Other methods, such as complex intervention modelling have also been proposed to help understand intervention impacts in dynamic complex systems (Greenwood-Lee, Hawe, Nettel-Aguirre, Shiell, & Marshall, 2016).

In comparison with the two previous perspectives discussed above, the systems perspective can be viewed as an approach that best attends to the principles of complexity theory, that is, an accretion of anti-reductionist scientific ideas on understanding the behaviours of systems as more than the sum their constituent parts (Reed & Harvey, 1992). While there are many different definitions of complex systems, the key scientific idea associated with the study of complex systems, as argued by Thrift, is the idea of the science of *holistic emergent order* in complex and unpredictable phenomena. This is a more open science, which prioritises *processes* over variables, *relationships* over entities, and *dynamics* over structures (Thrift, 1999). Box 1.2 attempts to summarise the key principles of the systems perspective (systems thinking) more broadly, and its implications for studying social systems more specifically. In general, a complex adaptive system can be described as a collection of intelligent agents whose actions are interconnected, so that one agent’s actions change the context for the other (Plsek & Greenhalgh, 2001). Agents can simultaneously be members of several systems, and they respond to their environment by using internalised rule sets that drive action.

Systems, on the other hand, are embedded within other systems, co-evolve and adapt their behaviour over time. The frequently described characteristics of complex adaptive systems, therefore, include multiple interactions, non-linearity of relationships, adaptation, emergence, and feedback loops (see Table 1.8) (Plsek & Greenhalgh, 2001). Although systems perspective, currently, largely remains a theoretical approach, it is worth highlighting that there is evolving methodological research aiming to identify practical applications of this perspective in intervention research both at the levels of primary research and systematic reviewing of evidence (Diez Roux, 2011; Galea et al., 2010; Petticrew, 2015; WHO, 2018). Furthermore, even at the theoretical level, a systems perspective can provide important concepts, as outlined in Table 1.8, to guide the thinking of the researchers. By way of illustration, two recent methodological projects aiming to develop guidance on the assessment of complex interventions, specifically, the INTEGRADE-HTA project and the WHO project on retrieval, synthesis and assessment of complex health interventions (see Table 1.4) draw on the characteristics of complex adaptive systems as concepts to guide the planning, conduct and interpretation of a complex intervention assessment (Wahlster et al., 2016; WHO, 2018).

Box 1.2. Broad principles of systems thinking, adapted from Capra (2014)

1. Shift of perspective from the parts to the whole
Social systems are integrated wholes, whose properties and functioning cannot be reduced to its parts, or interventions within it. Systematic properties are destroyed, when a system is dissected into isolated elements, either physically or conceptually.

2. Inherent transdisciplinarity
All living systems, including social systems, such as a family, worksite, community, share a set of common properties and principles of organisation. Studying social systems, therefore, requires integration of academic disciplines. This corresponds to Mode 2 type of research and practice characterised by transdisciplinary, social accountability, contextualisation, reflexivity, heterogeneity, and organisational diversity (Gibbons, 1994).

3. From objects to relationships
Living systems are nested in each other, such as families are nested within communities, which are nested within societies. A part is always a pattern in a web of relationships. In systems perspective, therefore, objects are seen as networks of relationships. For the systems thinker, the relationships are primary, while the boundaries of the discernible patterns, the so-called "objects", are secondary.

4. From measuring to mapping
Relationships in systems cannot be measured and weighed following the traditional scientific methods. The shift from objects to relationships goes in line with a change of methodology from measuring to mapping. When relationships are mapped in social systems, patterns in the form of social networks, cycles and boundaries can be observed.

5. From quantities to qualities
Mapping relationships and studying patterns is not a quantitative but a qualitative approach. Thus, systems perspective implies a shift from quantities to qualities (for example, employing mathematics of visual patterns to map patterns in social networks).

6. From structures to processes
In the framework of Cartesian science, there are fundamental structures, and then, there are forces and mechanisms through which these interact and give rise to processes. In systems perspective, every structure is seen as the product of underlying processes.

7. From objective to epistemic science
In Cartesian science, scientific observations are believed to be objective, that is, independent of the observer and the process of knowing. Systems perspective, by contrast implies that epistemology, the understanding of the process of knowing, has to be incorporated into the description of a phenomenon. In other words, it involves a shift from "objective" to "epistemic" science, in which the method of questioning becomes an integral part of the scientific knowledge. In intervention research, this corresponds to what Sanderson calls a "practical rationality", which takes into account a plurality of forms of knowledge and relevant normative considerations in decision making (Sanderson, 2006).

8. From certainty to approximate knowledge

Systems perspective recognises that all scientific concepts, models and theories are limited and approximate.

Table 1.8. Characteristics of complex adaptive systems, adapted from Patton (2011) and Petticrew et al. (2013)

Characteristic	Definition	Example
Non-linearity	Sensitivity to initial conditions, where the effect, or the scale of the effect does not appear to be directly related to the cause; small actions can stimulate large reactions (Gleick, 1987; Taleb, 2007). This may also include phase transitions or tipping points, where the system moves quickly from one state to a qualitatively different state, or where the effects appear (or disappear) suddenly as a result of achieving certain critical levels/thresholds.	Herd protection from sanitation interventions: community sanitation first need to reach thresholds in the order of 60% or higher, to optimise health and nutrition gains.
Emergence	Effects emerge from self-organisation among interacting agents. Agents in the system interact and effects emerge "unintentionally".	Herd immunity from vaccination
Adaptation	A process of change that results from interacting agents responding and adapting to each other.	The banning of smoking in public places may affect individual consumption; manufacturers may reformulate tobacco products
Feedback loops	Changes in the system create conditions for further changes.	Availability of healthy foods affects healthy diets, which further create a need for healthy foods (positive feedback)
Interactions	Synergistic or dysynergistic relationships among agents in the system, as well as parts of the system, such as intervention components. In synergistic interactions of intervention components, effects of the intervention are more than the sum of effects of the individual components; in dysynergistic interactions, intervention effects are less than the sum of the individual components.	Community capacity building interventions to improve the infrastructure of a community provide services and programmes that may vary between contexts, but the goals and theories are the same across settings

Perspective supported in the thesis

The above discussion presents a brief and necessarily selective account of how approaches to conceptualising complex interventions have evolved in the last ten years, mostly in health sciences. Three main perspectives on complexity in intervention research have been outlined, namely, "complex interventions", "complex interventions and their causal pathways" and "complex systems". These perspectives differ in the extent to which they incorporate principles of systems thinking (see Box 1.2) and in how they envisage interventions, such as "packages" of activities introduced into a fixed context from outside or as "events within the dynamic systems" (Hawe et al., 2009). These differences have important implications on the planning and evaluation of interventions. For example, taking a "complex systems perspective" would necessarily entail broadening the scope of an evaluation and, in addition to assessing the effectiveness of an intervention with regard to a subset of outcomes, also examining the evidence on how the wider system changes and adapts when an intervention is introduced (Greenwood-Lee et al., 2016; Hawe et al., 2009). As in the example of understanding the impact of health promotion interventions in schools presented above, this would include collecting qualitative evidence on social norms and on how students and teachers socialise in schools, as well as the dynamics of the entire school network.

Considering, however, the focus of this DPhil thesis on the GRADE approach as a framework for rating the certainty of evidence in the *estimates of effect* of an intervention, that is, a clear focus on the evidence of intervention effectiveness, the most relevant perspective for this thesis towards conceptualising complex interventions would be the Perspective 2 discussed above, namely, *"complex interventions and their causal pathways"* (see Table 1.7). While this perspective largely follows the dimensions of

complexity from the MRC guidance, it also extends the definition to incorporate concepts from complex systems perspective, such as non-linearity of the causal pathway and context-specificity of the effects. For this thesis, complex interventions can thus be conceptualised as interventions that involve multiple interacting components and different levels of target (e.g., effects at the individual level, family level, the community level, the societal level); furthermore, effects of these interventions are amenable to contextual factors and processes of implementation. Because of these characteristics the causal pathways (also referred to as programme theories, above) of these interventions may be longer and may involve non-linear relationships, phase transitions and feedback loops, where the intervention inputs may not be linearly linked to its outputs and where changes in the outcomes may further affect the intervention (Craig et al., 2008; Galea et al., 2010; Noyes et al., 2013; Petticrew, Anderson, et al., 2013). It is important to note here that this conceptualisation of complex interventions may apply to interventions with different mechanisms of action. In this regard, complex interventions can be both "biomedical", that is to say, those that operate through physiological processes, as well as "social", that is, those that operate through malleable social and psychological processes, including cognitions, emotions, behaviours, norms, interpersonal relationships and social environments (Grant, 2014).

It is also important to acknowledge that conceptualising interventions and their causal pathways as "complex" is a pragmatic decision to be made and a perspective to be adopted by a researcher in each specific case, and should be led by users' needs. In many instances, such as when assessing effectiveness of a drug with regard to certain outcomes, it may be most informative and practical to conceptualise the intervention as "simpler" in terms of its causal pathway, number of outcomes and interactions with the

context. In a number of social interventions, however, such as Smokefree legislation, a more complex perspective might be needed to appropriately describe and assess the effects of the intervention at individual, as well as population levels, so as not to produce simplistic and potentially misleading conclusions (Petticrew et al., 2017).

While complexity as a perspective to intervention evaluation has more frequently been discussed in the context of social interventions, it may, nonetheless, be equally applicable to biomedical interventions, if deemed useful. Finally, it is also worth noting that no strict boundaries can be drawn to distinguish "complex" interventions from "simple" interventions; but rather, this thesis supports a view of complexity as a continuum of ways of looking at interventions. In this regard, the broad sources of complexity outlined in Table 1.7 can serve as heuristics to help distinguish aspects of interventions, their causal pathways, and the systems which may require more complex description and interpretation. Not all dimensions of complexity will be relevant to every evaluation and systematic review. This approach to conceptualising complexity through a list of dimensions, aspects or sources of complexity has also been supported by the many methodological projects in the last decade aiming to develop guidance for assessing complex interventions (see Table 1.4).

Implications of complexity for systematic reviews and practice guidelines

As noted above, a number of methodological initiatives have been launched in the recent years, which discuss the implications of complexity for systematic reviews and practice guidelines. While these projects may have differences in how they conceptualise complexity (see Table 1.4), they have common support for a holistic approach to systematic reviewing and guideline development (Anderson, Petticrew, et al., 2013; Guise, Chang, et al., 2017; Petticrew, 2015). This frequently entails consideration of a wide range of questions and integration of different sources and types of evidence to inform a contextualised decision-making (Greenhalgh, 2013; Rutter et al., 2017). An illustration of this approach in practice is the model of health decision-making supported by the NICE Centre for Public Health Excellence (see Figure 1.5). Similar to assessing intervention efficacy in tightly controlled conditions versus intervention effectiveness in a real-world setting (Haynes, 1999), this approach aims to bring together different types of *context-free scientific evidence* (i.e., knowledge produced from a scientific framework, which assumes existence of an objective reality independent of the observer and context) and *context-sensitive scientific evidence* (i.e., knowledge produced in specific real-life circumstances) in an integrated model for making practical recommendations in specific contexts and circumstances (Davies, 2005; NICE, 2012). In this broad perspective, however, use of conventional procedures and methods of reviewing evidence is often perceived challenging as they may disregard the important sources of complexity and, therefore, yield oversimplified conclusions (Kelly et al., 2017; Petticrew et al., 2017).

Reported challenges and implications of complexity for each step of systematic reviewing and guideline development as outlined in Figure 1.3 are further discussed below.

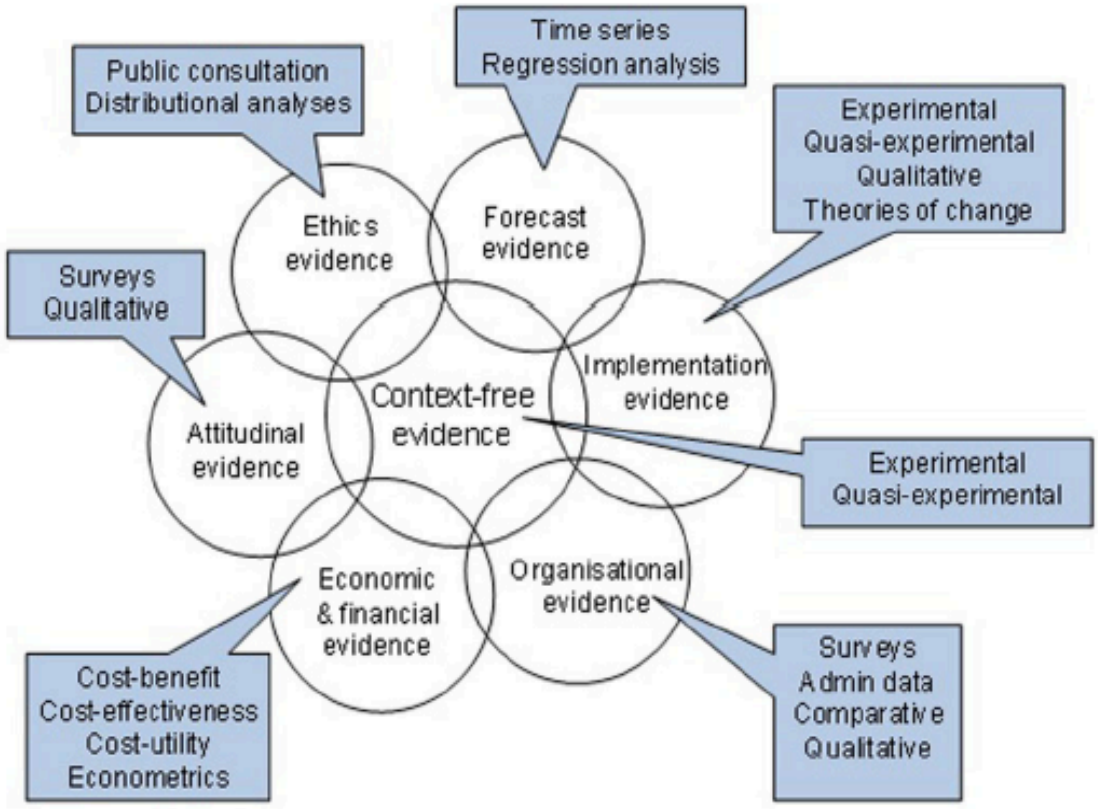


Figure 1.5. Scientific model of evidence integration for contextualised health decision-making, adapted from Davies (2005) and NICE (2012)

Framing systematic reviews of complex interventions

The first and most important decision in systematic reviewing or guideline development is to determine the focus and formulate key questions (Squires, Valentine, & Grimshaw, 2013). In clinical practice, detailed specification of a research question requires consideration of several key components (Richardson, Wilson, Nishikawa, & Hayward, 1995). The most commonly used approach for structuring questions, namely the "PICO" approach specifies the types of *Populations* (participants), *Interventions,*

Comparisons, and *Outcomes* that are of interest in a review or a guideline (O'Connor, Green, & Higgins, 2008). Framing reviews and guidelines based on PICO, therefore, implicitly privileges questions on intervention effectiveness and the types of evidence that are designed to answer them (i.e., RCTs). While PICO still serves as the most widely used approach to frame reviews and guidelines, it has come under criticism for failing to account for the mechanisms of action or causal pathways that mediate the outcomes and other contextual factors that may further moderate the outcomes (Anderson, Oliver, et al., 2013; Petticrew, 2015). As these interventions usually involve complex social change processes and hard-to-predict contextual interactions, it has been argued that simple *"does it work?"* review questions may yield uninformative findings (Petticrew et al., 2017). Instead, suggestions have been made to shift the aim of systematic reviews of complex interventions to assembling a range of evidence on *"what happens when interventions are implemented across different contexts, populations, and how do the effects come about?"* (Petticrew, 2015).

This shift in thinking about the aim and focus of reviews of complex interventions has prompted different proposals on how to frame systematic reviews (Butler et al., 2017; Squires et al., 2013). For example, many researchers have suggested to adapt PICO into PICOC or PICOTS to describe further sources of complexity by adding the dimensions of *Context*, *Time* and *Setting* (Butler et al., 2017; Kelly et al., 2017). Use of additional frameworks to better describe implementation and context are also widely advocated in systematic reviews of complex interventions, such as the Consolidated Framework for Implementation Research (CFIR) and the Context and Implementation of Complex Interventions (CICI) framework (Damschroder et al., 2009; Pfadenhauer et al., 2017). By way of illustration, the CICI framework provides a detailed description of the constructs

of context, implementation and setting that can be used to systematically identify and assess these dimensions in systematic reviews (see Table 1.9). Furthermore, because complex interventions in their design and implementation may involve multiple components and mechanisms of delivery, there are critical issues around specification of the interventions that review authors need to attend to (Squires et al., 2013).

Table 1.9. Dimensions and domains of the CICI framework, adapted from Pfadenhauer et al. (2017)

Dimension	Definition of Dimension	Domains
Context	Context refers to a set of characteristics and circumstances that consist of active and unique factors, within which the implementation is embedded. As such, context is not a backdrop for implementation, but interacts, influences, modifies and facilitates or constrains the intervention and its implementation. Context is usually considered in relation to an intervention, with which it actively interacts. It is an overarching concept, comprising not only a physical location but also roles, interactions and relationships at multiple levels.	- Geographical - Epidemiological - Socio-cultural - Socio-economic - Ethical - Legal - Political
Implementation	Implementation is an actively planned and deliberately initiated effort with the intention to bring a given intervention into policy and practice within a particular setting. These actions are undertaken by agents who either actively promote the use of the intervention or adopt the newly appraised practices. Usually, a structured implementation process consisting of specific implementation strategies is used and underpinned by an implementation theory.	- Implementation theory - Implementation process - Implementation strategies - Implementation agents - Implementation outcomes
Setting	Setting refers to the specific physical location, in which the intervention is put into practice and interacts with context and implementation.	

This involves the need to develop a working definition of the intervention of interest based on *pragmatic* (real-world) descriptions of intervention components as

opposed to using *theoretical* constructs (Michie et al., 2009; Squires et al., 2013). An example of this would be defining active learning in terms of interactive education sessions as opposed to problem-based learning, which is the theoretical construct behind active learning (Forsetlund et al., 2009). Finally, it is important that review authors distinguish between *prototypical* and *discretionary* intervention components. Prototypical components (or "active ingredients") are those that need to be present for the intervention to meet the working definition, while discretionary components may be present, but do not need to be present to meet the working definition of the intervention (Squires et al., 2013). The recently developed Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Complex Interventions (PRIMSA-CI) describes these components as essential and optional elements of complex interventions in more precisely reporting of I of PICOTS in systematic reviews (Guise, Butler, et al., 2017a; Guise, Butler, et al., 2017b). The items described in the Template for Intervention Description and Replication (TIDieR) checklist have also been proposed to guide comprehensive description of interventions in systematic reviews (Hoffmann et al., 2014).

Logic models have further gained popularity in reviews of complex interventions in the recent years. While there are different definitions of logic models, most frequently, they are defined as "a graphic description of a system identifying important elements and relationships within that system" (Kneale, Thomas, & Harris, 2015; Rehfuess et al., 2018). By describing the core components of an intervention, the connections between different components and outcomes, as well as contextual factors that may influence the outcomes, logic models serve as "frames" for systematic reviews and help to clarify the implicit or explicit theory of change for an intervention (Anderson et al., 2011). The

taxonomy of logic models recently developed by Rehfuess et al. (2017) differentiates between system-based and process-oriented logic models in systematic reviews: a system-based logic model attempts to unpick the complexity of an intervention and situates it within a broader context. It comprises a detailed description of the PICO elements as well as context and implementation elements; in this way, it depicts the system, in which interactions between participants, interventions, and the context takes place (Rohwer, Pfadenhauer, et al., 2017).

By comparison, a process-orientated logic model seeks to capture elements of the process within the intervention; it graphically displays the linear or non-linear causal pathways that lead from the intervention to its multiple outcomes (Rehfuess et al., 2018). System-based models can help to describe the entire system around a health or a social problem and, therefore, frame systematic reviews by suggesting important questions, such as the impact of the contextual complexity on the intervention and the nature of interaction between different intervention components. Process-oriented models, on the other hand, can have more instrumental use in terms of delineating the specific components, outcomes and contextual factors to consider in the review and the corresponding sources of evidence to search for. While these types of logic models differ in how they describe complexity, i.e., interventions embedded in complex systems (see Perspective 3 on complexity discussed above) and complex causal pathways (see Perspective 2 on complexity discussed above), they can be used in a complementary manner in reviews or guidelines. For example, in a review on e-learning of evidence-based healthcare (EBHC), authors first used a system-based logic model to define different intervention components, contextual factors and important interactions, i.e., to set the boundaries of the system around the problem and the intervention, and then

used a more-focused process-oriented model showing the pathway from e-learning to outcomes (Rohwer, Motaze, Rehfuess, & Young, 2017).

Another key debate around scoping systematic reviews of complex interventions is often referred to in the literature as “lumping” vs. “splitting” (Squires et al., 2013). While “splitters” argue that it is only appropriate to combine similar studies, i.e., studies using similar designs, participants, intervention characteristics and outcome measurements, “lumpers” argue that a systematic review aims to identify the common generalisable features within similar interventions, which warrants combining studies with minor differences. In their guidance on framing systematic reviews of complex interventions, Squires et al. (2013) support a lumping approach, as individual studies of complex interventions are almost always expected to vary in context, population and intervention components. As argued by the authors, the lumping approach reduces risk of bias and chance results by allowing assessment of the generalisability and consistency of research findings across a wide range of different populations, contexts and circumstances. This approach, thereby, might be better suited to answer broader research questions on which interventions (or components) work for whom and under what circumstances. Furthermore, lumping may also allow review authors to carry out explicit a priori subgroup analyses, thereby, enhancing the transparency of a systematic review. It is, however, crucial that the potential sources of heterogeneity are identified a priori to ensure that relevant data are extracted and to reduce the risk of data-driven analyses. An example of a lumping approach to review scoping can be found in the Cochrane review of audit and feedback interventions (Ivers et al., 2012). In this review, the authors adopted a lumping approach with explicit a priori subgroup analyses to explain variation in the results. Subgroup analyses answered secondary objectives of the

review on whether the effectiveness of audit and feedback interventions varied based on the format of feedback (e.g., verbal, written) or the frequency of feedback (e.g., weekly, monthly). Taking a broad lumping approach with a priori defined subgroup analyses may also be appropriate to avoid producing a series of empty reviews, especially when the evidence-base around a specific topic is expected to be scarce.

Finally, engaging stakeholders to better understand sources of complexity in each specific review and guideline has also been widely advocated in the recent years (Oliver & Dickson, 2016). While stakeholders may vary in their priorities and perspectives, involving people with expertise and those who would be the potential users of the evidence can help to refine and develop relevant review questions, as well as translate the results to a specific decision-making context (Guise et al., 2013; Kelly et al., 2017). For reviews of complex interventions, in particular, they can help to distil the sources of complexity of most relevance, therefore, save time and resources associated with the conduct of these reviews.

Synthesising evidence of complex interventions

Along with the need to broaden the scope of systematic reviews to account for complexity, a range of methods have been described to evidence synthesis beyond "traditional" meta-analysis for systematic reviews of complex interventions. These methods are broadly classified based on the purpose of evidence synthesis, such as to *generate*, *explore*, or *test* theories or some combination of these purposes (Anderson, Oliver, et al., 2013; Gough et al., 2012). Figure 1.6 provides an example of a typology of evidence synthesis methods in relation to the purpose of the synthesis. Methods such as statistical meta-analysis aim to test the effects of interventions and, therefore, need to

identify sufficient studies for unbiased *aggregation* of evidence; in the meantime, methods such as meta-ethnography or thematic synthesis aim to identify sufficient concepts for coherent *configuration* of new theories. Synthesis methods can also be located between these two extremes (i.e., configurative vs. aggregative) , such as framework synthesis that explores a range of questions, for example, those related to feasibility and acceptability of interventions (Anderson, Oliver, et al., 2013; Gough et al., 2012).

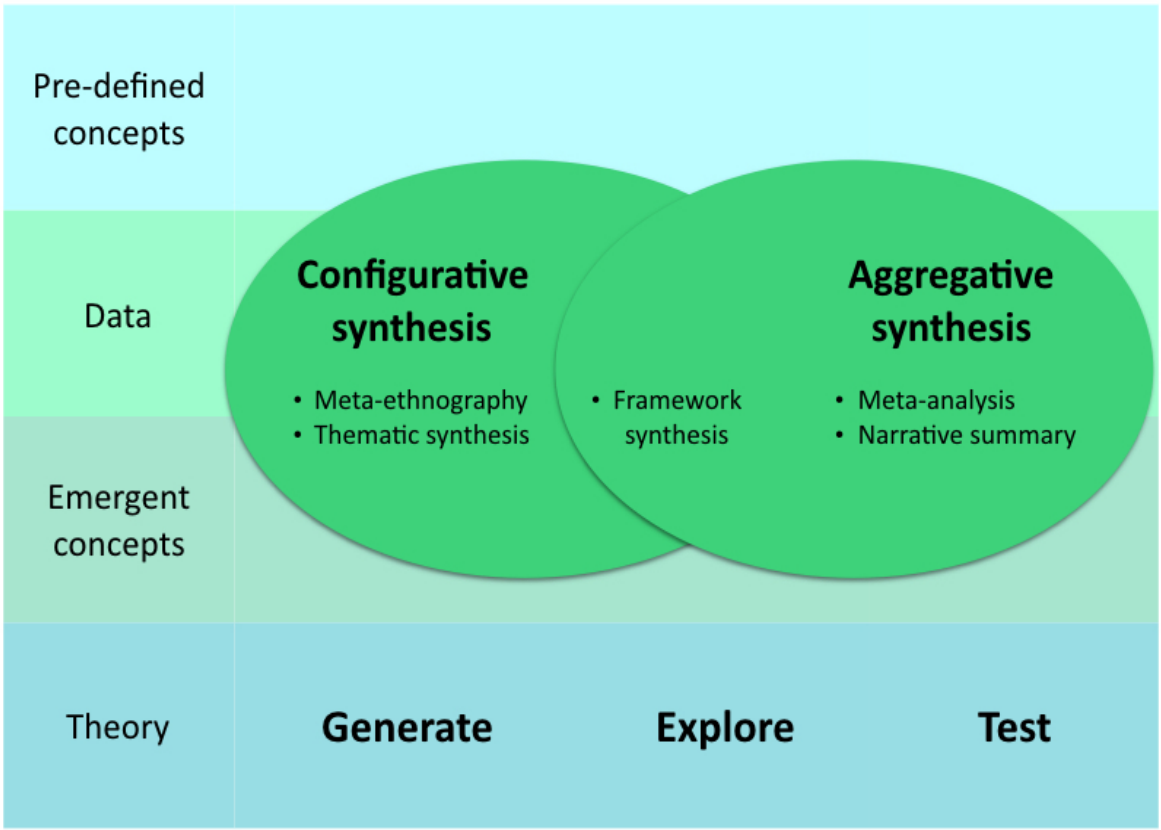


Figure 1.6. A typology of evidence synthesis methods based on the purpose of synthesis, adapted from Anderson, Oliver, et al. (2013) and Gough et al. (2012)

Typologies of evidence synthesis methods for complex interventions have also been described based on the types of data considered in the review, such as quantitative

(i.e., aiming to integrate quantitative evidence), qualitative (i.e., aiming to integrate qualitative evidence) and mixed-methods syntheses (aiming to integrate quantitative and qualitative evidence) (Petticrew, Rehfuess, et al., 2013). Moreover, within these broad categories there may be different approaches based on different research traditions and aims, such as Bayesian synthesis and realist synthesis within the category of mixed-methods synthesis. Use of this spectrum of methods and approaches in reviews of complex interventions allows answering broader sets of research questions beyond intervention effectiveness. For example, supplementing an aggregative synthesis of quantitative evidence that tests a theory on intervention effects with findings of a configurative qualitative synthesis that generates a theory (e.g., through "trial-sibling" or "unrelated" qualitative studies (Noyes et al., 2016)) may help to identify the "active ingredients" of multicomponent interventions, as well as provide further insights on why and how interventions work (Anderson, Oliver, et al., 2013; Thomas et al., 2004). This can be conducted within the framework of a single review or as two or more parallel reviews (Petticrew, Rehfuess, et al., 2013).

As analytic methods for synthesising evidence on complex interventions continue to evolve, an emphasis is put on the review questions as important guides for choosing the appropriate techniques (Tanner-Smith & Grant, 2018; Viswanathan et al., 2017). By way of illustration, in the recent AHRQ series on complex intervention systematic reviews, Pigott et al. (2017) suggest another typology of meta-analytic techniques for evidence synthesis based on the review question. Four review questions are described in this typology and analytic techniques are mapped accordingly: (1) how effective is the intervention? (2) For whom is the intervention effective and in what contexts? (3) What happens when the complex intervention is implemented? (4) What decisions are possible

given the results of the synthesis? An additional question is outlined in a similar typology by Tanner-Smith & Grant (2017) for understanding interventions' mechanisms of change (see Table 1.10).

Table 1.10. A typology of meta-analytic techniques for systematic reviews based on the review question, adapted from Pigott et al. (2017) and Tanner-Smith & Grant (2018)

Question	Analytic technique
1. How effective is the intervention?	• Random and fixed-effects meta-analysis • Network meta-analysis • Single subject studies meta-analysis • Multi-variate meta-analysis (using methods, such as multi-level modelling) • Robust variance estimation
2. For whom is the intervention effective and in what contexts?	• Sub-group analysis • Meta-regression - Frequentist - Bayesian Finite mixture models • Qualitative comparative analysis
3. What happens when the complex intervention is implemented?	• Choice of a wide range of approaches and methodologies
4. What decisions are possible given the results of the synthesis?	• Decision analysis
5. How does the intervention work?	• Path analysis and meta-analytic structural equation modelling

Analytic techniques have also been discussed for understanding the effects of individual components or their combinations in multi-component interventions. For example, meta-regressions can be used to create covariates for particular features of interventions, such as those related to the intensity of the intervention (as described by Belle et al., 2003). In addition to using meta-regressions to understand the effects of intervention components, Melendez-Torres et al. (2015) describe an alternative components-based approach, whereby the combination of components (based on similar

strategies or theoretical functions) are treated as their own "class" of interventions and are further analysed using network meta-analysis (as described in the review by Cooper et al. (2012) on interventions to promote uptake of smoke alarms). While very robust, these techniques require a sufficient number of studies with the same intervention to enable appropriate analyses. In this view, the utility of integrating qualitative review evidence with evidence on the effects of an intervention is increasingly discussed (Candy, King, Jones, & Oliver, 2013). Another approach is the Qualitative Comparative Analysis (QCA), which uses tabulation of evidence to identify those intervention components or combinations which are necessary and/or sufficient for a desired outcome (Thomas, O'Mara-Eves, & Brunton, 2014).

Rating the certainty of evidence of complex interventions

Challenges have also been reported in using the established approaches for assessing and rating evidence in systematic reviews of complex interventions. These challenges commonly relate to privileging evidence from randomised controlled trials by these methods and inadequate consideration of the rigour of specific types of nonrandomised studies, which are more frequently used to evaluate complex interventions (Bagshaw & Bellomo, 2008; Harder et al., 2015; Movsisyan et al., 2016b). Issues around how best to conceptualise the construct of "certainty" of evidence for complex interventions have also been raised (Rehfuess & Akl, 2013). Literature on these challenges, and specifically, those related to the use of the GRADE approach in reviews of complex interventions is further discussed below, which includes two publications by the DPhil candidate in the *Journal of Clinical Epidemiology*.

The first comprehensive study on the experiences of GRADE use in reviews and guidelines of public health interventions was conducted by Rehfuess & Akl (2013) . The authors conducted 18 interviews with individuals and groups (mostly within WHO) that have applied the GRADE approach in the field of public health. While all participants appreciated the systematic and transparent process of assessing the certainty of evidence, a range of challenges were reported relating to the complexity of public health interventions (see Table 1.11). Suggestions were made to develop guidance on applying GRADE to complex interventions and modify the current GRADE rating scheme (Rehfuess & Akl, 2013). The reported challenges and suggestions were classified as minor, such as those related to interpretation of the GRADE domains and aspects of GRADE implementation, and major, such as those related to introduction of new GRADE domains and use of observational evidence. For example, one of the major challenges reported in this study relates to the interpretation of the GRADE construct of "certainty evidence", that is ,confidence that the estimates of effect are correct, when the effects of complex interventions are contingent upon intervention programming, implementation and contextual factors. An alternative approach is suggested to conceptualise certainty of evidence for complex interventions as *"confidence that the effect is meaningful across a range of plausible implementation contexts"* (Rehfuess & Akl, 2013).

Table 1.11. Challenges of GRADE use in complex public health interventions as reported in Rehfuess & Akl (2013)

Challenge	Description	Type of challenge
Multi-component interventions	"Guideline developers and systematic reviews developing a PICO question for a specific intervention need to either consider the intervention as a whole or focus on a presumed active component. Either decision present challenges regarding which studies to include or exclude, how to interpret heterogeneity, and how to make careful judgment about the degree of indirectness."	Minor
Choice of outcome measures	"Reviews of public health interventions use (i) multiple outcomes for most interventions, (ii) outcomes at individual and group levels, (iii) reliance on short-term surrogate outcomes (iv) inconsistency associated with the use of competing measures or assessment scales, and (v) the need to group heterogeneous measures under 'umbrella outcomes' to make the review policy-relevant. The choice of outcomes and outcome measure has implications for indirectness of evidence and needs to be decided and judged carefully as part of GRADE."	Minor
Ability to discriminate between different types of observational studies	"With the quality of evidence for all types of nonrandomised studies starting as low, the GRADE approach is perceived as lacking the ability to distinguish between those public health interventions that are reasonably well-supported by evidence (e.g. by interrupted time series studies) and those that are less supported by evidence (e.g. by cohort studies). This may lead to misinterpretations of the evidence when communicating the message to policy-makers, and may even discourage the conduct of 'best possible' studies."	Major
Use of non-epidemiological evidence	"Judging the effectiveness of a public health intervention sometimes relies on sources of evidence outside of epidemiology. Physiological, physical or engineering principles and the insights gained through laboratory or animal studies can only be brought into the rating exercise as a separate very low-quality piece of evidence rather than lending additional credibility to epidemiological evidence."	Major
GRADE terminology	"Several groups stated that some of the GRADE terminology and definitions were not appropriate for public health interventions (e.g. definition of quality of evidence, use of the terms patients and clinicians, observational studies)."	Minor
Meaning of confidence in the evidence	GRADE defines quality of evidence as confidence in the pooled effect estimate. As the effectiveness of public health interventions is critically influenced by modes of delivery and contextual issues, a more relevant interpretation would be 'confidence that the effect is meaningful across a range of plausible implementation contexts'."	Major
Selection of the appropriate body of evidence for rating	"Several groups stated that it is challenging to grade the quality of evidence, in the presence of both a single or few randomised studies (rated as high quality, not downgraded for inconsistency or reporting bias) and a number of nonrandomised studies (rated as low quality, potentially downgraded for inconsistency and reporting bias)."	Minor
Insufficient possibilities to upgrade evidence	"Different groups suggested a broader interpretation of existing criteria, such as extending the concept of dose–response to the population level, and the addition of new criteria for upgrading, in particular consistency and analogy."	Major

Similar challenges were reported in an investigation of GRADE use in Cochrane systematic reviews led by the DPhil candidate (Movsisyan et al., 2016a, 2016b). In this study, 40 systematic reviews published in 3 Cochrane Review Groups from 2013 to May 2014, specifically the Cochrane Developmental, Psychosocial and Learning Problems Group (CDPLPG), the Cochrane Public Health Group (CPHG), and the Cochrane Depression, Anxiety, and Neurosis Group (CCDAN) were coded and classified in "complex" (n=24) and "simple" (n=16) intervention review groups. This classification was based on the predefined sources of complexity from the available literature at the time mapped into the PICOTS framework. Data on the GRADE ratings were then extracted and analysed in these two groups of reviews to help identify specific patterns of the GRADE ratings in reviews of complex interventions. This study found that outcomes of complex intervention reviews had higher proportions of "very low" certainty of evidence ratings compared with those of simple intervention reviews (37.5% vs. 9.1% for the primary benefit outcomes) and were more frequently downgraded for inconsistency, performance bias (as part of GRADE risk of bias assessment), and study design. In the meantime, none of the outcomes of complex intervention reviews (0%) were given "high" GRADE ratings (Movsisyan et al., 2016a).

Contacts established with 19 review authors further helped to interpret these findings and explore user perspectives on applying GRADE to systematic reviews of complex interventions (Movsisyan et al., 2016b). Review authors reported specific challenges in applying GRADE to reviews of complex interventions. Consistent with the findings of Rehfuess & Akl (2013), these challenges predominantly related to rating of nonrandomised studies and assessment of performance bias in GRADE. As argued by many researchers, population-level interventions often cannot be studies with RCTs

because of either ethical, practical and political considerations (Craig, Katikireddi, Leyland, & Popham, 2017; Rutter et al., 2017). Review authors, therefore, felt that GRADE frequently downgraded the *"best evidence possible"* for many complex interventions because of its initial categorisation of evidence based on study design, and lack of differentiation between more rigorous nonrandomised (e.g., interrupted time series studies) and less rigorous observational evidence (e.g., cross-sectional studies). Meanwhile, authors expressed uncertainties on how to adequately assess inconsistency (heterogeneity) of evidence in GRADE. For example, flexibility in the intervention implementation is a frequently reported source of complexity in systematic reviews (see Table 1.7); while in the current GRADE guidance, this flexibility is regarded as an unwanted heterogeneity and, therefore, a potential reason to downgrade evidence, from a complex intervention perspective, this flexibility might be viewed as a legitimate source of variation raising the question on how much variability in the studies can be tolerated before making a decision to downgrade the certainty of evidence (Movsisyan et al., 2016b). In this view, authors felt the need for further guidance on how to address these specific considerations of complex interventions when doing GRADE ratings.

While it can be reasonably argued that because of the complexities of the evidence of many social and public health intervention, one may never have "high" confidence in the effects of these interventions, two main concerns should be discussed with regard to the frequently observed "low" and "very low" labelling of evidence. First, as demonstrated by Weiss yet in 1997 evidence may not always be used directly in policymaking, but rather most frequently it serves to provide an intellectual background of concepts, ideas and overall orientations (this is referred to as an "enlightenment model" of research use). In this view, Cairney and Oliver (2017) argue that successful

engagement in evidence-based policymaking requires pragmatism, which combines scientific evidence with governance principles, and persuasion to translate complex evidence into simple stories. Scientific evidence in the form of "stories" and "ideas" is, therefore, seen to have better chances to travel to policy (Smith, 2013). It should be noted that, in principle, this approach is akin to tenets of the EBP model, whereby scientific evidence needs to be integrated with professional expertise and normative judgments to inform decisions (see Figure 1.1 above, Haynes et al., 2002).

In this light, there are concerns that using evidence labels such as "low" and "very low" certainty may cause possible misinterpretation, and be potentially used to justify inaction by policymakers, who might not be in the position to understand the nuances of these ratings as might be expected by the scientific community (Cairney & Oliver, 2017). In fact, research on this topic demonstrates that GRADE certainty of evidence ratings are the only statistically significant factor associated with the strength of recommendations in practice guidelines (Djulbegovic, Kumar, Kaufman, Tobian, & Guyatt, 2015). Furthermore, a study assessing the extent to which GRADE ratings of evidence affect uptake of WHO recommendations in national guidelines shows that recommendations underpinned by higher certainty of evidence ratings are associated with greater uptake: among strong recommendations, the odds of uptake of recommendations based on "high" or "moderate" certainty of evidence was 2.0 (95% CI: 1.4–2.8) compared to those based on "low" or "very low" certainty of evidence (Nasser et al., 2015). A related concern of frequently observed low ratings of evidence for many complex interventions is the potential for "evaluative bias", that is to say, prioritisation of those interventions in research and policy which are more conducive to experimentation (Ogilvie, Egan, Hamilton, & Petticrew, 2005). A few researchers argue that this may be

one of the main reasons why system-level interventions have not gained much traction in areas of public health and social policy , as these interventions cannot be easily evaluated using RCTs (Parkhurst & Abeysinghe, 2014; Rutter et al., 2017; Smith, 2013).

As outlined above, the GRADE Working Group is a collaboration of researchers who actively work to advance the GRADE approach and promote evidence-based practice guidelines. For example, in response to the need for an advanced framework for rating confidence in the qualitative evidence synthesis findings, members of the GRADE Working Group have recently developed the Confidence in the Evidence from Reviews of Qualitative research (CERQual) approach, which largely builds on the GRADE domains of evidence (Lewin et al., 2018; Lewin et al., 2015). Similarly, in recent years, the Group has come to further acknowledge that—alongside the certainty of evidence on intervention effectiveness—several other factors of relevance to decision-making should be systematically considered. This has led to the development of the GRADE Evidence-to-Decision (EtD) frameworks (Alonso-Coello et al., 2016). As shown in Figure 1.3, these factors include: values and preferences (in relation to outcomes), balance of benefits and harms, resource implications, priority of the problem, equity and human rights, acceptability and feasibility. No guidance, however, has been published by the GRADE Working Group on how to consider complexity in systematic reviews and practice guidelines. In the meantime, as discussed above, many researchers have proposed that complex intervention reviews in public health and social policy may benefit from a tailored guidance on how to use GRADE in their specific areas. Furthermore, considering the increasing interest in complexity in the recent years, this guidance can serve as a timely addition to the ongoing efforts to develop methods for incorporating a complexity perspective in systematic reviews and practice guidelines (see Table 1.4).

References

Alonso-Coello, P., Schunemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., . . . GRADE Working Group (2016). GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ, 353*, i2016.

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. *J Clin Epidemiol, 66*(11), 1223-1229.

Anderson, L. M., Petticrew, M., Chandler, J., Grimshaw, J., Tugwell, P., O'Neill, J., . . . Shemilt, I. (2013). Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol, 66*(11), 1205-1208.

Anderson, L. M., Petticrew, M., Rehfuess, E., Armstrong, R., Ueffing, E., Baker, P., . . . Tugwell, P. (2011). Using logic models to capture complexity in systematic reviews. *Res Synth Methods, 2*(1), 33-42.

Bagshaw, S. M., & Bellomo, R. (2008). The need to reform our assessment of evidence from clinical trials: a commentary. *Philos Ethics Humanit Med, 3*, 23.

Balshem, H., Helfand, M., Schunemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., . . . Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol, 64*(4), 401-406.

Belle, S. H., Czaja, S. J., Schulz, R., Zhang, S., Burgio, L. D., Gitlin, L. N., . . . Investigators, R. (2003). Using a new taxonomy to combine the uncombinable: integrating results across diverse interventions. *Psychol Aging, 18*(3), 396-405.

Butler, M., Epstein, R. A., Totten, A., Whitlock, E. P., Ansari, M. T., Damschroder, L. J., . . . Guise, J. M. (2017). AHRQ series on complex intervention systematic reviews-paper 3: adapting frameworks to develop protocols. *J Clin Epidemiol, 90*, 19-27.

Cairney, P., & Oliver, K. (2017). Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy? *Health Res Policy Syst, 15*(1), 35.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ, 321*(7262), 694-696.

Candy, B., King, M., Jones, L., & Oliver, S. (2013). Using qualitative evidence on patients' views to help understand variation in effectiveness of complex interventions: a qualitative comparative analysis. *Trials, 14*, 179.

Capra, F. (2014). *The systems view of life: a unifying vision*. Cambridge: Cambridge University Press.

Carnap, R. (1947). On the application of inductive logic. *Intern Phenomen Soc, 8*, 133-148.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing to better*. New York: Oxford University Press.

Cook, D. J., Mulrow, C. D., & Haynes, R. B. (1997). Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med, 126*(5), 376-380.

Cooper, H. (1984). *The integrative research review: a systematic approach*. Newbury Park, CA: Sage Publications Ltd.

Cooper, N. J., Kendrick, D., Achana, F., Dhiman, P., He, Z., Wynn, P., . . . Sutton, A. J. (2012). Network meta-analysis to evaluate the effectiveness of interventions to increase the uptake of smoke alarms. *Epidemiol Rev, 34*, 32-45.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, P. (2008). Developing and evaluating complex interventions: new guidance. Retrieved 8 Feb, 2018 from https://www.mrc.ac.uk/documents/pdf/developing-and-evaluating-complex-interventions/

Craig, P., Katikireddi, S. V., Leyland, A., & Popham, F. (2017). Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. *Annu Rev Public Health, 38*, 39-56.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci, 4*, 50.

Datta, J., & Petticrew, M. (2013). Challenges to evaluating complex interventions: a content analysis of published papers. *BMC Public Health, 13*, 568.

Davies, P. (2005). *Evidence-based policy at the Cabinet Office: Impact and Insight Series*: Overseas Development Institute.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., . . . European Carotid Surgery Trial Collaborative, G. (2003). Evaluating non-randomised intervention studies. *Health Technol Assess, 7*(27), iii-x, 1-173.

Diez Roux, A. V. (2011). Complex systems thinking and current impasses in health disparities research. *Am J Public Health, 101*(9), 1627-1634.

Djulbegovic, B., Kumar, A., Kaufman, R. M., Tobian, A., & Guyatt, G. H. (2015). Quality of evidence is a key determinant for making a strong GRADE guidelines recommendation. *J Clin Epidemiol, 68*(7), 727-732.

Douthwaite, B., Kuby, E., Fliert, v. d., & Schultz, S. (2003). Impact pathway evaluation: an approach for achieving and attributing impact in complex systems. *Agricultural systems, 78*,243-65.

Ebell, M. H., Siwek, J., Weiss, B. D., Woolf, S. H., Susman, J., Ewigman, B., & Bowman, M. (2004). Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician, 69*(3), 548-556.

Field, A., & Hole, G. (2003). *How to design and report experiments*. London: Sage.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: criteria for efficacy, effectiveness and dissemination. *Prev Sci, 6*(3), 151-175.

Forsetlund, L., Bjorndal, A., Rashidian, A., Jamtvedt, G., O'Brien, M., & Wolf, F. (2009). Continuing education meetings and workshops: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev, 15*(2), CD003030.

Funnell, S. (1997). Program logic: an adaptable tool for distinguishing and evaluating programs. *Evaluation News and Comment, 6*(1), 5-7.

Galea, S., Riddle, M., & Kaplan, G. A. (2010). Causal thinking and complex system approaches in epidemiology. *Int J Epidemiol, 39*(1), 97-106.

Gambrill, E. (2006). Evidence-Based Practice and Policy: choices ahead. *Res. Soc. Work Pract., 16*(3), 338-357.

Gerhardus, A. (2016). on behalf of the INTEGRATE-HTA project team. Integrated health technology assessment for evaluating complex technologies (INTEGRATE-HTA): An introduction to the guidances. Retrieved 15 Jan, 2018 from http://www.integrate-hta.eu/downloads/

Gibbons, M. (1994). *The new production of knowledge: the dynamics of science and research in contemporary societies*. London: Sage Publications Ltd.

Gleick, J. (1987). *Chaos: making a new science*. New York: Penguin Books.

Glouberman, S., & Zimmerman, B. (2002). *Complicated and complex systems: What would successful reform of Medicare look like?* Commission on the Future of Health Care in Canada.

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. London, UK: SAGE Publications Ltd.

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. (2017). Retrieved 1 Feb, 2018 from http://gradeworkinggroup.org/

Grant, S. (2014). *Development of a CONSORT extension for social and psychological interventions.* DPhil thesis. Department of Social Policy and Intervention, Oxford: University of Oxford.

Gray, J., A., M. (1997). *Evidence-based healthcare: how to make health policy and management decisions*. London: Churchill Livingstone.

Greenhagh, T. (2013). Why do we always end up here? Evidence-based medicine's conceptual cul-de-sacs and some off-road alternative routes. *Int J Prosthodont, 26*(1), 11-15.

Greenwood-Lee, J., Hawe, P., Nettel-Aguirre, A., Shiell, A., & Marshall, D. A. (2016). Complex intervention modelling should capture the dynamics of adaptation. *BMC Med Res Methodol, 16*, 51.

Gruer, L., Tursan d'Espaignet, E., Haw, S., Fernandez, E., & Mackay, J. (2012). Smoke-free legislation: global reach, impact and remaining challenges. *Public Health, 126*(3), 227-229.

Guise, J. M., Butler, M., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017a). AHRQ series on complex intervention systematic reviews-paper 7: PRISMA-CI elaboration and explanation. *J Clin Epidemiol, 90*, 51-58.

Guise, J. M., Butler, M. E., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017b). AHRQ series on complex intervention systematic reviews-paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol, 90*,43-50.

Guise, J. M., Chang, C., Butler, M., Viswanathan, M., & Tugwell, P. (2017). AHRQ series on complex intervention systematic reviews-paper 1: an introduction to a series of articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol, 90*,6-10.

Guise, J. M., O'Haire, C., McPheeters, M., Most, C., LaBrant, L., Lee, K., . . . Graham, E. (2013). A practice-based tool for engaging stakeholders in future research: a synthesis of current practices. *J Clin Epidemiol, 66*(6), 666-674.

Guyatt, G. H., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., . . . Schunemann, H. J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol, 64*(4), 383-394.

Guyatt, G. H., Oxman, A. D., Kunz, R., Atkins, D., Brozek, J., Vist, G., . . . Schunemann, H. J. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol, 64*(4), 395-400.

Guyatt, G. H., Oxman, A. D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., . . . Schunemann, H. J. (2011). GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol, 64*(12), 1283-1293.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., . . . GRADE Working Group (2011). GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol, 64*(12), 1303-1310.

Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., . . . GRAD Working Group (2011). GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol, 64*(12), 1294-1302.

Guyatt, G. H., Oxman, A. D., Montori, V., Vist, G., Kunz, R., Brozek, J., . . . Schunemann, H. J. (2011). GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol, 64*(12), 1277-1282.

Guyatt, G. H., Oxman, A. D., Sultan, S., Glasziou, P., Akl, E. A., Alonso-Coello, P., . . . GRADE Working Group (2011). GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol, 64*(12), 1311-1316.

Guyatt, G. H., Oxman, A. D., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., . . . Schunemann, H. J. (2011). GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol, 64*(4), 407-415.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., . . . GRADE Working Group (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*(7650), 924-926.

Guyatt, G. H., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J., & Cook, R. J. (1995). Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA, 274*(22), 1800-1804.

Harder, T., Abu Sin, M., Bosch-Capblanch, X., Bruno, C., de Carvalho Gomes, H., Duclos, P., . . . Zuiderent-Jerak, T. (2015). Towards a framework for evaluating and grading evidence in public health. *Health Policy, 119*(6), 732-736.

Hawe, P., & Ghali, L. (2008). Use of social network analysis to map the social relationships of staff and teachers at school. *Health Educ Res, 23*(1), 62-69.

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: how "out of control" can a randomised controlled trial be? *BMJ, 328*(7455), 1561-1563.

Hawe, P., Shiell, A., & Riley, T. (2009). Theorising interventions as events in systems. *Am J Community Psychol, 43*(3-4), 267-276.

Haynes, R. B. (1999). Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. *BMJ, 319*,652-653.

Haynes, R. B., Devereaux, P. J., & Guyatt, G. H. (2002). Clinical expertise in the era of evidence-based medicine and patient choice. *ACP J Club, 136*(2), A11-14.

Higgins, J., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. Retrieved 8 Feb, 2018 from http://www.handbook.cochrane.org/

Hill, A., B. (1965). The Environment and Disease: Association or Causation? *Proc R Soc Med, 58*, 295-300.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., . . . Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*, g1687.

Howick, J. H. (2011). *Philosophy of evidence-based medicine*. Chichester: Wiley-Blackwell.

Hultcrantz, M., Rind, D., Akl, E. A., Treweek, S., Mustafa, R. A., Iorio, A., . . . Guyatt, G. (2017). The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol, 87*, 4-13.

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J., M., Odgaard-Jensen, J., & French, S. D. (2012). Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev, 6,* CD000259.

Jadad, A. R., Cook, D. J., & Browman, G. P. (1997). A guide to interpreting discordant systematic reviews. *CMAJ, 156*(10), 1411-1416.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., & McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials, 17*(1), 1-12.

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 Med Res Methodol, 4*, 22.

Kellert, S. (1993). *In the wake of chaos: unpredictable order in dynamical systems*. Chicago, IL: University of Chicago Press.

Kelly, M. P., Noyes, J., Kane, R. L., Chang, C., Uhl, S., Robinson, K. A., . . . Guise, J. M. (2017). AHRQ series on complex intervention systematic reviews-paper 2: defining complexity, formulating scope, and questions. *J Clin Epidemiol, 90*,11-18.

Kneale, D., Thomas, J., & Harris, K. (2015). Developing and Optimising the Use of Logic Models in Systematic Reviews: Exploring Practice and Good Practice in the Use of Programme Theory in Reviews. *PLoS One, 10*(11), e0142187.

Kuhne, F., Ehmcke, R., Harter, M., & Kriston, L. (2015). Conceptual decomposition of complex health care interventions for evidence synthesis: a literature review. *J Eval Clin Pract, 21*(5), 817-823.

Leischow, S., Best, A., & Trochim, W. (2008). Systems thinking to improve the public's health. *Am J Prev Med, 35*(2S), S196-203.

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., . . . Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to series. *Implement Sci, 13, Suppl 1*(2).

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gulmezoglu, M., . . . Rashidian, A. (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 Med, 12*(10), e1001895.

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., . . . Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

Linsley, P., Howard, D., & Owen, S. (2015). The construction of context-mechanisms-outcomes in realistic evaluation. *Nurse Res, 22*(3), 28-34.

Lipsey, M., W., & Wilson, D., B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage Publications Ltd.

Meadows, D. (2008). *Thinking in systems*. A Primer London: Earthscan.

Melendez-Torres, G. J., Bonell, C., & Thomas, J. (2015). Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Med Res Methodol, 15*, 47.

Michie, S., Fixsen, D., Grimshaw, J. M., & Eccles, M. P. (2009). Specifying and reporting complex behaviour change interventions: the need for a scientific method. *Implement Sci, 4*, 40.

Miller, J. (2007). *Complex adaptive systems: an introduction to computational models of social life* (Vol. Princeton University Press): Princeton, NJ; Oxford.

Moher, D., Glasziou, P., Chalmers, I., Nasser, M., Bossuyt, P. M., Korevaar, D. A., . . . Boutron, I. (2016). Increasing value and reducing waste in biomedical research: who's listening? *Lancet, 387*(10027), 1573-1586.

Möhler, R., Kopke, S., & Meyer, G. (2015). Criteria for Reporting the Development and Evaluation of Complex Interventions in healthcare: revised guideline (CReDECI 2). *Trials, 16*, 204.

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., . . . Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ, 350*, h1258.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016a). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol, 78*, 22-33.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016b). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

Munro, S., Lewin, S., Swart, T., & Volmink, J. (2007). A review of health behaviour theories: how useful are these for developing interventions to promote long-term medication adherence for TB and HIV/AIDS? *BMC Public Health, 7*, 104. doi:10.1186/1471-2458-7-104

Nagarajan, N., & Vanheukelen, M. (1997). *Evaluating EU expenditure programmes: a guide*. Brussels: Directorate-general for budgets of the European Union.

Nasser, S. M., Cooke, G., Kranzer, K., Norris, S. L., Olliaro, P., & Ford, N. (2015). Strength of recommendations in WHO guidelines using GRADE was associated with uptake in national policy. *J Clin Epidemiol, 68*(6), 703-707.

NICE. (2012). *Methods for the development of NICE public health guidance: Process and methods guides* (3rd ed.) National Institute for Health and Care Excellence. Retrieved 8 Feb, 2018 from https://www.nice.org.uk/process/pmg4/resources/methods-for-the-development-of-nice-public-health-guidance-third-edition-pdf-2007967445701/

Noyes, J., Gough, D., Lewin, S., Mayhew, A., Michie, S., Pantoja, T., . . . Welch, V. (2013). A research and development agenda for systematic reviews that ask complex questions about complex interventions. *J Clin Epidemiol, 66*(11), 1262-1270.

Noyes, J., Hendry, M., Lewin, S., Glenton, C., Chandler, J., & Rashidian, A. (2016). Qualitative "trial-sibling" studies and "unrelated" qualitative studies contributed to complex intervention reviews. *J Clin Epidemiol, 74*, 133-143.

O'Connor, D., Green, S., & Higgins, J., P., T. (2008). *Defining the review question and developing criteria for including studies:* In Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions. Chichester, UK: Wiley-Blackwell.

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. (2005). Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? *J Epidemiol Community Health, 59*(10), 886-892.

Oliver, S., & Dickson, K. (2016). Policy-relevant systematic reviews to strengthen health systems: models and mechanisms to support their production. *Evid Policy, 12*(2), 235-259.

Parkhurst, J., O., & Abeysinghe, S. (2014). What constitutes 'good' evidence for public health and social policy making? From hierarchies to appropriateness. *SERRC, 3*(4), 34-46.

Patton, M. (1997). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage Publications Ltd.

Patton, M. (2011). *Developmental evaluation: applying complexity concepts to enhanve innovation and use*. New York, NY: A Division of Guilford Publications, Inc.

Pawson, R. (2006). *Evidence-based policy: a realist perspective*. London: Sage Publications Ltd.

Petticrew, M. (2011). When are complex interventions 'complex'? When are simple interventions 'simple'? *Eur J Public Health, 21*(4), 397-398.

Petticrew, M. (2015). Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev, 4*(36).

Petticrew, M., Anderson, L., Elder, R., Grimshaw, J., Hopkins, D., Hahn, R., . . . Welch, V. (2013). Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol, 66*(11), 1209-1214.

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P., Mayhew, A., Pantoja, T., . . . Sowden, A. (2013). Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol, 66*(11), 1230-1243.

Petticrew, M., & Roberts, H. (2003). Evidence, hierarchies, and typologies: horses for courses. *J Epidemiol Community Health, 57*(7), 527-529.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in social sciences: a practical guide*. Malden, MA; Oxford: Blackwell Pub.

Petticrew, M., & Roberts, H. (2008). Systematic reviews--do they 'work' in informing decision-making around health inequalities? *Health Econ Policy Law, 3*(Pt 2), 197-211.

Petticrew, M., Shemilt, I., Lorenc, T., Marteau, T. M., Melendez-Torres, G. J., O'Mara-Eves, A., . . . Thomas, J. (2017). Alcohol advertising and public health: systems perspectives versus narrow perspectives. *J Epidemiol Community Health, 71*(3), 308-312.

Pfadenhauer, L. M., Gerhardus, A., Mozygemba, K., Lysdahl, K. B., Booth, A., Hofmann, B., . . . Rehfuess, E. (2017). Making sense of complexity in context and implementation: The Context and Implementation of Complex Interventions (CICI) framework. *Implement Sci, 12*(1), 21.

Pigott, T., Noyes, J., Umscheid, C. A., Myers, E., Morton, S. C., Fu, R., . . . Beretvas, S. N. (2017). AHRQ series on complex intervention systematic reviews-paper 5: advanced analytic methods. *J Clin Epidemiol, 90*, 37-42.

Plsek, P., & Greenhalgh, T. (2001). The challenge of complexity in health care. *BMJ, 323*,625.

Prigogine, I. (1996). *The end of certainty: time, chaos, and the new laws of nature*. New York, NY: The Free press.

Reed, M., & Harvey, D. (1992). The new science and the old: comlexity and realism in the social sciences. *J Theor Soc Behav, 22*,353-380.

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rehfuess, E. A., Booth, A., Brereton, L., Burns, J., Gerhardus, A., Mozygemba, K., . . . Rohwer, A. (2017). Towards a taxonomy of logic models in systematic reviews and health technology assessments: A priori, staged, and iterative approaches. *Res Synth Methods*.

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP J Club, 123*(3), A12-13.

Rogers, P. (2008). Using program theory to evaluate complicated and complex aspects of interventions. *Evaluation, 14*(1), 29-48.

Rohwer, A., Motaze, N., V., Rehfuess, E., & Young, T. (2017). E-learning of evidence-based health care (EBHC) to increase EBHC competencies in healthcare professionals: a systematic review. *Campbell Syst Rev, 4*.

Rohwer, A., Pfadenhauer, L., Burns, J., Brereton, L., Gerhardus, A., Booth, A., . . . Rehfuess, E. (2017). Series: Clinical Epidemiology in South Africa. Paper 3: Logic models help make sense of complexity in systematic reviews and health technology assessments. *J Clin Epidemiol, 83*, 37-47.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T., . . . White, M. (2017). The need for a complex systems model of evidence for public health. *Lancet*, 9;390(10112), 2602-2604.

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health, 56*(2), 119-127.

Sackett, D., L., Straus, S., E., Richardson, M., Rosenberg, W., S., & Haynes, R., B. (2000). *Evidence-Based Medicine: how to practice and teach EBM* (2nd ed.). London: Churchill Livingstone.

Sanders, M. R., Turner, K. M., & Markie-Dadds, C. (2002). The development and dissemination of the Triple P-Positive Parenting Program: a multilevel, evidence-based system of parenting and family support. *Prev Sci, 3*(3), 173-189.

Sanderson, I. (2006). Complexity, "practical rationality" and evidence-based policy making. *The Policy Press, 34*(1), 115-132.

Satterfield, J. M., Spring, B., Brownson, R. C., Mullen, E. J., Newhouse, R. P., Walker, B. B., & Whitlock, E. P. (2009). Toward a transdisciplinary model of evidence-based practice. *Milbank Q, 87*(2), 368-390.

Schunemann, H., Hill, S., Guyatt, G., Akl, E. A., & Ahmed, F. (2011). The GRADE approach and Bradford Hill's criteria for causation. *J Epidemiol Community Health, 65*(5), 392-395.

Shekelle, P. G., Woolf, S. H., Eccles, M., & Grimshaw, J. (1999). Clinical guidelines: developing guidelines. *BMJ, 318*(7183), 593-596.

Shiell, A., Hawe, P., & Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *BMJ, 336*(7656), 1281-1283.

Smith, K., E. (2013). *Beyond evidence-based policy in public health: the interplay of ideas*. Basingstoke: Palgrave Macmillan.

Squires, J. E., Valentine, J. C., & Grimshaw, J. M. (2013). Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol, 66*(11), 1215-1222.

Sur R., L., & Dahm, P. (2011). History of evidence-based medicine. *Indian J Urol, 27*(4), 487-489.

Taleb, N. (2007). *The black swan: the impact of highly improbable*. New York: Random House.

Tanner-Smith, E., E., & Grant, S. (2018). Meta-analysis of complex interventions. *Annu Rev Public Health, 39*, 16.11-16.17.

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., . . . Kavanagh, J. (2004). Integrating qualitative research with trials in systematic reviews. *BMJ, 328*(7446), 1010-1012.

Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Syst Rev, 3*, 67.

Thrift, N. (1999). The place of complexity. *Theor Cult Soc, 3*,31-69.

Uttley L., Montgomery P. (2017). The influence of the team in conducting a systematic review. *Syst Rev, 6*(1),149.

Victora, C. G., Habicht, J. P., & Bryce, J. (2004). Evidence-based public health: moving beyond randomized trials. *Am J Public Health, 94*(3), 400-405.

Viswanathan, M., McPheeters, M. L., Murad, M. H., Butler, M. E., Devine, E. E. B., Dyson, M. P., . . . Morton, S. C. (2017). AHRQ series on complex intervention systematic reviews-paper 4: selecting analytic approaches. *J Clin Epidemiol, 90*,28-36.

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. *J Epidemiol Community Health, 67*(1), 98-104.

Wahlster, P., Brereton, L., Burns, J., Hofmann, B., Mozygemba, K., Oortwijn, W., . . . Gerhardus, A. (2016). Guidance on the integrated assessment of complex health technologies - the INTEGRATE-HTA Model. Retrieved 19 Jan, 2018 from http://www.integrate-hta.eu/downloads/

Weiss, C. (1998). *Evaluation: methods for studying programs and policies*. Englewood Cliffs: Prentice Hall.

Weiss, C., H. (1977). Research for policy's sake: the enlightment function of social research. *Policy Anal, 3*(4), 531-545.

West, S., King, V., & Carey, T. e. a. (2002). *Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47* (Prepared by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). Rockville, MD: Agency for Healthcare Research and Quality.

WHO. (2018). Retrieval, Synthesis and Assessment of Evidence on Complex Health Interventions. World Health Organization Guidelines on Maternal, Newborn, Child and Adolescent Health. Retrieved from 28 Jan, 2018 http://www.who.int/maternal_child_adolescent/guidelines/development/complex-health-interventions/en/

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: realist syntheses. *BMC Med, 11*, 21.

Wortman, P., M. (1994). *Judging research quality.* In Cooper H., Hedges L., V. (Eds). The handbook of research synthesis. New York: Russell Sage Foundation.

Chapter 2. Thesis methodology

Chapter overview

The previous chapter provided a broad overview of the GRADE approach and the recent initiatives on considering complexity in systematic reviews. In light of the reported challenges of using GRADE in systematic reviews of complex interventions, a case has been made for a new guidance on how to consider sources of complexity when applying the GRADE domains to rate the certainty of evidence in systematic reviews. This chapter further outlines the aims, specific research questions and methods of this thesis work.

The proposed thesis methodology largely draws on the recommended strategy for developing and disseminating research reporting guidelines in line with the principles of EBP (Moher et al., 2010). Specific techniques of this strategy have previously been used to develop research reporting guidelines, such as the CONSORT statement and its extensions, as well as to advance the GRADE guidance for use in different contexts (Akl et al., 2017). This chapter describes the specific methods and phases of research within the broad thesis methodology, and how the thesis chapters contribute to these phases. It also describes how this thesis fits within the wider research project on developing *GRADE Guidance for Complex Interventions*. The specific roles of the DPhil candidate as a student and a project Research Assistant are also outlined. Ethical approval for different phases of the thesis research has been obtained from the Departmental Research Ethics Committee at the University of Oxford.

Thesis logic and research questions

As has been discussed in Chapter 1, there is a growing interest in methods for incorporating a complexity perspective in systematic reviews and practice guidelines. While a number of initiatives have been launched in the recent decade on how to consider complexity in formulating review questions and synthesising evidence (Anderson, Oliver, et al., 2013; Petticrew, Rehfuess, et al., 2013; Pigott et al., 2017; Squires et al., 2013), no guidance is currently available on how to address complexity when assessing and rating evidence in systematic reviews, and, specifically, when using the GRADE approach (Rehfuess & Akl, 2013). In the meantime, challenges have been reported in using GRADE in reviews of complex interventions, and concerns have been raised that GRADE often downgrades *"the best evidence possible"* for complex interventions; the frequently observed "low" certainty of evidence ratings are perceived to potentially divert decision-makers from implementing wider system-level interventions, drawing the focus to those interventions, which are easier to implement and evaluate (Ogilvie et al., 2005; Rutter et al., 2017). In this view, researchers in public health and social policy have called for an extended guidance on how to use GRADE in the context of complex interventions (Movsisyan et al., 2016b).

By using a rigorous and transparent mixed-methods approach, this thesis seeks to further investigate the challenges of using GRADE in systematic reviews of complex interventions and explore how the GRADE approach can be advanced to address these challenges. In this way, this thesis aims to inform the write-up of a new GRADE guidance for complex interventions. To address these broad aims, the thesis work draws on the recommended methodology for developing research reporting guidelines based on a

collective experience from the development of more than 20 research reporting guidelines (Moher et al., 2010). For a transparent and efficient development of a new guidance, the methodology proposed by Moher et al. (2010) promotes a linear process, which emphasises the importance of incorporating existing methodological knowledge by way of conducting a literature review, as well as consultation with relevant experts in the field. For the latter, it recommends employing consensus development methods, including a Delphi process followed by a face-to-face expert meeting. This methodology also highlights the importance of post-meeting activities, including the write-up and pilot testing of the new guidance and dissemination activities, such as translation of the guidance and development of a dedicated website. This thesis work focuses on the phases of research leading to the write-up of the GRADE guidance for complex interventions. Figure 2.1 outlines the logic of this thesis work showing how each chapter contributes to the phases of research suggested by Moher et al. (2010). Specific research questions within each phase include:

Phase 1: Preparatory activities (Chapters 1 and 3)

- What are the challenges of using the GRADE approach for complex interventions?
- What is the empirical evidence supporting the need for a new GRADE guidance for complex intervention?

Phase 2: Pre-meeting activities (Chapters 4 and 5)

- How should the challenges of using GRADE in systematic reviews of complex interventions be addressed?
- What domains of evidence should be considered in the GRADE guidance for complex interventions?

Phase 3: Face-to-face expert meeting (Chapter 6)

- How should the construct of "certainty of evidence" be conceptualised in the GRADE guidance for complex interventions?
- What domains of evidence should be included in the guidance, and how should these domains be operationalised?

Phase 4: Post-meeting activities (Chapter 7)

- How should the GRADE guidance for complex interventions be written and disseminated?

Following the traditional monograph format, this thesis includes 4 chapters (specifically, chapters 3 to 6) describing the results from the thesis research phases. These include (1) a systematic review of existing evidence rating systems, (2) semi-structured interviews with important stakeholders, including review authors and GRADE methodologists, (3) a Delphi-based online expert panel exploring agreement and disagreement regarding the content and structure of the GRADE guidance for complex interventions, and, finally, (4) a three-day face-to-face expert meeting with a select group of stakeholders to finalise the guidance. Building on each other, these chapters provide empirical evidence for the content and dissemination of the GRADE guidance for complex interventions. Each of these chapters includes a background and methods sections justifying and detailing the conduct of the specific phase of the research; meanwhile, results and discussion sections summarise the findings in each chapter and contextualise them in light of the broader thesis aims. These chapters have been adapted into four corresponding manuscripts and submitted for peer-review publication (see Figure 2.1).

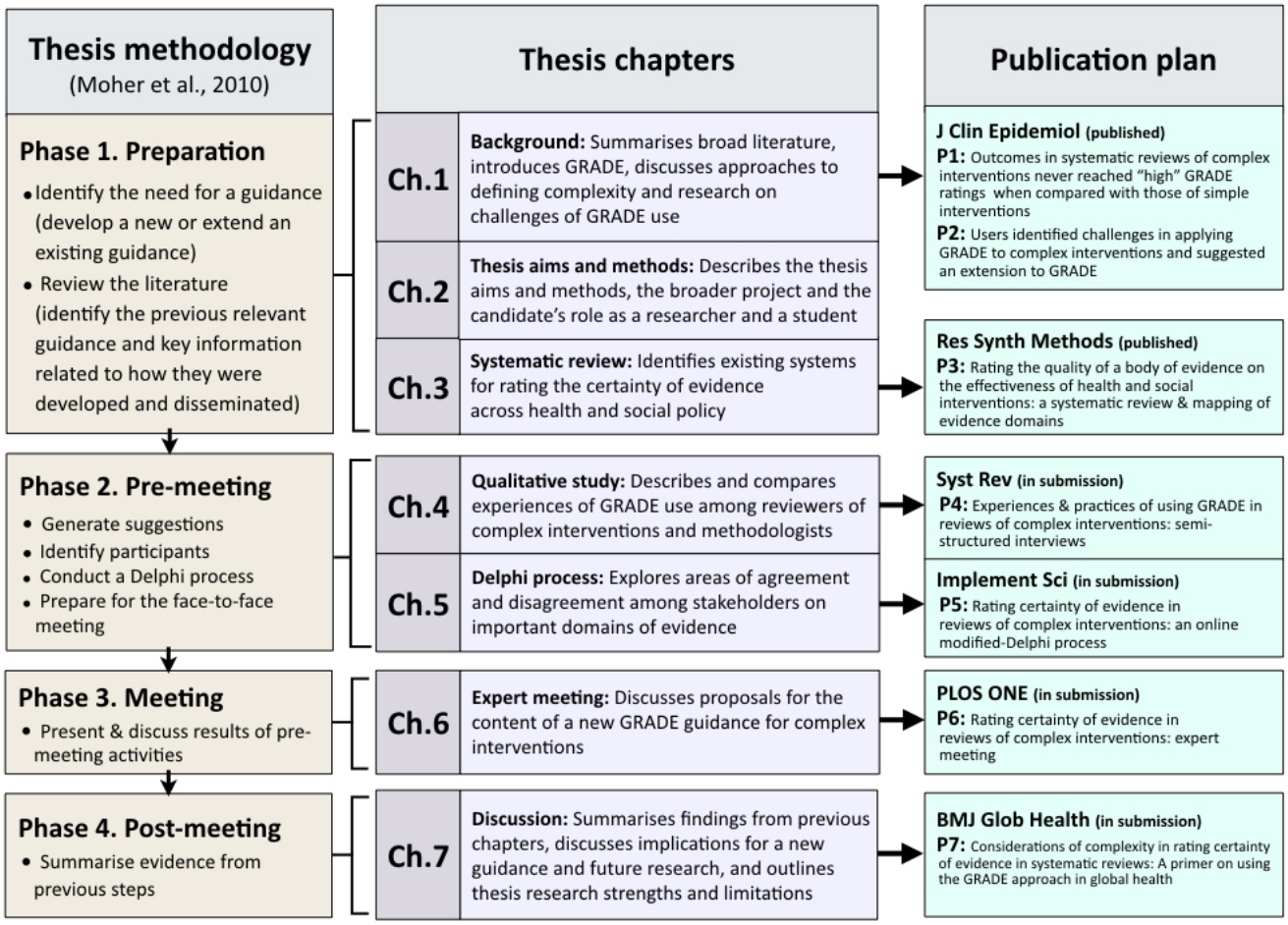


Figure 2.1. Thesis logic model outlining how the thesis chapters contribute to the thesis research phases (Moher et al., 2010) and papers published or in submission

Thesis research phases and methods

As noted above, this thesis work extends on the methodology described by Moher et al. (2010) for developing research reporting guidelines. This methodology was adopted in this thesis because it has proven helpful in establishing transparent techniques for developing research reporting guidance, including the CONSORT statement and its extensions (Altman et al., 2001; Boutron et al., 2008). Specific techniques of this methodology have also been used to advance GRADE guidelines for different applications, such as the GRADE Equity Guidelines (Akl et al., 2017). These techniques are as important for development of guidance on rating the certainty of evidence as for research reporting, as they follow the key principles of EBP (Sackett, 2000). Similar to the EBP model discussed in Chapter 1 (see Figure 1.1), the methodology by Moher et al. (2010) recommends integration of the best available empirical evidence with knowledge obtained from rounds of stakeholder consultations to develop the new guidance. The overall process includes several phases of research (see Figure 2.1).

Phase 1. Preparatory activities: Systematic review

Developers of research reporting guidelines, including members of the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network ("EQUATOR", 2018), recommend a few initial activities to provide a strong rationale for a new guidance (Moher et al., 2010). This includes a comprehensive literature review to examine whether an adequate guidance already exists in relation to the considered topic and scope. A literature review is also recommended to identify empirical evidence on the potential problems with the existing guidance, and to serve as initial mapping of important stakeholder contributors and a range of considerations to be further examined

for the new guidance. As a result, researchers may choose to pursue one of the following three options. Specifically, researchers may choose to develop an entirely new guidance, if no existing guidance was established on the topic under consideration. Researchers may alternatively choose to extend an existing guidance, if they establish that the existing broad guidance can be augmented by additional guidance for a specific topic; an example of such an extension in research reporting is the recently published PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for Complex Interventions (Guise, Butler, et al., 2017a; Guise, Butler, et al., 2017b). Finally, researchers can choose to illustrate how an existing guidance can be implemented for a specific topic of research or domain of practice. However, it should be noted that since such implementations generally involve a need for additional topic-specific considerations, the distinction between an extension and an implementation is often arbitrary (Moher et al., 2010).

In line with this process, Chapter 1 of this thesis provides a broad literature review on the topic, including discussion of the existing empirical research on challenges of using GRADE in systematic reviews of complex interventions (Barbui et al., 2010; Movsisyan et al., 2016a, 2016b; Rehfuess & Akl, 2013). The DPhil candidate herself has contributed to this literature through two publications in the *Journal of Clinical Epidemiology* at the outset of her DPhil project investigating experiences with GRADE use in reviews of complex interventions. In the meantime, Chapter 3 of this thesis describes a systematic review led by the DPhil candidate on existing evidence rating systems relevant for complex interventions. The systematic review maps the domains of evidence described in the identified systems and assesses reported procedures for development and dissemination of these domains. It establishes that the majority of the domains in

the existing evidence rating systems are comprehensively covered by the GRADE approach, however, it also reveals ad-hoc attempts to modify GRADE to address sources of complexity in public health interventions. Along with the empirical evidence discussed in Chapter 1, this systematic review provides a rationale for the need to augment the existing GRADE guidance to address the sources of complexity in interventions and their causal pathways. Specific methods employed in the systematic review, including the search strategy and data extraction are further described in Chapter 3.

Phase 2. Pre-meeting activities: Stakeholder consultations

After the need for a new guidance is established, the methodology by Moher et al. (2010) recommends specific activities to prepare for a face-to-face expert meeting, where the content and dissemination strategy of the guidance is to be finalised. These activities involve identification and consultation with key stakeholders and generation of a list of considerations to further discuss at the expert meeting. It is suggested that the expertise of individuals involved in the consultations reflects the particular guidance under consideration. For a new GRADE guidance, participants would include methodologists, epidemiologists, content experts, journal editors and consumer representatives, that is, representatives from all key stakeholder groups in the entire machinery of systematic reviewing. Since not all potential participants can be invited and will be available to come to the main guidance development meeting, it is recommended to conduct a Delphi process to engage a wider group of stakeholders (Hopewell et al., 2008). The Delphi method provides a structured process of collecting information from a group of stakeholders by means of a series of questionnaires, each one refined based on the feedback from participants on a previous version (Murphy et al., 1998).

Chapters 4 and 5 of this thesis summarise the results from consultations with key stakeholders conducted by the DPhil candidate to inform the development of the GRADE guidance for complex interventions. Specifically, Chapter 4 describes the results of semi-structured interviews conducted with 10 Cochrane and Campbell review authors and 5 GRADE methodologists from the GRADE Working Group regarding their views on GRADE use in the same sample of systematic reviews of complex interventions. Through comparison of practices and experiences of these two important stakeholder groups, that is, users of GRADE (i.e., review authors) and developers of the GRADE guidelines (i.e., GRADE methodologists), this qualitative research helps to separate the "real" challenges of the GRADE approach when used in reviews of complex interventions from those related to lack of skills and inadequate implementation of GRADE. This cross-examination also allowed to further specify the need and appropriate suggestions for developing a new GRADE guidance tailored to complex interventions.

In accordance with the recommendations described above, a list of considerations for a new GRADE guidance was generated based on the findings of research described in Chapters 3 and 4. These considerations were further examined for agreement in a multidisciplinary group of 114 stakeholders in an online-modified Delphi process. Chapter 5 further details the methods and findings from this Delphi-based process.

Phase 3. Face-to-face expert meeting

The key phase in the methodology for developing research reporting guidance is the face-to-face expert meeting, where a group of 20-25 key stakeholders are brought together over the course of 2-3 days to discuss and finalise the content of the guidance

(Moher et al., 2010). It is recommended that this meeting closely follows the pre-meeting preparation plans and structure, including a meeting agenda developed in advance. The substantive meeting is suggested to begin with formal presentation of background topics, summary of evidence from the previous research, including the results of the Delphi process. The meeting is usually divided into sessions. While most of the chairs of sessions should include members of the project team, other participants of the meeting may also be invited to moderate the sessions, especially if they have previous expertise in chairing meetings.

In line with this recommendation (Moher et al., 2010), Chapter 6 of this thesis describes the key themes from discussions of a three-day face-to-face expert meeting organised by the DPhil candidate and the research team leading the project on developing *GRADE Guidance for Complex Interventions* (see below). The meeting was held in Oxford, UK in May 2017, where a group of 28 stakeholders with a range of content expertise and professional roles, including guideline developers, systematic reviewers, methodologists of complex interventions, journal editors and research funders discussed and made decisions on the content and dissemination of the GRADE guidance for complex interventions. The discussions were grounded in the evidence from the previous research phases, and the meeting sessions were moderated by the DPhil candidate and the project co-investigators. Further details on the recruitment of participants, specific procedures and the meeting agenda are described in Chapter 6.

Phase 4. Post-meeting activities

After the meeting, developers of guidance should start drafting the guidance paper (or series of papers). This process may take several iterations, which entails taking

the discussions from the face-to-face expert meeting on specific aspects of the guidance content and appropriately wording it in the paper with examples of application (Moher et al., 2010). It is recommended that the efforts aimed at extending or implementing an existing guidance make it very clear which parts of the new paper have remained the same as in the original guidance and which aspects have been modified or added. In the meantime, for transparency of research, it is suggested that a short document is written, in addition to the guidance paper itself, reporting on the rationale for developing the new guidance and the development process. After drafting the guidance paper (or papers depending on the adopted publication strategy), it is recommended that the paper is circulated among the meeting participants for feedback and pilot testing. The received comments should be incorporated into the guidance revisions (Moher et al., 2010). Finally, it is also recommended that developers of a new guidance have a publication and implementation strategy, which will promote the uptake of the new guidance and enhance its visibility. This may include activities, such as multiple publications of the guidance in relevant journals and development of a web page dedicated to the guidance.

While this thesis work focuses only on the phases of research leading to the write-up of the GRADE guidance for complex interventions, Chapter 7 provides a discussion of the implications of the findings from the three research phases outlined above. These implications aim to provide important basis for the project team tasked with the write-up and dissemination of the GRADE guidance for complex interventions.

The project on developing *GRADE Guidance for Complex Interventions*

The DPhil research reported in this thesis is embedded within a broader project on developing *GRADE Guidance for Complex Interventions*. This project is funded by the Department for International Development (DFID) and Economic and Social Research Council (ESRC) grant (ref number: ES/N012267/1). The project was launched in January, 2016 and was initially based in the Department of Social Policy & Intervention, University of Oxford; however, it has now moved to the Department of Social Policy, Sociology and Criminology, University of Birmingham with the change of the affiliation of the project Principal Investigator, Prof Paul Montgomery. The project is expected to be complete by July, 2018.

Project aims and scope

The project represents multiple collaborations among relevant stakeholder groups, including the GRADE Working Group, the World Health Organization and the RAND Corporation. It aims to develop a new guidance on how to use GRADE in systematic reviews of complex interventions. For this, the paper (or the series of papers) produced by the project team needs to be internally reviewed and approved by the members of the GRADE Working Group. The overall project strategy is the same as that described in this thesis work, that is, the project draws on the recommended techniques for developing research reporting guidelines to develop the guidance (Moher et al., 2010). In this view, the research presented in this thesis will be instrumental in informing the write-up and dissemination of the GRADE guidance for complex interventions. The guidance itself, however, is an expected output of the project and not this thesis work.

Project team

The project involves 6 co-investigators from a range of disciplines and professional roles: Dr Erik von Elm, Dr Eva Rehfuess, Prof Geraldine Macdonald, Dr Jane Dennis, Dr Sean Grant and Dr Susan Norris. These co-investigators are involved to various degrees in advising on different phases of research, recommending stakeholders for the Delphi process and the expert meeting, writing up the guidance papers and promoting the guidance dissemination and implementation. The DPhil candidate serves as a Research Assistant (RA) on the project and has had significant contribution to all phases of the research, including drafting grant proposals for project funding. Her specific roles are further described below.

The GRADE Working Group

To enhance the uptake and visibility of the GRADE guidance for complex interventions, this project closely collaborates with the GRADE Working Group to develop papers that are *official GRADE guidance*. The project team has, therefore, sought feedback and engagement of the interested members and chairs of the GRADE Working Group from the conception of the project idea and throughout all the following project phases; the project team itself includes members of the GRADE Working Group. The structure and procedures of the GRADE Working Group and the publication of the official GRADE papers are discussed in detail in Chapter 6.

The candidate's roles as a DPhil student and a project Research Assistant

I commenced my DPhil research in October 2014 and was involved in setting up the project on developing *GRADE Guidance for Complex Interventions* from the beginning, including drafting grant applications and liaising with potential co-investigators. I am currently registered as a Research Assistant (RA) on the project, and the salary I receive contributes to my living expenses as a DPhil student in Oxford. The project Principal Investigator, Prof Paul Montgomery, has been the official supervisor of my DPhil thesis before moving to the University of Birmingham. I am currently officially supervised by Dr Dave Humphreys, however, I still continue working and hold regular meetings with Prof Paul Montgomery. The idea for the project was conceived by Prof Paul Montgomery and myself, and I conducted a preliminary empirical work (reported in Chapter 1 of the thesis) on using GRADE in systematic reviews of complex interventions independently from the project. I also commenced working on the systematic review prior to the project launch (see Chapter 3).

As a project RA, I helped to develop the conceptual rationale for the project, as well as drafted grant applications and established the project team. Throughout the project execution, I have taken substantive administrative roles by maintaining effective communication among project co-investigators and collaborators, preparing materials and setting up project meetings, and attending and presenting at different seminars and workshops, including the annual meetings of the GRADE Working Group (see dissemination of the thesis work). I have been responsible for drafting the five papers produced so far by the project, and I am a lead author on three of them. I am also

expected to contribute to the write-up of the project final paper (or series of papers), that will comprise the GRADE guidance for complex interventions.

As a DPhil student, I have been leading the design, execution and data analysis of all phases of the research described above and reported in the thesis chapters. I undertook the systematic review, including searches in databases and key agency websites, data extraction and data analysis. I also undertook all the interviews with review authors and GRADE methodologists, transcribed and analysed the data. Further, I designed the online-modified Delphi process, including recruiting participants, developing the questionnaires, sending reminders for Delphi rounds and data analysis. Finally, I also drafted the agenda for the expert face-to-face meeting and actively participated in the moderation of the sessions at the meeting. I double checked the transcription of the recordings of this meeting (the initial transcription was conducted by another project RA) and further analysed the data. It is, however, worth noting that throughout these different research phases, I had an advantage as a DPhil student to consult the project co-investigators for their expertise and feedback in addition to the guidance received from my supervisors. The co-investigators also contributed to the rigour of this thesis research by helping with procedures, such as double data extraction and reading the preliminary drafts of the analysed data.

While my DPhil thesis uses the resources from the project to enable transparent and rigorous execution of different phases of the research to inform the development of the guidance, it is important to note that my DPhil thesis does not aim to produce an official GRADE guidance for complex interventions (that is, approved by the GRADE Working Group). Such guidance is an expected output of the project instead and will require further phases of write-up and pilot testing. I have chosen the traditional

monograph route for my thesis to highlight the independence of my thesis outputs and conclusions from those of the project, which might be subject to funder expectations and further communication and collaboration with key stakeholders, such as the GRADE Working Group.

Finally, in addition to my roles as an RA and a DPhil student, it is also important to note my professional interests and personal characteristics as these may have inevitably shaped the thesis research methods and its findings. Much of the personal investment in this thesis derives from my background in public health and interest in complex systems thinking. In the meantime, I am also a proponent of the principles of the evidence-based practice (EBP) model and, specifically, how it can be applied to social areas of practice, such as public health policy. I have started working on the issues around GRADE use in complex public health interventions when still reading for a Master of Science (MSc) in Evidence-based Social Intervention in the same Department at the University of Oxford. During the first few months of my DPhil project, I worked to extend my MSc thesis into manuscripts, which I subsequently published in the *Journal of Clinical Epidemiology* (these are discussed in Chapter 1). At the outset of my DPhil project, I also attended a few meetings of the GRADE Working Group as one of the major entities promoting the development and application of the EBP principles to healthcare decisions. I am currently a member of this group and have frequent contacts with other interested researchers in the Group. Throughout the coordination of my DPhil thesis research, I have (consciously) aimed to make judgments based on academic arguments and principles. Nevertheless, there are personal characteristics, including my ethnicity (white Armenian) and gender (female), which may have (unconsciously) influenced my decisions, such as recruitment of participants and interpretation of the research findings.

References

Akl, E. A., Welch, V., Pottie, K., Eslava-Schmalbach, J., Darzi, A., Sola, I., . . . Tugwell, P. (2017). GRADE equity guidelines 2: considering health equity in GRADE guideline development: equity extension of the guideline development checklist. *J Clin Epidemiol, 90*, 68-75.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., . . . Consort, G. (2001). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med, 134*(8), 663-694.

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. *J Clin Epidemiol, 66*(11), 1223-1229.

Barbui, C., Dua, T., van Ommeren, M., Yasamy, M. T., Fleischmann, A., Clark, N., . . . Saxena, S. (2010). Challenges in developing evidence-based recommendations using the GRADE approach: the case of mental, neurological, and substance use disorders. *PLoS Med, 7*(8).

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., Ravaud, P., & Group, C. (2008). Methods and processes of the CONSORT Group: example of an extension for trials assessing nonpharmacologic treatments. *Ann Intern Med, 148*(4), W60-66.

Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network (2018). Retrieved 20 Feb, 2018 from http://www.equator-network.org/

GRADE Guidance for Complex Interventions. (2016). Retrieved 15 Feb, 2018 from http://tinyurl.com/GRADE-Extension

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. (2017). Retrieved 1 Feb, 2018 from http://gradeworkinggroup.org/

Guise, J. M., Butler, M., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017a). AHRQ series on complex intervention systematic reviews-paper 7: PRISMA-CI elaboration and explanation. *J Clin Epidemiol, 90*, 51-58.

Guise, J. M., Butler, M. E., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017b). AHRQ series on complex intervention systematic reviews-paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol, 90*, 43-50.

Hopewell, S., Clarke, M., Moher, D., Wager, E., Middleton, P., Altman, D. G., . . . Group, C. (2008). CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med, 5*(1), e20.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2), e1000217.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016a). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol, 78*, 22-33.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016b). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technol Assess, 2*(3), i-iv, 1-88.

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. (2005). Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? *J Epidemiol Community Health, 59*(10), 886-892.

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P., Mayhew, A., Pantoja, T., . . . Sowden, A. (2013). Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol, 66*(11), 1230-1243.

Pigott, T., Noyes, J., Umscheid, C. A., Myers, E., Morton, S. C., Fu, R., . . . Beretvas, S. N. (2017). AHRQ series on complex intervention systematic reviews-paper 5: advanced analytic methods. *J Clin Epidemiol, 90*, 37-42.

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rutter, H., Savona, N., Glonti, K., Bibby, J., Cummins, S., Finegood, D. T., . . . White, M. (2017). The need for a complex systems model of evidence for public health. *Lancet*, 9;390(10112), 2602-2604.

Sackett, D., L. (2000). *Evidence-based medicine: how to practice and teach EBM*. Edinburgh: Churchill Livingstone.

Squires, J. E., Valentine, J. C., & Grimshaw, J. M. (2013). Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol, 66*(11), 1215-1222.

Chapter 3. A systematic review and mapping of evidence domains

Rating the certainty of evidence on the effectiveness of health and social interventions

A paper adaptation of this chapter has been published in the *Research Synthesis Methods*

Chapter overview

Through a systematic review, involving a multi-component search strategy and critical appraisal of existing evidence rating systems from health and social policy, this chapter examines the need for a new GRADE guidance for rating the certainty of evidence in systematic reviews of complex interventions. Thirteen discrete domains of evidence identified across 17 systems are described: study design, study execution, consistency, measures of precision, directness, publication bias, magnitude of effect, dose-response, plausible confounding, analogy, robustness, applicability and coherence.

While the majority of these domains are comprehensively covered by the GRADE approach, a few represent ad-hoc modifications of the GRADE approach for wider public health interventions. However, little reporting of rigorous procedures was found underpinning the development and dissemination of these systems. While GRADE is the most comprehensive and rigorous evidence rating system in its guidance, development and dissemination, it has been developed and mainly applied in biomedical interventions. A new GRADE guidance, which considers sources of complexity in interventions and their causal pathways will, therefore, be most appropriate to ensure widespread dissemination and uptake of the GRADE approach by relevant stakeholders in wider areas of practice, such as public health and social policy.

Introduction

The literature discussed in Chapter 1 suggests a need for a new guidance, which incorporates a complexity perspective into an evidence synthesis and, specifically, into the process of rating the certainty of evidence. Rating the certainty of evidence is an important procedure, which aims to translate evidence from systematic reviews to the immediate users, including guideline developers, practitioners and policy makers (Guyatt et al., 2008). Challenges, however, have been described by those attempting to rate the certainty of evidence for interventions, which are frequently denoted as “complex” (Rehfuess & Akl, 2013). While there are different potential sources of complexity (see Chapter 1 for a discussion on different perspectives to defining complexity), some sources of it are commonly ascribed to aspects of the interventions themselves (Craig et al., 2008; Lewin et al., 2017), such as those with multiple components that aim to address different and multiple causes of problems (for example, biological and social). Other sources of complexity are seen as emanating from system properties (Hawe, Shiell, & Riley, 2009), that is to say, long, non-linear, and dynamic relationships between interventions and outcomes, interactions and interdependences between different components of interventions and levels of target (Diez Roux, 2011).

The GRADE approach is currently the pre-eminent system for rating the certainty of evidence in healthcare, and has been widely adopted by systematic reviewers and guideline developers, including over 100 organisations worldwide ("GRADE Working Group", 2017). Despite the impact of the GRADE approach, empirical investigation into how GRADE is used in systematic reviews of social and public health interventions reported in Chapter 1 suggests that GRADE does not adequately address the complexity

of these interventions (Movsisyan, Melendez-Torres, & Montgomery, 2016a, 2016b). There are concerns that this might hinder uptake and implementation of GRADE beyond biomedical areas of practice, thereby, undermining the quality of systematic reviewing in public health and social policy. In the meantime, concerns are raised that use of the GRADE approach, as it stands, might produce misleading certainty of ratings for the outcomes of complex interventions. For example, frequently observed "low" and "very low" certainty of evidence ratings may deflect decision-makers from considering many important complex interventions for practice and policy (Movsisyan et al., 2016a, 2016b; Ogilvie, Egan, Hamilton, & Petticrew, 2005; Rehfuess & Akl, 2013).

Drawing on the best-practice techniques for developing research reporting guidelines (Moher, Schulz, Simera, & Altman, 2010), before developing a new guidance or extending an existing one, researchers need to undertake a few preliminary activities to justify such a project. Specifically, a comprehensive literature review should be carried out first of all, to identify whether an adequate guidance already exists relating to the considered scope. This initial needs assessment will then facilitate identification of stakeholder groups, stakeholder contributors, and a spectrum of considerations to be further examined for scientific agreement for the new guidance (Moher et al., 2010). A systematic review of the content, development and dissemination of systems for rating the certainty of evidence on intervention effectiveness should, therefore, indicate whether any single existing system already sufficiently addresses the sources of complexity, and whether a new guidance is needed.

The previous systematic reviews investigating evidence rating systems have mainly focused on scientific evidence in clinical medicine contexts and have not included systems from social policy domains, such as public health, education, international

development and crime and justice (Bai, Shukla, Bak, & Wells, 2012; West, King, & Carey, 2002). Moreover, these reviews did not investigate how rigorously the evidence ratings systems were developed and disseminated. A systematic review with a larger scope to include systems from both health and social policy is needed and would have the ability to demonstrate that existing systems do not adequately integrate a complexity perspective, when rating the certainty of evidence, and that a new guidance is needed for that purpose. This review would also provide a comprehensive synthesis of the previous work of the field to guide the development of the new guidance.

Correspondingly, this chapter reports on a systematic review that aims to assess the need for a new guidance for rating the certainty of evidence for complex interventions. The specific objectives of this review are to (1) identify existing systems for rating the certainty of evidence on intervention effectiveness across health and social policy, (2) examine how existing systems describe the construct of "certainty of evidence" on intervention effectiveness and map out a discrete set of evidence domains, (3) examine the extent to which the identified domains of evidence address sources of complexity as presented in Table 3.1, and, finally, (4) describe the reported procedures used to develop and disseminate the identified systems. A paper adaptation of this chapter has been published in *Research Synthesis Methods*.

Table 3.1. Frequently described sources of complexity in systematic reviews, adapted from Petticrew et al., 2013

1. Characteristics of the intervention itself
Number of components
Number of groups or organisational levels targeted
Degree of flexibility or tailoring permitted
Self-organisation, adaptivity and changes over time
2. Characteristics of the intervention's causal pathway
Number and variability of outcomes
Non-linear relationships and phase changes
Number of mediators and moderators of effect
Positive/negative feedback loops
Synergies/dysynergies between components
Interaction with context

Methods

Eligibility criteria

To be included in this review, a system had to (1) comprise a full-length document reporting a procedure for rating the certainty of a body of evidence derived from evidence synthesis integrating results across individual studies on the effectiveness of health or social interventions, and (2) have been published in English from 1995 onward, when evidence rating was first proposed as a stage of systematic reviewing (Guyatt et al., 1995). Where a document discussed a system developed by others (e.g., a literature review), the original documents reporting those systems were retrieved and examined for eligibility. Documents were excluded, if they described a procedure for rating the certainty of evidence on intervention effectiveness for a specific clinical topic (e.g., systems used in specific guidelines on osteoarthritis and brain injury), as these are

extensively described by the previous systematic reviews (Bai et al., 2012; West et al., 2002). Systems that were no longer used by an organisation were also excluded from the review (e.g., the systems previously used by the Scottish Intercollegiate Guidelines Network and the Institute for Clinical Systems Improvement, before these organisations adopted GRADE). Information on suspended use of these systems was either directly available on the organisation's website, or was obtained through email communication with representatives of the organisation.

Systematic search strategy

A multi-component search strategy was applied with multiple sources in an attempt to maximise the sensitivity of the search. First, search strategies used in the previous systematic reviews (Bai et al., 2012; West et al., 2002) were updated and expanded to include key social science databases. Searches were then run on 2 June 2016 in the following databases: Applied Social Sciences Index (ASSIA), Cochrane Methodology Register (Cochrane Library), EMBASE, MEDLINE, PsycINFO, SCIE Social Care Online, Scopus Social Sciences, and Social Sciences Citation Index (Web of Knowledge). Next, using the personal network of experts in the area and through bibliography searches of the related general literature, websites of 83 key stakeholder organisations were located and searched that specifically aim to aggregate, review and assess evidence across social policy domains, such as child and family welfare, international development, crime and justice, public health, and education (see Appendix 1 for the detailed search strategy). Third, bibliographies of all the included documents and literature reviews containing secondary reporting of potentially eligible systems were searched. Finally,

experts identified from the website searches were consulted to check whether any systems were missed.

Screening of all titles, abstracts and full-text documents was conducted by the DPhil candidate by using the Rayyan web application for systematic reviews (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016). A subset of randomly chosen titles (10%) was independently screened by a co-investigator of the project on developing *GRADE Guidance for Complex Interventions*. All discrepancies were discussed until agreement was reached.

Data extraction

Data on 4 types of information were extracted. First, descriptive information was extracted about each included system, namely the author, year, title, publication source, and eligibility criteria. Then information was extracted from each system on how its authors defined the construct of "certainty of evidence". Details of specific domains within the system used to rate the certainty of evidence were further extracted, as well as how these domains were defined, and how the certainty of evidence ratings were categorised in the system (for example, "high", "moderate", or "low"). Based on a template of pre-specified domains related to the development and dissemination of research reporting guidelines (Grant, Mayo-Wilson, Melendez-Torres, & Montgomery, 2013; Moher et al., 2010), data were extracted on whether included systems reported any preparatory activities, such as a review of literature on existing domains for rating evidence and consensus based activities, such as a Delphi exercise and expert meetings to develop the system. Finally, information on how the documents describing the systems were written-up and disseminated was sought, such as whether the authors of

the systems described how they planned to address criticism and feedback for the system or whether the system was available on an open-access website (see Appendix 2 for the data extraction template).

For quality assurance, the DPhil candidate and a project co-investigator extracted information about the content, development and dissemination of the included systems into a Microsoft Excel spreadsheet. Three independent reviewers (the DPhil candidate and 2 co-investigators of the project) piloted and revised the data extraction form on the same evidence rating system before continuing the extraction with the remaining systems.

Data synthesis

The review employed a four-step procedure to describe the domains of evidence for rating the certainty of evidence in the included systems. First, an inventory of all identified domains was created by using cross-case tables. These tables were examined to compare how the domains for rating the certainty of evidence were labelled, defined, and operationalised across included systems (Miles & Huberman, 1994). The cross-case tables also outlined how each system defined the construct of "certainty of evidence". After examining these tables, a discrete (i.e., nonredundant) list of domains of evidence was compiled along with their definitions. The systems used different terminology to denote similar constructs and domains of evidence (e.g., aspects of the domain that is referred to as "imprecision" in the GRADE approach (Guyatt, Oxman, Kunz, Brozek, et al., 2011) were covered by "precision" in the system adopted by the Agency for Healthcare Research and Quality (AHRQ; Berkman et al., 2013) and fell under the domain termed "clinical impact" in the system used by the National Health and Medical Research

Council (NHMRC) of Australia (Hillier et al., 2011)). Where such overlap existed, the review followed the terminology of the GRADE approach to describe the discrete set of domains. This was supplemented with a list of additional domains that are not currently considered in the GRADE approach, but were found in other systems. Terminology used in the systems was followed to describe these domains.

To help readers visualise findings, a heat map was developed summarising how the systems reported the identified discrete domains of evidence. By using different colour shades, the heat map describes whether these domains of evidence are reported in each included system or not. If the system reported the domain, however, did not provide specific criteria and guidance for rating that domain, the map denotes those as a different category of reporting (i.e., with a brighter shade). Similar to this, a second heat map was developed describing how authors reported activities underpinning the development and dissemination of the included systems. Both of these heat maps were developed by the DPhil candidate and further verified by a second reviewer (a co-investigator of the project). Finally, to identify the extent to which the included systems incorporated a complexity perspective, the identified discrete set of domains were further examined in relation to the sources of complexity outlined in Table 3.1 (Petticrew et al., 2013).

Results

The review identified 11,758 records after duplicates were removed. After title and abstract screening, the DPhil candidate assessed the full-texts of 141 records, from which 28 records were found to be eligible for inclusion in this review. Overall, these 28

records describe 17 evidence rating systems (see Figure 3.1 for the PRISMA flow diagram).

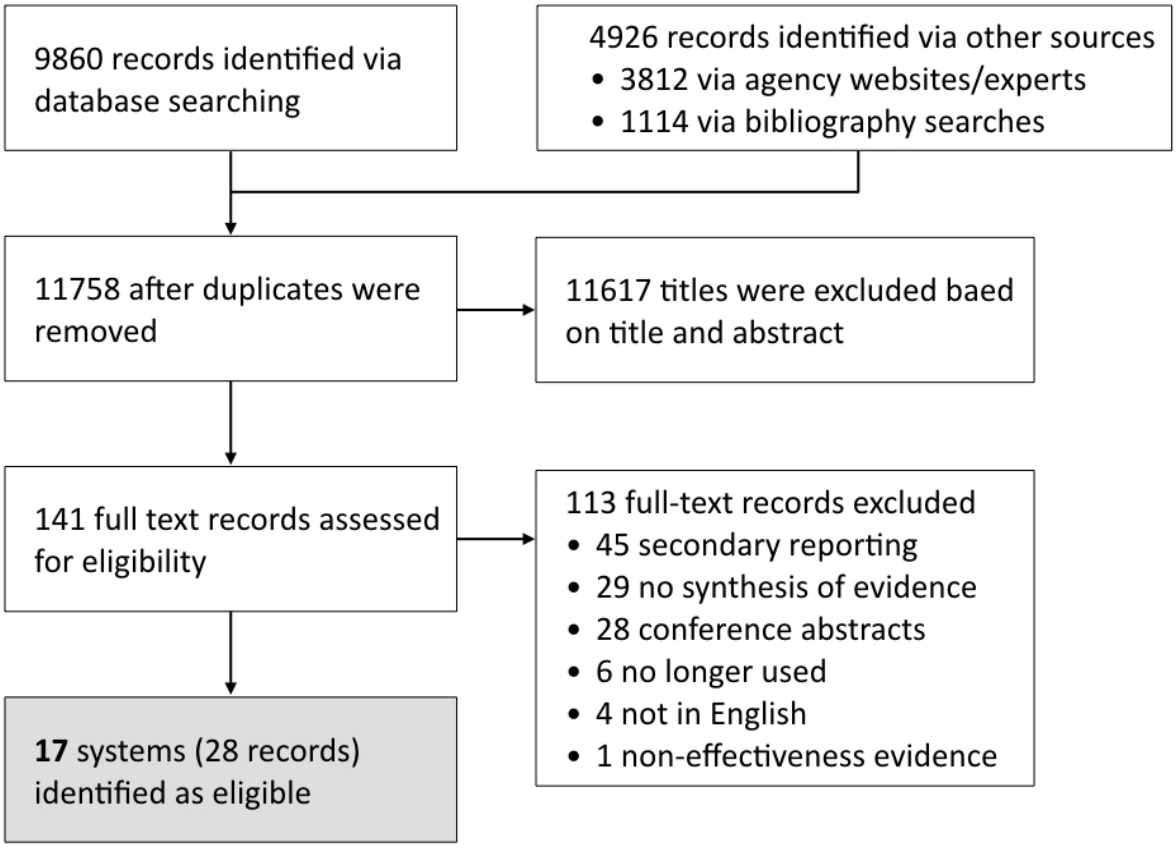


Figure 3.1. Systematic review PRISMA flow diagram

Excluded studies

Of the 113 records excluded at full text screening, 45 involved literature reviews of evidence rating systems, 28 were editorials or conference abstracts, and 4 records were not in English (Chinese, French, Portuguese, and Spanish). Twenty-nine records described procedures and domains for categorising interventions on websites of different "what works" organisations, also known as evidence clearinghouses or evidence-based programme registers (Burkhardt, Schroter, Magura, Means, & Coryn, 2015). Because these procedures and corresponding domains of evidence did not

consider a "certainty of evidence" as defined in this review, they were also excluded from this review (see Appendix 3 for the full list of these systems and their specific domains). Through website searches and contacts with experts, 6 systems were identified, which have been superseded by either an updated version (Weightman, Ellis, Cullum, Sander, & Turley, 2005) or another system (Eccles, Freemantle, & Mason, 1998; Greer, Mosser, Logan, & Halaas, 2000; Guyatt et al., 1995; Harbour & Miller, 2001; Liddle, Williamson, & Irwig, 1996). A further system, namely the Confidence in the Evidence from Reviews of Qualitative Research (CERQual), which is designed for sole application to a body of qualitative evidence, was not eligible for use in assessment of effectiveness evidence (Lewin et al., 2015).

Characteristics of the included studies

Fourteen of the included systems were developed for healthcare, including general clinical and public health interventions (see Table 3.2). Only 3 systems were developed for other policy domains—specifically, education, criminology and international development (DFID, 2014; Gough, 2007; Johnson, Tilley, & Bowers, 2015). Three of the included systems were largely based on the GRADE approach (Guyatt et al., 2011), but introduced modifications that warrant their classification as separate systems (Berkman et al., 2013; Bruce et al., 2014; JBI, 2014). Ten systems mentioned specific research synthesis methods for which the system was developed; most referred to a meta-analysis or a "narrative synthesis" without a single pooled effect estimate (Popay et al., 2006) to synthesise data on the effects of an intervention. Only 1 system developed in criminology was explicitly described for use with a mixed-method approach to evidence synthesis (Johnson et al., 2015). Eight of the systems described rating the

certainty of evidence within the context of systematic reviewing only, while 8 others described rating the certainty of evidence for a guideline development context. Only the GRADE approach addressed the conceptual and procedural differences of rating the certainty of evidence in systematic reviews versus guideline development contexts (Guyatt et al., 2011). By way of illustration, the GRADE approach specifies 2 definitions for the construct of "certainty of evidence" for a systematic review and a guideline development context (see section 3.3 below).

Inconsistencies were identified in how included systems labelled and defined rating of the certainty of evidence, overall, and the components of that rating, more specifically. The most frequently used terms to describe the overall rating of certainty of evidence were *strength of evidence*, *grades of evidence, quality, confidence,* or *certainty of evidence* (Berkman et al., 2013; Briss et al., 2000; Clark, Burkett, & Stanko-Lopp, 2009; DFID, 2014; Ebell et al., 2004; Guyatt et al., 2011; NICE, 2012; Sawaya et al., 2007). In contrast, the most commonly used terms for assessing the conduct of individual studies were *levels of evidence, critical appraisal, quality appraisal, study limitations, risk of bias,* and *study quality* (Clark et al., 2009; DFID, 2014; Ebell et al., 2004; Hillier et al., 2011). From these, terms such as *levels of evidence*, *risk of bias* and *study limitations* were mainly discussed with regard to assessing studies for bias and internal validity, while *study quality*, *quality appraisal and critical appraisal* were used to denote study execution more broadly with regard to eliminating threats to both internal and external validities.

Defining certainty of evidence

Only 6 systems—3 of which are largely based on the GRADE approach—provided a definition for the construct of "certainty of evidence" on intervention effectiveness (Berkman et al., 2013; Briss et al., 2000; Bruce et al., 2014; Guyatt et al., 2011; JBI, 2014; Sawaya et al., 2007). In a systematic review context, the GRADE approach and 3 other systems based on it defined certainty of evidence as *"the extent of confidence that an estimate of the effect is correct"* (Berkman et al., 2013; Guyatt et al., 2011; JBI, 2014). The Guide to Community Preventive Services defined certainty of evidence as *"confidence that changes in outcomes are attributable to the interventions"* (Briss et al., 2000), and the U.S. Preventive Services Task Force (USPSTF) — as the *"likelihood that the assessment of the net benefit (i.e., benefits minus harms) of a preventive service is correct"* (Sawaya et al., 2007). The USPSTF definition is similar to how the GRADE approach defines the overall certainty of evidence in the context of guideline development, when considering all important outcomes associated with the intervention, including harms. In this context GRADE defines the overall certainty of evidence as *"the extent of confidence that an estimate of the effect is adequate to support a particular decision or recommendation"*(Guyatt et al., 2013).

In order to assess the net benefit of a preventive service, the USPSTF system uses analytic frameworks, also called "chain of evidence" diagrams to map out the specific linkages in the overall chain of evidence that must be present for a preventive service to be considered effective (Sawaya et al., 2007). The system assesses the certainty of evidence for each separate linkage in the chain of evidence to draw conclusions about the overall effectiveness of a preventive service (see Figure 3.2 for an example of an analytic framework). This approach is very similar to that adopted by the GRADE-

modified Grading of Evidence for Public Health Interventions (GEPHI) system (Bruce et al., 2014). In addition to rating the certainty of evidence for the estimates of the effect of an intervention (which corresponds to the approach described in GRADE), the GEPHI system suggests to also rate the evidence for the overall causal chain of an intervention. This rating of the confidence in the overall causal chain of an intervention is referred to as "coherence of evidence" assessment in the GEPHI system (Bruce et al., 2014).

Mapping of evidence domains

The evidence domains used to rate the certainty of evidence were often similar in concept across systems yet different in how they were described and operationalised. Table 3.2 and Figure 3.3 should, therefore, be used as 2 complementary sources of information on the identified evidence rating systems to examine the discrepancies in labelling and describing those domains. Table 3.2 provides an overview of the domains of evidence as they are reported in the original studies, while Figure 3.3 maps the thirteen discrete domains identified in included systems and presents how they are reported in each of the included systems. More information on how the specific evidence domains were defined and operationalised in each system is presented in Appendix 4. The sections below, briefly summarise the identified discrete set of domains of evidence (see Figure 3.3), as well as the extent to which they address sources of complexity in systematic reviews, and whether the authors of these systems report recommended activities underpinning the development and dissemination of these domains (see Figure 3.4).

Study design

Twelve systems included an evidence domain related to the design of the individual studies constituting the body of evidence. All but 4 of these systems (Gough, 2007; Johnson et al., 2015; NICE, 2012; Turner-Stokes, Harding, Sergeant, Lupton, & McPherson, 2006) described an "evidence hierarchy" approach that influenced how the overall certainty of evidence was assessed. Procedurally, this entailed initially privileging a body of evidence from certain study designs (namely RCTs) as providing a higher certainty of evidence (compared with other study designs) before assessing other evidence domains. While all systems with an evidence hierarchy approach placed evidence from RCTs at the top of this hierarchy, many systems further privileged specific nonrandomised study designs over others. For example, the system used by the Joanna Briggs Institute (JBI, 2014) suggests initial ratings of the certainty of evidence depending on whether a body of evidence consists of experimental (Level 1), quasi-experimental (Level 2), or observational studies (Level 3). Similarly, the GRADE-modified GEPHI system for public health interventions recommends that a body of evidence consisting of nonrandomised studies with controls or before and after [uncontrolled] studies has an initial rating of "moderate" certainty, if included studies use appropriate methods to minimise selection bias and confounding (Bruce et al., 2014).

Table 3.2. Overview of the included evidence rating systems

First author (year) *Name of the system/organisation*	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Baral (2012) ***The Highest Attainable Standard of Evidence (HASTE)*** *"…focuses on triangulation of 3 distinct categories of evidence" (p. 572)*	1. Efficacy data 2. Implementation data[a] 3. Plausibility data[a]	1. Consistent; inconsistent; limited 2. Consistent; inconsistent; limited[a] 3. High; low; undefined[a]	• Grade 1 (strong) • Grade 2 (conditional) • Grade 2a (probable) • Grade 2b (possible) • Grade 2c (pending) • Grade 3 (insufficient) • Grade 4 (inappropriate)	Not specified	Guideline development in public health (specific focus on HIV/AIDS interventions)
Berkman (2013) ***Agency for Healthcare Research and Quality (AHRQ)*** *"…confidence in systematic review conclusions so that decision-makers can use them effectively" (p. 1)*	1. Study design 2. Study limitations 3. Directness 4. Consistency 5. Precision 6. Reporting bias 7. Dose-Response 8. Plausible confounding 9. Magnitude of effect 10. Applicability[a]	1. High (RCTs); Low (non-RCTs) 2. Risk of bias in RCTs/non-RCTs 3. Divergence from the outcomes & comparisons of interest 4. Consistency in magnitude or direction 5. Sample size, width of 95% CI 6. Publication bias; selective outcome reporting bias; selective analysis reporting 7. Dose-response relationship 8. Counteracting confounding 9. Size of the estimate of the effect 10. Likelihood of expected results under the "real-world" conditions[a]	• High • Moderate • Low • Insufficient	Quantitative: meta-analysis or narrative synthesis	Evidence synthesis in clinical medicine
Briss (2000) ***The Guide to Community Preventive Services*** *"…confidence that changes in outcomes are attributable to the interventions" (p. 38)*	1. Design suitability 2. Quality of study execution 3. Number of studies 4. Consistent 5. Effect size 6. Applicability[a] 7. Barriers to implementation[a] 8. Economic evaluations[a] 9. Other effects[a]	1. Greatest (concurrent comparison); moderate (comparison, but not concurrent); least (single group) 2. 6 categories of threats to validity: good, fair or limited 3. – 4. Consistent in direction & size 5. Defined on a case-by-case basis 6. Applicability to local situations[a] 7. – 8. – 9. Evidence on harms[a]	• Strong • Sufficient • Insufficient	Quantitative: meta-analysis or narrative synthesis	Guideline development in public health

Table 3.2. (Continued)

First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Bruce (2014) ***Grading of Evidence for Public Health Interventions (GEPHI)*** *"…it is useful to make a distinction between: (a) strength of evidence for causal inference, for which Bradford Hill viewpoints for distinguishing causation from association in environmental epidemiology are often referred to…, and (b) the quality of evidence for the intervention effect size (confidence in the estimate), for which GRADE may be used" (p. 11)*	1. Study design 2. Study limitations 3. Indirectness 4. Inconsistency 5. Imprecision 6. Reporting bias 7. Dose-response 8. Plausible confounding 9. Magnitude of effect 10. Analogy 11. Consistency 12. Coherence	1. High (RCTs); Moderate (quasi-experimental designs); Low (other observational designs) 2. Risk of bias in RCTs/non-RCTs 3. Divergence from the PICO elements 4. Heterogeneity in the effect estimates 5. Sample size, width of 95% CI 6. Failure to identify studies 7. Dose-response relationship 8. Counteracting plausibility 9. Size of the estimate of the effect 10. Supporting evidence with similar mechanisms 11. Consistent evidence across different settings 12. Coherence in the overall causal chain: high, moderate, weak (separate rating)	• High • Moderate • Low • Insufficient	Quantitative: meta-analysis or narrative synthesis	Guideline development in public health
Clark (2009) ***Let Evidence Guide Every New Decision (LEGEND)*** *"The term 'level' was important to nurses to indicate the quality of an individual article; while 'grade' was more familiar to doctors and was adopted to indicate the quality of a body of evidence" (p. 1057)*	1. Study quality 2. Consistency 3. Number of studies	1. The aggregate quality ratings for individual studies (categorised based on an evidence hierarchy; e.g. 1a, good quality systematic review, 1b, lesser quality systematic review, 2a, good quality RCT/CCT; 2b, lesser quality RCT/CCT; etc.) 2. The extent to which similar findings are reported: yes; no; not available 3. –	• High • Moderate • Low • Grade-Not-Assignable	Not specified	Guideline development in clinical medicine
DFID (2014) ***How to Note*** *"This Note assumes that the overall 'strength' of a body of evidence is determined by the "avoidance of bias" of studies that constitute it, and by the size, context and consistency" (p. 3)*	1. Quality 2. Size of the body of evidence 3. Consistency 4. Context of the body of evidence	1. Assessed regarding 7 domains: high; moderate; low 2. Large; medium small 3. Consistent; inconsistent; mixed 4. Global; context-specific	• Very Strong • Strong • Medium • Limited • No evidence	Not specified	Evidence synthesis in international development

Table 3.2. (Continued)

First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Ebell (2004) ***Strength of Recommendation Taxonomy (SORT)*** *"We use the term level of evidence to refer to individual studies. The strength (or grade) of a recommendation for clinical practice is based on a body of evidence" (p. 60)*	1. Study quality 2. Consistency	1. Study quality (combined with study design considerations based on an evidence hierarchy) 2. Consistent; inconsistent	• Level 1 (good quality) • Level 2 (limited quality) • Level 3 (other evidence)	Quantitative: meta-analysis	Guideline development in clinical medicine
Gough (2007) ***Weight of Evidence: a framework for the appraisal of the quality and relevance of evidence*** *"Weight of evidence is a useful heuristic for considering how to make separate judgements on different criteria" (p. 11)*	1. Relevance of research design 2. Study execution 3. Relevance of the focus/context of evidence	1. A review-specific judgement about the appropriateness of that form of evidence for answering the review question 2. Generally accepted criteria for evaluating the quality of evidence 3. A review specific judgement about the relevance of the focus of the evidence for the review question	• Weight of evidence A • Weight of evidence B • Weight of evidence C • Weight of evidence D	Not specified (different quantitative and qualitative approaches)	Evidence synthesis in education
Guyatt (2011) GRADE ***Grading of Recommendations Assessment, Development and Evaluation (GRADE)*** *"The extent to which one can be confident that the estimates of the effect are correct" (p. 394)*	1. Study design 2. Study limitations 3. Indirectness 4. Inconsistency 5. Imprecision 6. Publication bias 7. Dose-response 8. Plausible confounding 9. Magnitude of effect	1. High (RCTs); low (non-RCTs) 2. Risk of bias in RCTs/non-RCTs 3. Divergence from the PICO elements 4. Heterogeneity in effect estimates 5. Sample size, width of 95% CI 6. Failure to identify studies 7. Dose-response relationship 8. Counteracting confounding 9. Size of the estimate of the effect	• High • Moderate • Low • Very low	Quantitative: meta-analysis or narrative synthesis	Evidence synthesis & guideline development in clinical medicine (also mentions potential applicability to public health)
Hillier (2011) ***FORM: An Australian method for formulating and grading recommendations in evidence-based guidelines*** *"…considering all of these elements across all of the research studies addressing the clinical question as a whole (the 'body of evidence')" (p. 2)*	1. Evidence base 2. Consistency 3. Clinical Impact 4. Applicability 5. Generalisability	1. Quality; quantity and study design (level I, SR of RCTs; level II, RCTs, level III-1, pseudorandomised trial; level III-2, comparative study with concurrent control; level III-3, comparative study without concurrent controls) 2. Excellent; good; poor 3. Very large; substantial; moderate; slight 4. Excellent; good; satisfactory; poor 5. Excellent; good; satisfactory; poor	• A (evidence trusted) • B (evidence mostly trusted) • C (some support) • D (weak evidence)	Not specified	Guideline development in clinical medicine

Table 3.2. (Continued)

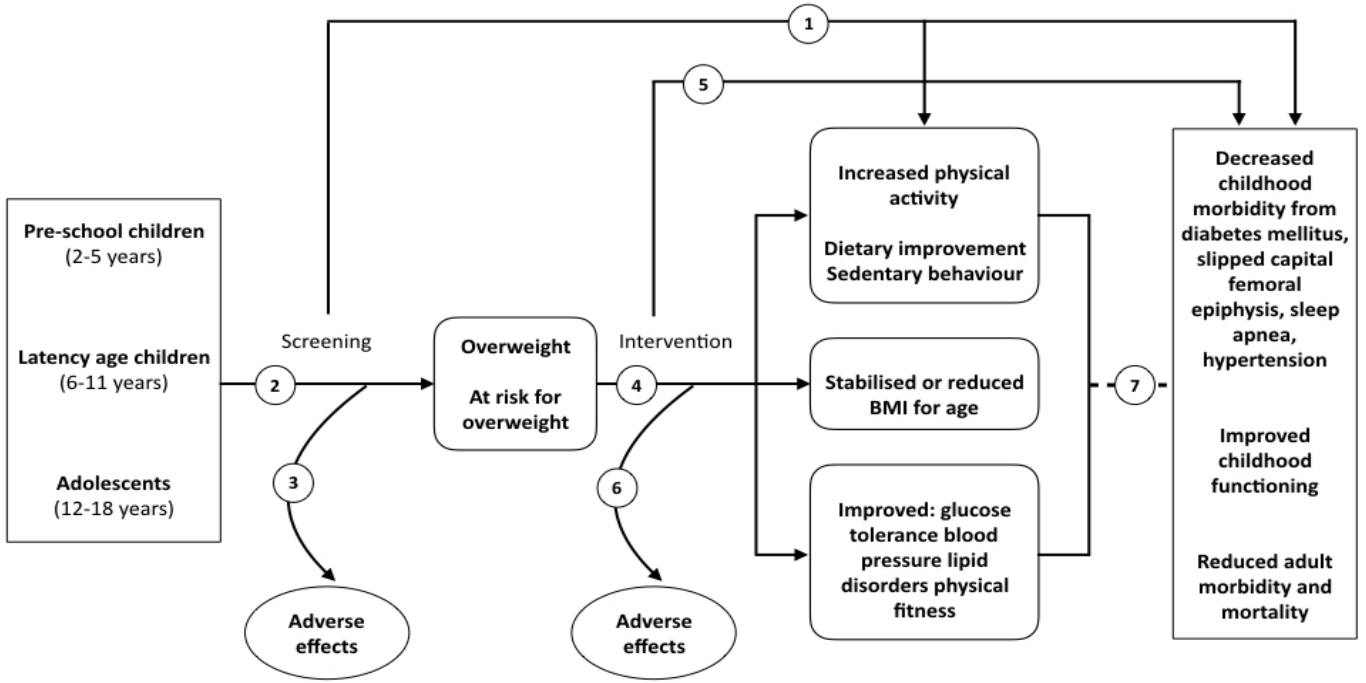
First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Joanna Briggs Institute (2014) ***Levels of evidence and grades of recommendations*** *"One of the main reason for continuing with Levels of Evidence system is to assist in assigning GRADE pre-rankings…" (p. 4)*	1. Study design 2. **The remaining domains of evidence follow those of the GRADE approach**	1. Level 1: experimental; level 2: quasi-experimental; level 3: Observational-Analytic; level 4: Observational-Descriptive; level 5: Expert opinion	• High • Moderate • Low • Insufficient	Quantitative: meta-analysis or narrative synthesis	Evidence synthesis in clinical medicine & public health
Johnson (2015) ***Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis*** *"…in addition to considering the extent to which evaluations manage to rule out biases that might distort estimates of effect size, we also need to gauge the extent to which they contribute to understanding of the contexts/moderators" (p. 462)*	1. Effects 2. Mechanisms/mediators[a] 3. Moderators/contexts[a] 4. Implementation[a] 5. Economic analysis[a]	1. Consideration of evidence validity elements 2. Reference to and/or test of theory of change[a] 3. Reference to and/or analysis of data relating to pre-defined moderators[a] 4. Account of implementation or implementation challenges[a] 5. Estimation of marginal, total or opportunity costs[a]	Promotes descriptive profiles rather than a single overall score	Mixed-method synthesis	Evidence synthesis in criminology
NICE (2012) ***Methods for the development of NICE public health guidance*** *"strength of evidence – reflecting the appropriateness of the study design to answer the question and the quality, quantity and consistency of evidence" (p. 89)*	1. Study design 2. Quality 3. Quantity 4. Consistency 5. Direction of the effect[a] 6. Size of the effect[a] 7. Applicability[a]	1. Appropriateness to answer the question 2. Assessment of both internal and external validity 3. – 4. – 5. Positive; negative; mixed; none[a] 6. Small; medium; large[a] 7. Applicability of evidence to PICO[a]	• No evidence • Weak evidence • Moderate evidence • Strong Evidence • Inconsistent Evidence	Quantitative: meta-analysis or narrative synthesis Qualitative for questions other than intervention effectiveness	Guideline development in public health
Sawaya (2007) ***U.S. Preventive Services Task Force (USPSTF)*** *"The U.S. Preventive Services Task Force (USPSTF) defines certainty as "likelihood that the USPSTF assessment of the net benefit of a preventive service is correct" (p. 873)*	1. Study Design 2. Study Quality 3. Generalisability 4. Quantity 5. Consistency 6. Other	1. Level I, RCT; level II-1, controlled trial without randomisation; level II-2, cohort and case-control; level II-2, multiple time series; level III, opinions) 2. Design specific: good; fair; poor 3. – 4. – 5. – 6. Dose-response; fit within a biological model	Chain of evidence: • High • Moderate • Low	Not specified	Guideline development in clinical medicine

Table 3.2. (Continued)

First author (year) Name of the system/organisation	Domains of evidence	Notes on the domains of evidence	Evidence ratings	Evidence synthesis approach	Context of application
Tang (2008) ***Grading of evidence of the effectiveness of health promotion interventions*** *"…the strength of evidence can be graded by using 3 criteria" (p. 832)*	1. Association 2. Repeatability 3. How it works	1. High (risk ratio of 2 or more) and statistically significant association: high, low, none 2. Reflects the consistency of the findings in different settings: wide, limited, none 3. Reflects the known cause-effect mechanism for the intervention under study: known; not known	• Grade 1 (strong) • Grade 2A (probable) • Grade 2B (possible) • Grade 2C (limited) • Grade 3 (insufficient)	Not specified	Evidence synthesis in public health
Treadwell (2006) ***A system for rating the stability and strength of medical evidence*** *"Our system draws a distinction between 2 types of conclusions: quantitative and qualitative…a quantitative conclusion characterises the size of the effect, whereas a qualitative conclusion characterises the direction of the effect" (p. 6)*	1. Quality 2. Quantity 3. Informativeness 4. Homogeneity 5. Robustness	1. High; moderate; low; very low 2. Criterion met; criterion not met (at least 3 studies and 80% having calculable effect sizes) 3. Effect size 4. Homogeneous; heterogeneous 5. Tested through sensitivity analysis: robust; not robust	• Strong • Moderate • Weak • Inconclusive	Quantitative: meta-analysis	Evidence synthesis in clinical medicine
Turner-Stokes (2006) ***Generating the evidence base for the National Service Framework for long term conditions: a new research typology*** *"Each individual recommendation is then given an overall 'grade of research evidence' rating of A, B or C based on the quality of all the evidence supporting it and how much of it was directly relevant" (p. 97)*	1. Type of evidence 2. Study quality 3. Applicability	1. Primary research-based; secondary research-based; review-based (no classification based on an evidence hierarchy) 2. Quality is assessed on the basis of 5 questions to reach a max. score of 10 (includes a question on the appropriateness of the study design) 3. Population context of the study: direct; indirect	• GRADE A • GRADE B • GRADE C	Quantitative: meta-analysis or narrative synthesis Qualitative for questions other than intervention effectiveness	Evidence synthesis in clinical medicine

[a]These domains of evidence go beyond rating the certainty of evidence on intervention effectiveness and are used in systems to further inform grading of the recommendations for practice. In the GRADE approach, these domains are separately specified as "Evidence to Decision" criteria (see Alonso-Coello et al., 2016).

Notes: CCT: Controlled Clinical Trial; CI: Confidence Interval; DFID: Department for International Development; NICE: National Institute for Health and Care Excellence; PICO: Population, Intervention, Comparison, Outcomes; RCT: Randomised Controlled Trial; SR – Systematic Review.



Arrow 1: Is there direct evidence that screening (and intervention) for overweight in childhood improves age-appropriate behavioural or physiologic measures, or health outcomes?

Arrow 2: **a.** What are appropriate standards for overweight in childhood and what is the prevalence of overweight based on these?
b. What clinical screening tests for overweight in childhood are reliable and valid in predicting obesity in childhood?
c. What clinical screening tests for overweight in childhood are reliable and valid in predicting poor health outcomes in adulthood?

Arrow 3: What are the adverse effects of screening, including labelling? Is screening acceptable to patients?

Arrow 4: **a.** Do weight control interventions lead to improved intermediate outcomes?
b. What are common behavioural and health system elements of efficacious interventions?
c. Are there differences in efficacy between patient subgroups?

Arrow 5: Do weight control interventions lead to improved health outcomes and/or improved functioning?

Arrow 6: What are the adverse effects of interventions? Are interventions acceptable to patients?

Arrow 7: Are improvements in intermediate outcomes associated with improved health outcomes? (only evaluated if there is no direct evidence for link 1 or link 5 and if there is sufficient evidence for link 4)

Figure 3.2. Screening and interventions for overweight in childhood: analytic framework and evidence links (Whitlock, Williams, Gold, Smith, & Shipman, 2005)

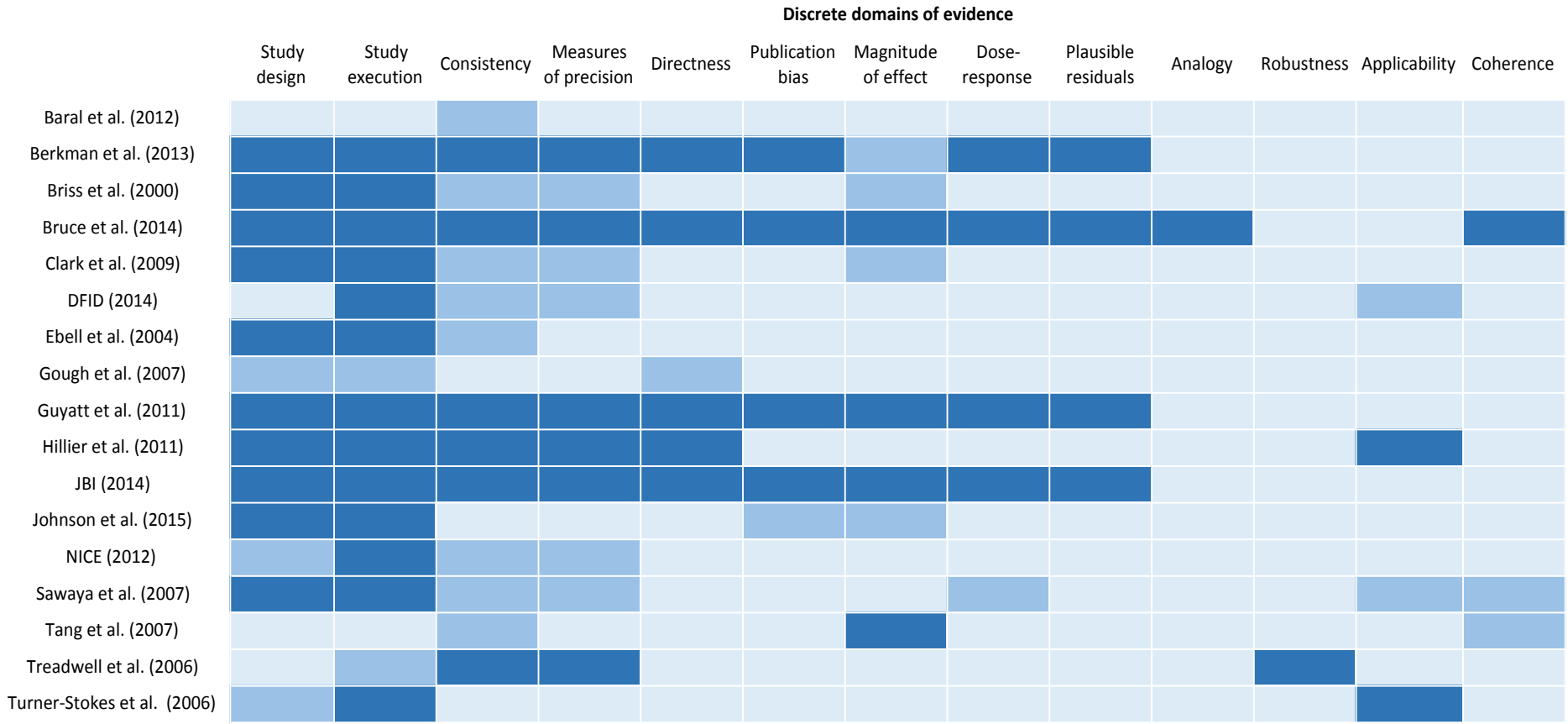


Figure 3.3 Reporting of the domains of evidence for rating the certainty of evidence on intervention effectiveness

Notes: DFID: Department for International Development; JBI: Joanna Briggs Institute; NICE: National Institute for Health and Care Excellence

1. Reported
2. Reported, but does not describe specific criteria or guidance for assessing the domain
3. Not reported

Figure 3.4. Reporting of the activities for developing and disseminating the evidence rating systems

Notes: DFID: Department for International Development; JBI: Joanna Briggs Institute; NICE: National Institute for Health and Care Excellence

Study execution

Fifteen systems included an evidence domain related to assessing how well studies constituting the body of evidence were executed to minimise threats to internal and/or external validities (also labelled as "quality of study execution", "risk of bias", "study limitations", and "study quality"). In most instances, systems mainly included criteria to assess risks of bias or threats to the internal validity for assessing study execution. A few systems, however, also included specific criteria for assessing the generalisability of the study results, that is, criteria related to the external validity of the individual study.

Systems varied in how they operationalised assessment of study execution. Some systems used design-specific criteria, such as checklists or signalling questions for appraising RCTs (Berkman et al., 2013; Clark et al., 2009; Guyatt et al., 2011; Johnson et al., 2015; Viswanathan et al., 2012) or longitudinal studies (Clark et al., 2009). Most systems, however, described more generic criteria to assess study execution across various study designs included in the body of evidence (Briss et al., 2000; DFID, 2014; Hillier et al., 2011; NICE, 2012; Turner-Stokes et al., 2006; Zaza et al., 2000).

Consistency

Fourteen systems included an evidence domain related to the consistency of evidence. Generally, systems defined consistency as *"the extent to which findings are similar across included studies"* in a body of evidence (Hillier et al., 2011), usually in reference to the degree of similarity in the magnitude and/or direction of effect estimates. Most systems, however, did not report any specific criteria on how to rate

consistency in the body of evidence. Only a few systems discussed specific procedures, such as statistical testing for heterogeneity to rate consistency in the body of evidence.

The GRADE-modified GEPHI approach distinguished between 2 types of consistency ratings (Bruce et al., 2014): The first type was identical to the domain of the GRADE approach termed as "inconsistency" and was defined as *"assessment of statistical heterogeneity in the estimates of the effect"* (Guyatt, Oxman, Kunz, Brozek, et al., 2011). The second type of consistency rating was specified in the system as "consistency" assessment and was defined as a presence of *"consistent evidence across a large number of settings, geographical locations and diverse epidemiological study designs".* The system argued that the fact that an intervention effect is reproducible under highly variable conditions suggests reduced likelihood that the observed effect is attributable to confounding or bias. This was viewed to increase a reviewer's confidence in the body of evidence regarding the overall effect of an intervention (Bruce et al., 2014).

Measures of precision

Eleven systems included an evidence domain that this review has classified as relating to measures of precision of the body of evidence: i.e., considerations of the impact that random error may have on effect estimates. Systems differed widely in the level of specification and sophistication they required for assessing precision of the body of evidence. For instance, many systems recommend only considering the number of studies in the body of evidence as a measure of precision (Briss et al., 2000; Clark et al., 2009; DFID, 2014; NICE, 2012; Treadwell, Tregear, Reston, & Turkelson, 2006); however, only 1 of these systems specifies a threshold for the minimum number of studies to be included in the body of evidence (Treadwell et al., 2006). Furthermore, only the GRADE

approach and its variants described specific criteria for assessing precision regarding the sufficiency of the sample size of the body of evidence (Berkman et al., 2013; Bruce et al., 2014; Guyatt et al., 2011; JBI, 2014). These systems assessed sufficiency of the sample size relative to an "optimal information size": i.e., *"a number of patients (for continuous outcomes) and events (for dichotomous outcomes) that would be needed to regard a body of evidence as having adequate power"*(Guyatt, Oxman, Kunz, Brozek, et al., 2011). In addition, these systems also considered the boundaries of confidence intervals for an effect estimate in relation to a null effect and a clinically important effect threshold to make an overall judgement about the precision of a body of evidence. The estimate of the effect of an intervention is judged to be less precise, if the confidence interval is wide to include a null effect or a threshold, which is considered as clinically unimportant (Guyatt, Oxman, Kunz, Brozek, et al., 2011).

Directness

In general, the systems used concepts of "directness", "applicability" and "generalisability" of evidence interchangeably and inconsistently—often without providing clear definitions or specific criteria to guide the assessment (DFID, 2014; Gough, 2007; Hillier et al., 2011; Sawaya et al., 2007; Turner-Stokes et al., 2006). In addition, these terms were not necessarily used as synonyms across the systems. For example, the system endorsed by the National Health and Medical Research Council (NHMRC) of Australia used the term "applicability" to address whether the body of evidence was relevant to the local context (including the organisational and cultural contexts), while the term "generalisability" was used to refer to how precisely a body of evidence answered a review or a guideline question in terms of populations and settings

of interest (Hillier et al., 2011). To disentangle the discrepancies in the terminology, this review used the terminology of the GRADE approach, namely "directness" of evidence to describe the domains of evidence from the included systems related to the notion of comparability of the evidence to the original research question. Six systems were identified that used this domain of evidence to assess how directly the available evidence answers a review or a guideline question regarding Population, Intervention, Comparison and Outcomes (PICO) elements of the question (Berkman et al., 2013; Briss et al., 2000; Gough, 2007; Guyatt, Oxman, Kunz, Woodcock, et al., 2011; Hillier et al., 2011; JBI, 2014).

Publication bias

Five systems included publication bias as a domain for rating the certainty of evidence on intervention effectiveness (Berkman et al., 2013; Bruce et al., 2014; Guyatt, Oxman, Montori, et al., 2011; JBI, 2014; Johnson et al., 2015). All of these systems, except one, followed a definition of publication bias used within the GRADE approach, that is, "a failure to identify studies as a result of studies remaining unpublished or obscurely published" (Guyatt, Oxman, Montori, et al., 2011). The system used by AHRQ on the other hand, considered publication bias as only one type of potential bias within a broader domain of reporting biases, which was itself defined as a decision by authors or journals to report research findings based on their direction and magnitude of effect (Berkman et al., 2013). Selective outcome reporting and selective analysis reporting were the other types of reporting biases described in this system.

Magnitude of effect

This review identified 7 systems, which included "magnitude of effect" as a distinct domain to rate the certainty of evidence on intervention effectiveness (Berkman et al., 2013; Briss et al., 2000; Bruce et al., 2014; Clark et al., 2009; Guyatt, Oxman, Sultan, et al., 2011; JBI, 2014; Johnson et al., 2015; Tang, Choi, & Beaglehole, 2008). However, only 4 of these systems specified the thresholds for what they considered to be a "large" magnitude of effect (Bruce et al., 2014; Guyatt, Oxman, Sultan, et al., 2011; JBI, 2014; Tang et al., 2008).

Dose-response

Overall, 5 systems considered dose-response as a distinct domain of evidence, when rating the certainty of evidence on intervention effectiveness (Berkman et al., 2013; Bruce et al., 2014; Guyatt, Oxman, Sultan, et al., 2011; JBI, 2014; Sawaya et al., 2007). The systems commonly defined dose-response as a *"pattern of a larger effect with greater exposure to an intervention"* (Berkman et al., 2013).

Plausible residuals

All systems that followed the structure of the GRADE approach (overall 4 systems, including GRADE itself) considered counteracting confounding, as a domain to upgrade the certainty of evidence, when a body of evidence is mainly composed of observational studies (Berkman et al., 2013; Bruce et al., 2014; Guyatt, Oxman, Sultan, et al., 2011; JBI, 2014). Two possibilities were commonly applied: *"If all plausible residual biases would diminish the observed effect, or if all plausible residual biases would suggest a spurious effect when no effect is observed"* (Guyatt, Oxman, Sultan, et al., 2011).

Analogy

Only 1 system, the GEPHI system, included an evidence domain related to analogous evidence. The GEPHI system operationalised analogous evidence as supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanisms, which, if present, could lead to a higher certainty of evidence rating (Bruce et al., 2014). In the context of the WHO guidelines on indoor air quality, the system discusses the example of how the certainty in the effects of household air pollution from solid fuel can be enhanced by strong empirical evidence about the effects of second-hand or active smoking. In this example, both household air pollution and second-hand or active smoking are seen to expose individuals to similar combustion mixtures, and, therefore, can be viewed as analogous pieces of evidence (Bruce et al., 2014).

Robustness

Robustness of evidence was described as a domain to rate the certainty of evidence on intervention effectiveness by 1 system (Treadwell et al., 2006). The system suggests that reviewers measure robustness of evidence through sensitivity analysis with a pre-specified definition for "robust" effects. A priori specifications can guard against biases involved in judgements during a review process. For example, a review author may decide a priori that a threshold for robustness assessment is one in which *"confidence intervals of the last 3 cumulative, random-effects meta-analyses remain fully on the same side of zero after removing of the study with the smallest weight"* (Treadwell et al., 2006).

Applicability

Four systems described domains of evidence when rating the certainty of evidence on intervention effectiveness, which can be related to the assessment of the applicability of the available evidence regarding the context of evidence use (DFID, 2014; Hillier et al., 2011; Sawaya et al., 2007; Turner-Stokes et al., 2006). It is worth highlighting that 3 additional systems were identified (Berkman et al., 2013; Briss et al., 2000; NICE, 2012), which considered applicability as a separate judgement when making recommendations for practice. In these systems, discussion of applicability was held separately from other domains of evidence, and largely within a context of guideline development. For example, the GRADE-based system endorsed by AHRQ clearly separates judgements of directness of evidence from that of applicability assessment. In this system, directness of evidence is defined to express "how closely available evidence measures an outcome of interest", and relies on 2 judgements (Berkman et al., 2013): the directness of the employed outcomes (i.e., whether the available evidence is in fact only a proxy for an ultimate outcomes of interest) and the directness of the comparisons (i.e., whether evidence derives from head-to-head comparisons). Meanwhile, the system defines applicability as the external validity of the evidence base regarding different populations, and is considered explicitly, but separately from the overall certainty of evidence rating (Berkman et al., 2013).

Coherence of the causal pathway

Only 3 systems included an evidence domain related to assessing the coherence of the causal pathway of an intervention (Bruce et al., 2014; Sawaya et al., 2007; Tang et al., 2008). This relates to the assessment of a theory of change or a mechanism whereby

an intervention is expected to operate. The GEPHI system recommends assessing confidence in the overall causal pathway (termed as rating of "coherence") between an intervention and distal outcomes regarding the evidence contributing to each individual link in the causal pathway (Bruce et al., 2014). Similarly, by using analytic frameworks, the USPTSF system rates the certainty of evidence in the overall chain of evidence for a specific preventive service (Sawaya et al., 2007). The system described by Tang et al. (2008) included assessment of the known mechanisms of action as a separate domain of evidence for rating the certainty of evidence on intervention effectiveness: "*if the theoretical basis is not known, the strength of evidence will be less convincing"* (Tang et al., 2008).

Sources of complexity in the included systems

In general, this review did not find many instances, when systems attended to the sources of complexity as outlined in Table 3.1. From the sources related to the characteristics of the intervention itself, systems mainly discussed population-level targets of interventions as a dimension that complicates assessment of these interventions through a single study design, namely a randomised controlled trial (this was discussed under the "study design" domain in the included systems). Two systems in social policy made an argument against using "evidence hierarchies" from the field of clinical medicine, which do not differentiate between *"true observational study designs"* (e.g., cohort, case-control and cross-sectional studies) and *"nonrandomised experimental study designs"* (e.g., before-and-after studies and quasi-experimental study designs; (Bruce et al., 2014; JBI, 2014); while 3 systems replaced the evidence hierarchy approach with the assessment of the "design suitability" instead (Briss et al., 2000; Gough, 2007;

NICE, 2012):

> *"The Guide's approach to suitability of study design allows studies to collect either individual or ecological data. This choice was made because many community interventions are applied across populations and could be difficult, impossible, or inappropriate to study with individual-level data"* (Briss et al., 2000, p. 41).

In a few instances, systems in social practice areas reported adapting aspects of the existing methods from clinical medicine to better describe the variety of study designs in areas of social policy:

> *"Actions in the field of environmental health—for practical, ethical, political and cost reasons—are rarely amenable to randomised controlled trials that directly assess the health impacts of interventions. Following the current GRADE approach, a majority of studies providing information on environmental health interventions would, therefore, start off as low-quality … The revisions to GRADE take account of nonrandomised experimental study designs by adding these designs into the modified table at the level of moderate evidence"* (Bruce et al., 2014, p. 9).

Regarding the characteristics of the intervention's causal pathway (see Table 3.1), *"coherence of evidence"* was described in 2 systems as a domain of evidence to address the long and dynamic causal pathways of interventions, as well as multiple mediators and outcomes associated with the intervention (see Figure 3.2; (Bruce et al., 2014; Sawaya et al., 2007). This domain was found useful for conceptualising the effectiveness of social and public health interventions and for enabling integration of different types of evidence when rating the certainty of evidence:

> *"Reducing the complexity of environmental health interventions by picking out a single link in the causal chain (e.g., only considering the direct RCT-evidence, which links the intervention with the distal health outcomes) may under- or overestimate actual effectiveness in the field"* (Bruce et al., 2014, p. 9).

The importance of mediators and moderators of an intervention was also highlighted in the Effect, Mechanisms, Moderators, Implementation, Economic costs

(EMMIE) system, which suggests to equally attend to mediators and moderators along with the overall estimates of the effect of an intervention when rating the certainty of evidence. The authors of this system argue that each dimension described by the acronym EMMIE may speak to a different element of a body of evidence in a systematic review and may inform different stages of the policy-making process. The system suggests to assess the certainty of evidence by way of creating descriptive profiles for each dimension, instead of producing overall ratings (Johnson et al., 2015).

Development and dissemination of the included systems

Figure 3.4 describes how authors report activities underpinning the development and dissemination of the included systems. Regarding the preparatory activities for developing the system, only 4 systems empirically demonstrated the need for developing a new evidence rating system by referring to a separate publication by the same research team providing a critical appraisal of the existing systems (Clark et al., 2009; Guyatt et al., 2011; Hillier et al., 2011; NICE, 2012). More frequently, the systems reported participants involved in the development of the system, which was mainly limited to the immediate author team. Only 4 systems described obtaining funding for developing the system (Hillier et al., 2011; Johnson et al., 2015; Treadwell et al., 2006; Turner-Stokes et al., 2006). None reported conducting a Delphi process to develop the system, and only 5 reported hosting an expert meeting. However, with the exception of the GRADE approach, these systems did not provide further details on how these meetings were organised (Baral et al., 2012; Ebell et al., 2004; Guyatt et al., 2011; Hillier et al., 2011; Turner-Stokes et al., 2006). The GRADE Working Group, on the other hand, organises annual meetings lasting 2 to 3 days, where members of the group have an opportunity to

meet face-to-face and further discuss, develop and refine aspects of the GRADE methodology ("GRADE Working Group", 2017).

Regarding the write-up and dissemination activities, only 3 systems described how the publication introducing the system was developed (Ebell et al., 2004; Hillier et al., 2011; Turner-Stokes et al., 2006), while instructions for using the systems were predominantly described in the same document that introduced the system, and only 7 systems provided examples on how to apply the system in specific interventions. In 6 instances, willingness to incorporate the feedback of users and update the systems was mentioned (Berkman et al., 2013; Clark et al., 2009; DFID, 2014; Ebell et al., 2004; Guyatt et al., 2011; Hillier et al., 2011). Finally, although most systems are available online, information regarding seeking adherence to or translation of the systems—was not reported for any system except for GRADE. The GRADE approach was also unique in involving ongoing working groups aiming to continually advance and expand the applicability of its methodology in step with the developments in the area of evidence synthesis and assessment ("GRADE Working Group", 2017).

Discussion

The findings of this review demonstrate that evidence rating systems have more frequently been designed and used in the context of healthcare interventions than social policy, such as education, criminology and international development. This systematic review supports the need for a new guidance, which addresses important sources of complexity, when rating the certainty of evidence on intervention effectiveness. The previous evidence rating systems, particularly the GRADE approach and its modifications have made useful contributions to rating the certainty of evidence on intervention

effectiveness. However, no current system appears sufficient in its included domains, development and dissemination for incorporating a complexity perspective. In general, evidence ratings systems in health and social policy would benefit from including a wider group of experts in their collaboration beyond the immediate author team and conducting preliminary literature reviews to inform the content of the system. The systems would also benefit from implementing consensus-based procedures, such as a Delphi process and an expert meeting (Moher et al., 2010). Implementing rigorous activities to develop and widely disseminate an evidence rating system across various groups of users, will ensure that the system is useful and accessible by relevant stakeholders.

Domains of evidence for rating the certainty of evidence

This review identified several domains of evidence that suggest relevant modifications to the existing GRADE guidance when applied beyond clinical medicine and to public health interventions. If a new GRADE guidance were to be developed with the aim to incorporate the sources of complexity (see Table 3.1), certain aspects of the present guidance might require further elaboration and modification. Most frequently reported modifications in included systems related to the conceptualisation of the construct of "certainty of evidence", specifically, a few systems proposed to conceptualise the effects of interventions with long and dynamic causal pathways through a "chain of evidence" approach. Other modifications to the GRADE approach related to the considerations of study design and evidence hierarchy, a broadened interpretation of the domain of consistency of a body of evidence across different study designs and introduction of a new domain of analogous evidence.

Defining the certainty of evidence

This review identified very few instances where systems provided a definition for the construct of "certainty of evidence". The few reported definitions mainly focus on the confidence in the estimate of the effect of an intervention—a definition initially proposed by the GRADE approach (Guyatt et al., 2008). It is, however, worth noting here, that the GRADE Working Group has recently revised their conceptualisation of certainty of evidence based on pre- defined thresholds and the context of the review (Hultcrantz et al., 2017). The revised guidance suggests 3 types of ratings for a certainty of evidence: noncontextualised, partially contextualised and fully contextualised. For noncontextualised ratings, which refer to systematic reviews with no specified decision-making context, reviewers can rate the certainty that either the effect lies in a specified range (for example, a 95% confidence interval) or that the effect of one intervention differs from another by way of using a threshold of the non-null effect. In partially contextualised ratings a net benefit of an intervention as a whole is not possible to estimate, however, a minimal important difference can be selected for individual outcomes considered in the review. In this case reviewers can rate the certainty in a pre-specified magnitude of effect (for example, trivial, small, medium or large effect). Finally, for fully contextualised ratings, in clearly specified decision-making contexts, reviewers can rate the certainty that the effect lies above a threshold that makes the intervention worthwhile to implement (Hultcrantz et al., 2017). Certainty of evidence ratings are, therefore, contingent upon these thresholds, which may vary depending on the context and purpose of the review.

Two systems, namely the USPTSF and the GRADE-modified GEPHI system used a broader "chain of evidence" approach to enable assessment of the overall certainty of

evidence on intervention effectiveness. Because many public health and social interventions may involve long and variable causal pathways, researchers argue that in addition to rating the certainty of evidence for the intervention effect estimates, it might be informative to also assess the strength of evidence for causal inference more broadly, for example through a chain of evidence approach (see Figure 3.2). For this, tools such as logic models, which represent a graphic description of the links in the causal pathway of an intervention can be very useful in guiding the entire review process, including framing the review questions and informing the important pieces of evidence that should be searched for and synthesised (Rohwer et al., 2017). If reviewers manage to populate the different links in the entire causal chain of an intervention with rigorous evidence, then this may increase their confidence in the overall effects of an intervention (Bruce et al., 2014).

Study design and evidence hierarchy

One of the most contested topics in the discussions of the certainty of evidence relates to the hierarchy of evidence initially described in the paradigm of evidence-based medicine as an approach to differentiate between weak and strong study designs for assessing intervention effectiveness (Guyatt & Drummon, 2002). The classification of the study designs in this hierarchy is based on the ability of the design to minimise selection bias and confounding, thereby, assuring that intervention and control groups are comparable with regard to all observed and unobserved variables. While different versions of the evidence hierarchy have been described in clinical medicine, all of them place study designs, such as case series, which are considered relatively weaker in terms of protecting against threats to internal validity in the bottom of the hierarchy, followed

by case-control and cohort studies in the middle and RCTs at the top (Murad, Asi, Alsawas, & Alahdab, 2016). As this review demonstrates, an evidence hierarchy approach is still used in many evidence rating systems, especially those in clinical medicine, including the GRADE approach, which uses 2 broad categories of study designs as a starting point of the body-of-evidence rating process (RCT evidence starting as "high" and non-RCT evidence as "low" (Guyatt et al., 2008). By contrast, findings show that systems, which are used in other policy areas, such as public health tend to allow more flexibility for differentiating between the many types of non-RCT designs within their constructions of evidence hierarchies. This practice is commensurate with a view that quasi-experimental approaches should be given appropriate provisions in evidence rating systems as valuable methods for making causal inferences for public health interventions (Geldsetzer & Fawzi, 2017). Quasi-experimental studies are most frequently defined as observational studies with an exogenous variable that the investigator does not control (King, Keohane, & Verba, 1995). Many arguments have been put forward for quasi-experimental studies; for example, it is asserted that if certain assumptions are met they can generate evidence of causal strength similar to RCTs (Cook, Shadish, & Wong, 2008). This is critically important for contexts, when RCTs are not feasible or ethical, as is the case with many social and health system interventions. Quasi-experimental studies are also argued to generate evidence with higher degree of external validity and they may also be better suited for evaluating long-term health, economic and social outcomes (Barbero et al., 2015).

Although these discussions highlight the importance of better integration of quasi-experimental methods in evidence synthesis, it is also important that researchers conduct a tailored assessment of the execution of these studies before making

conclusions about the certainty of evidence produced by these designs (Rockers, Rottingen, Shemilt, Tugwell, & Barnighausen, 2015). As this review shows, the assessment of the execution of individual studies constituting the body of evidence is the most frequently reported domain of evidence in the evidence rating systems. While criteria for assessing execution of RCTs are well established in both health and social policy, methods for assessing execution and risk of bias of nonrandomised studies are still in development (Waddington et al., 2017). Ongoing initiatives such as that supported by the Cochrane Collaboration to develop tools for assessing "risk of bias in nonrandomised studies of interventions" (ROBINS-I) may provide further opportunities for the field to move away from the contested hierarchies of evidence when rating the certainty of evidence (Sterne et al., 2016).

Consistency

Consistency of the body of evidence was another most frequently reported domain of evidence in the included systems. Review findings demonstrate that evidence rating systems most frequently conceptualise consistency of a body as similarity in the magnitude and/or direction of effect estimates across studies (of same or similar design) included in the body of evidence. Concerns, however, have been raised that this approach only partly reflects the central tenet of the scientific method and the concept of consistency of evidence supported by Sir Austin Bradford Hill, which he defines as replicability of findings across *"a variety of situations and techniques"* (Fedak, Bernal, Capshaw, & Gross, 2015; Hill, 1965). In this view, researchers argue for a broader interpretation of the consistency of evidence to also consider "triangulation of evidence" across different methodological approaches when arriving at overall conclusions about

intervention effectiveness (Vandenbroucke, Broadbent, & Pearce, 2016). Triangulation has been defined as integration of evidence from several different methodological approaches (different study designs and analytical approaches), which address the same underlying causal question, however, which vary in terms of key sources of potential bias (e.g., multivariate regression, instrumental variables and RCTs; (Lawlor, Tilling, & Davey Smith, 2016). The importance of evidence triangulation has been cogently argued in the context of public health interventions involving longer causal pathways and multiple targets and behaviours, such as smoking, alcohol consumption, which are difficult (or impossible) to evaluate with RCTs alone. In order to adequately evaluate these interventions, it is often necessary to consider multiple evidence types, study designs and analytical approaches (Lorenc et al., 2016). When the results from these different methodological approaches are consistent, that is, they all point to the same conclusion, this is argued to strengthen the confidence in the overall findings (see Lawlor et al, 2016 for examples of evidence triangulation). This review identified only 1 system, which extended the domain of consistency to consider evidence from different study designs (Bruce et al., 2014); its broad interpretation, which looks at evidence from different methodological approaches to inform the certainty of evidence rating was unique within the findings of this review.

Analogous evidence

This review identified another modification to the GRADE approach employed by the GRADE-modified GEPHI system, namely introduction of a new domain of “analogy”. This domain corresponds to Sir Bradford Hill’s criterion of causation of the same name (Hill, 1965), which has been interpreted to mean that when one causal agent is known,

the standards of evidence can be lowered for a second causal agent that is similar in some way (Susser, 1991). Use of the domain of analogous evidence in systematic reviews is in line with the recent calls for incorporating complex systems perspectives in systematic reviews to better understand the effects of wider population-level interventions. By way of illustration, researchers suggest that understanding of alcohol marketing interventions can be enhanced, if systematic reviews take a broader perspective to incorporate a larger and related body of evidence, such as evidence on tobacco advertising (Petticrew et al., 2017).

Adherence to the best-practice techniques in development and dissemination

This review demonstrates that the previous evidence rating systems, which have addressed sources of complexity, when rating the certainty of evidence on intervention effectiveness would benefit from a more rigorous and transparent techniques for development and dissemination. These techniques are discussed below extending on the established best practices for developing and disseminating research reporting guidelines (Moher et al., 2010).

Recommended procedures for preparation and consensus development

According to the best practices for developing research reporting guidelines, the initial steps for developing a guidance for an evidence rating system should involve careful planning of the entire development process, including a comprehensive literature review identifying relevant guidance and domains of evidence from existing systems (Moher et al., 2010). Though developers of the systems included in this review did

conduct some preliminary searches to establish the need for a new guidance, they did not as systematically search for existing systems or report limitations of those. The GRADE approach has had widespread impact in clinical medicine, perhaps because it has used empirical evidence to identify the limitations of the previous evidence rating systems and established a need for a comprehensive guidance to be used consistently across reviewers and guideline developers around the world (Atkins et al., 2004).

Other techniques for preparation include obtaining adequate funding and identifying relevant stakeholders to participate in the development process. The expertise of these stakeholders should reflect the particular guidance under consideration (Moher et al., 2010). This review, however, shows that developers of the evidence rating systems mainly reported the involvement of the intermediate author team. More stakeholders could have been recruited from relevant content areas, such as researchers and methodologists of complex interventions across different practice areas beyond clinical epidemiology. To enable engaging with a diverse group of stakeholders, as well as a comprehensive literature review and further activities for disseminating the guidance widely, developers would require sufficient funding. As shown in this review, however, very few authors report on obtaining funding for developing and disseminating the system.

While an inclusive consensus development is considered a key procedure in developing a guidance, for the exception of few, the evidence rating systems included in this review did not report recommended techniques for consensus development, such as a Delphi process or an expert meeting. These procedures on the other hand, will enable incorporating knowledge and expertise of the various disciplines that use complex interventions and ensure that all important considerations are included in the guidance.

A face-to-face meeting of experts is considered the crucial element in developing consensus around the new guidance as it allows for in-depth discussions and clarification of any uncertainties and concerns. Not all participants, however, will be able to participate in that meeting; a Delphi consensus method is, therefore, recommended as a preparatory activity for the meeting and can help obtain input from a large number of relevant stakeholders (Moher et al., 2010). Given the breadth of possible considerations to examine, consensus development procedures would be essential for developing a comprehensive guidance on rating the certainty of evidence in complex interventions.

Recommended procedures for publication and dissemination

The most immediate activity after an expert meeting is to draft the paper describing the new system (Moher et al., 2010). In addition to preparing a manuscript outlining the system, the rationale for its development and the actual development process, it is recommended to publish an accompanying explanatory document providing in-depth instructions for using the system. For example, the GRADE Working Group provides guidelines for applying the GRADE approach in a dedicated series published in the Journal of Clinical Epidemiology (Guyatt et al., 2011). However, the examples in these guidelines mainly include clinical interventions and do not address sources of complexity of wider health and social policy interventions. This review did not identify a case, where authors would provide separate publications, one describing the process of development of the evidence rating system and the other, providing a detailed guidance for using the system. Evidence rating systems were described in 1 document and often lacked worked examples of applying the system in interventions from a range of health and social policy areas. An open access document explicating each domain of evidence included in the

system with examples of appropriate judgements can be essential for effective adherence to the guidance and for further updating of the guidance through feedback from users.

The final phase of developing a new evidence rating system would involve activities for widespread dissemination to ensure that the new system is available among users. The recommended techniques include translation of the guidance into other languages, as well as development of a strategy to incorporate feedback from users and to update the guidance as needed. Online open access can be a crucial publication and dissemination strategy to enhance the uptake of the evidence rating system. Involvement of editors in the development procedures can also be a potentially useful strategy to ensure a wide endorsement of the system by leading organisations involved in evidence synthesis, such as the Cochrane and the Campbell Collaborations.

Strengths and limitations of the review

This review's unique contribution may lie in its thorough exploration of the content, development and dissemination of the existing systems for rating the certainty of evidence across a range of policy areas, following systematic searches of bibliographic databases and sources of grey literature. Consequently, this review provides a comprehensive inventory of evidence domains for rating the certainty of evidence on intervention effectiveness across both health and social policy.

One of the acknowledged methodological challenges of the previous reviews has been associated with locating evidence rating systems through formal literature searches. For example, the vast majority of the systems analysed in the AHRQ review in 2002, were identified through bibliography searches and contacts with relevant experts

(West et al., 2002). This challenge was mainly attributed to the lag in the development of specific Medical Subject Headings (MeSH) for appropriate indexing of terms related to evidence-based practice and evidence assessment methods at the time. In order to overcome this challenge a Quality Assessment Tool (QAT) project review conducted in 2005 and updated in 2007, first ran a very sensitive search for systematic reviews of evidence rating systems followed by a search for individual systems complemented by expert consultation (Bai et al., 2012). Considering these challenges and recommendations highlighted in the AHRQ report regarding the search strategy for evidence rating systems, the review reported in this chapter aimed to balance the searches of scientific databases with an extensive search of grey literature, including 83 websites and databases of key stakeholder organisations. Furthermore, these searches were complemented with expert consultations to help locate additional relevant sources.

Nevertheless, several limitations are worth considering when interpreting the findings of this review. First, the scope of this review has been limited in several aspects because of practical considerations. For instance, it included documents published in English only, and, therefore, might have missed relevant work from the non-English literature. In order to identify evidence rating systems from grey literature, this review largely relied on the network and expertise of the DPhil candidate and the co-investigators of the project for developing *GRADE Guidance for Complex Interventions*; it is, therefore, likely that agencies and websites containing information on relevant evidence rating systems beyond this network have been missed. As this review comprises only one chapter of this thesis, the DPhil candidate had to limit searching and contacts with experts to progress into further research phases to meet the timelines of the DPhil

programme. This may have further compromised the comprehensiveness of the search strategy.

Furthermore, given the identified differences in the terminology and the varying levels of specification of the evidence domains, the mapping of these domains necessarily involved a degree of interpretation on the part of the DPhil candidate and the co-investigators from the project, who assisted in double data extraction and development of the heat maps. For example, another review team may have interpreted the broad evidence domains of the "efficacy data" of the Highest Attainable Standard of Evidence (HASTE) system (Baral et al., 2012) as referring to the strength of association and, therefore, in the map classified under the category of the "measures of precision", rather than consistency as is currently presented in this review. To mitigate the level of interpretation involved in the description of the evidence rating systems, the initial mapping of evidence domains by the DPhil candidate was independently verified by a second reviewer, and all issues were further discussed and clarified. Finally, considering the broad aims of this DPhil thesis to examine the use of GRADE in the context of complex interventions, it is worth noting that this review largely followed the terminology and the structure of the GRADE approach to describe the included evidence rating systems. It is, therefore, possible that another team of reviewers might have produced a slightly different mapping of the domains.

Conclusions

The systematic review and mapping of the evidence domains presented in this chapter aim to clarify how domains of evidence for rating the certainty of evidence on intervention effectiveness have been specified, developed and disseminated across

health and social policy, as well as how they address sources of complexity. While the findings demonstrate that the GRADE approach is one of the most comprehensive and transparent evidence rating systems in its guidance, development and dissemination, the systematic review reported in this chapter also shows that a few evidence rating systems, including the GRADE-modified GEPHI system (Bruce et al., 2014) suggest important insights into considering sources of complexity, when rating the certainty of evidence on intervention effectiveness. However, these systems have not been developed and disseminated using as rigorous and transparent methods as GRADE. The GRADE guidance on the other hand, has mainly been developed and validated in the context of biomedical interventions (Guyatt et al., 2008); its application to wider policy areas can be enhanced by accounting for the sources of complexity of broader social and public health interventions, such as use of different study designs and analytic approaches to evaluate the effectiveness of these interventions. Considering the status of the GRADE approach among all other evidence rating systems and adoption of GRADE by the leading organisations in evidence synthesis, including the Cochrane Collaboration and WHO, an official GRADE guidance for complex interventions might be most appropriate to ensure widespread dissemination and uptake by stakeholders.

The GRADE Working Group encourages initiatives to advance the GRADE guidance for application in different contexts ("GRADE Working Group", 2017). Developing a GRADE guidance for complex interventions could address the challenges of using GRADE in social and public health interventions as discussed in Chapter 1 (Movsisyan et al., 2016a, 2016b; Rehfuess & Akl, 2013), while also incorporating insights from a few relevant evidence rating systems discussed in this chapter. As the methodological research on applying a complexity perspective in evidence synthesis continues (Anderson

et al., 2013; MACH project, 2017), this new guidance for complex interventions can serve as a timely consolidation of current theoretical and practical thinking on addressing complexity in evidence assessment. Following rigorous and transparent procedures for development and dissemination (Moher et al., 2010), the GRADE guidance for complex interventions can provide a harmonised framework that can be used in the same transparent way within evidence synthesis in health and social policy.

References

Anderson, L. M., Petticrew, M., Chandler, J., Grimshaw, J., Tugwell, P., O'Neill, J., Shemilt, I. (2013). Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol, 66*(11), 1205-1208.

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., GRADE Working Group. (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 Serv Res, 4*(1), 38.

Bai, A., Shukla, V., Bak, G., & Wells, G. (2012). *Quality Assessment Tools Project Report*. Ottawa: Canadian Agency for Drugs and Technologies in Health.

Baral, S. D., Wirtz, A., Sifakis, F., Johns, B., Walker, D., & Beyrer, C. (2012). The highest attainable standard of evidence (HASTE) for HIV/AIDS interventions: toward a public health approach to defining evidence. *Public Health Rep, 127*(6), 572-584.

Barbero, C., Gilchrist, S., Schooley, M. W., Chriqui, J. F., Luke, D. A., & Eyler, A. A. (2015). Appraising the evidence for public health policy components using the quality and impact of component evidence assessment. *Glob Heart, 10*(1), 3-11.

Berkman, N. D., Lohr, K., Ansari, M., McDonagh, M., Balk, E., Whitlock, E., Chang, S. (2013). *Grading the strength of a body of evidence when assessing health care interventions for the effective health care program of the Agency for Healthcare Research and Quality: An update*. Rockville, MD: Agency for Healthcare Research and Quality.

Briss, P. A., Zaza, S., Pappaioanou, M., Fielding, J., Wright-De Aguero, L., Truman, B. I., Harris, J. R. (2000). Developing an evidence-based Guide to Community Preventive Services--methods. The Task Force on Community Preventive Services. *Am J Prev Med, 18*(1 Suppl), 35-43.

Bruce, N., Pruss-Ustun, A., Pope, D., Heather, A., & Rehfuess, E. (2014). *WHO indoor air quality guidelines: household fuel combustion: Methods used for evidence assessment*. Switzerland: World Health Organization.

Burkhardt, J., T., Schroter, D., C., Magura, S., Means, S., N., & Coryn, C., L., S. (2015). An overview of evidence-based program registers (EBPRs) for behavioural health. *Eval Program Plann, 48*, 92-99.

Clark, E., Burkett, K., & Stanko-Lopp, D. (2009). Let Evidence Guide Every New Decision (LEGEND): an evidence evaluation system for point-of-care clinicians and guideline development teams. *J Eval Clin Pract, 15*(6), 1054-1060.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates:

New findings from within-study comparisons. *J Policy Anal Manage, 27*(4), 724-750.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, P. (2008). Developing and evaluating complex interventions: new guidance. Medical Research Council (MRC). Retrieved 20 Jun, 2017, from https://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/

Department for International Development. (2014). *How to Note: Assessing the strength of evidence*. Retrieved 10 Jan, 2017 from https://www.gov.uk/government/publications/how-to-note-assessing-the-strength-of-evidence

Diez Roux, A. V. (2011). Complex systems thinking and current impasses in health disparities research. *Am J Public Health, 101*(9), 1627-1634.

Ebell, M. H., Siwek, J., Weiss, B. D., Woolf, S. H., Susman, J., Ewigman, B., & Bowman, M. (2004). Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician, 69*(3), 548-556.

Eccles, M., Freemantle, N., & Mason, J. (1998). North of England evidence based guidelines development project: methods of developing guidelines for efficient drug use in primary care. *BMJ, 316*(7139), 1232-1235.

Fedak, K. M., Bernal, A., Capshaw, Z. A., & Gross, S. (2015). Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol, 12*, 14.

Geldsetzer, P., & Fawzi, W. (2017). Quasi-experimental study designs series-paper 2: complementary approaches to advancing global health knowledge. *J Clin Epidemiol,* 89, 12-16.

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. Research Papers in Education, 22(2), 213-228.

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. London, UK: SAGE Publications Ltd.

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. (2017). Retrieved 25 Jun, 2017 from http://gradeworkinggroup.org/

Grant, S. P., Mayo-Wilson, E., Melendez-Torres, G. J., & Montgomery, P. (2013). Reporting quality of social and psychological intervention trials: a systematic review of reporting guidelines and trial publications. *PLoS One, 8*(5).

Greer, N., Mosser, G., Logan, G., & Halaas, G. W. (2000). A practical approach to evidence grading. *Jt Comm J Qual Improv, 26*(12), 700-712.

Guyatt, G., & Drummon, R. (2002). *Users' guide to the medical literature: a manual for evidence-based clinical practice.* Chicago: American Medical Association.

Guyatt, G., H., Oxman, A., D., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., Schunemann, H., J. (2011). GRADE guidelines 6. Rating the quality of evidence--imprecision. *J Clin Epidemiol, 64*(12), 1283-1293.

Guyatt, G., H., Oxman, A., D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., GRADE Working Group. (2011). GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol, 64*(12), 1303-1310.

Guyatt, G., H., Oxman, A., D., Montori, V., Vist, G., Kunz, R., Brozek, J., Schunemann, H., J. (2011). GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol, 64*(12), 1277-1282.

Guyatt, G., H., Oxman, A., D., Sultan, S., Glasziou, P., Akl, E. A., Alonso-Coello, P., GRADE Working Group. (2011). GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol, 64*(12), 1311-1316.

Guyatt, G., H., Oxman, A., D., Vist, G., E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., GRADE Working Group. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*(7650), 924-926.

Guyatt, G., H., Sackett, D., L., Sinclair, J., C., Hayward, R., Cook, D., J., & Cook, R., J. (1995). Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. *JAMA, 274*(22), 1800-1804.

Guyatt, G., Oxman, A., D., Akl, E., A., Kunz, R., Vist, G., Brozek, J., Schunemann, H., J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol, 64*(4), 383-394.

Guyatt, G., Oxman, A., D., Sultan, S., Brozek, J., Glasziou, P., Alonso-Coello, P., Schunemann, H., J. (2013). GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *J Clin Epidemiol, 66*(2), 151-157.

Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *BMJ, 323*(7308), 334-336.

Hawe, P., Shiell, A., & Riley, T. (2009). Theorising interventions as events in systems. *Am J Community Psychol, 43*(3-4), 267-276.

Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proc R Soc Med, 58*, 295-300.

Hillier, S., Grimmer-Somers, K., Merlin, T., Middleton, P., Salisbury, J., Tooher, R., & Weston, A. (2011). FORM: An Australian method for formulating and grading

recommendations in evidence-based clinical guidelines. *BMC Med Res Methodol, 11*, 23.

Hultcrantz, M., Rind, D., Akl, E. A., Treweek, S., Mustafa, R. A., Iorio, A., . . . Guyatt, G. (2017). The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol, 87*, 4-13.

Joanna Briggs Institute. (2014). *Supporting document for the Joanna Briggs Institute levels of evidence and Grades of Recommendations*. Retrieved 20 Jan, 2017 from http://joannabriggs.org/assets/docs/approach/Levels-of-Evidence-SupportingDocuments.pdf

Johnson, S., D., Tilley, N., & Bowers, K., J. (2015). Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis reviews. *J Exp Criminol, 11*, 459-473.

King G, Keohane RO, & Verba, S. (1995). The importance of research design in political science. *Am Polit Sci Rev, 89*(2), 475-481.

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2016). Triangulation in aetiological epidemiology. *Int J Epidemiol, 45*(6), 1866-1886.

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gulmezoglu, M., Rashidian, A. (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 Med, 12*(10).

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

Liddle, J., Williamson, M., & Irwig, L. (1996). *Method for evaluating research and guideline evidence*. Sydney: NSW Health Department.

Lorenc, T., Felix, L., Petticrew, M., Melendez-Torres, G. J., Thomas, J., Thomas, S., . . . Richardson, M. (2016). Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Syst Rev, 5*(1).

Miles, B. M., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2).

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016a). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol, 78*, 22-33.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016b). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evid Based Med, 21*(4), 125-127.

National Institute for Health and Care Excellence. (2012). *Methods for the development of NICE public health guidance: Process and methods guides* (3rd ed.). Retrieved 19 Dec, 2016 from https://www.nice.org.uk/process/pmg4/chapter/introduction

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. (2005). Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? *J Epidemiol Community Health, 59*(10), 886-892.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Syst Rev, 5*(1), 210.

Petticrew, M., Anderson, L., Elder, R., Grimshaw, J., Hopkins, D., Hahn, R., Welch, V. (2013). Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol, 66*(11), 1209-1214.

Petticrew, M., Shemilt, I., Lorenc, T., Marteau, T. M., Melendez-Torres, G. J., O'Mara-Eves, A., Thomas, J. (2017). Alcohol advertising and public health: systems perspectives versus narrow perspectives. *J Epidemiol Community Health, 71*(3), 308-312.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rogers, M., Duffy, S. (2006). *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme*. Lancaster, UK.

Meta-analysis, Complexity and Heterogeneity (MACH) project. (2017). Retrieved 18 May, 2017 from http://evaluation.lshtm.ac.uk/2016/12/14/2667/

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rockers, P. C., Rottingen, J. A., Shemilt, I., Tugwell, P., & Barnighausen, T. (2015). Inclusion of quasi-experimental studies in systematic reviews of health systems research. *Health Policy, 119*(4), 511-521.

Rohwer, A., Pfadenhauer, L., Burns, J., Brereton, L., Gerhardus, A., Booth, A., . . . Rehfuess, E. (2017). Series: Clinical Epidemiology in South Africa. Paper 3: Logic models help make sense of complexity in systematic reviews and health technology assessments. *J Clin Epidemiol, 83*, 37-47.

Saldana, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Los Angeles, Calif.; London: SAGE publications.

Sawaya, G. F., Guirguis-Blake, J., LeFevre, M., Harris, R., Petitti, D., & Force, U. S. P. S. T. (2007). Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med, 147*(12), 871-875.

Sterne, J. A., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919.

Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am J Epidemiol, 133*(7), 635-648.

Tang, K. C., Choi, B. C., & Beaglehole, R. (2008). Grading of evidence of the effectiveness of health promotion interventions. *J Epidemiol Community Health, 62*(9), 832-834.

Treadwell, J. R., Tregear, S. J., Reston, J. T., & Turkelson, C. M. (2006). A system for rating the stability and strength of medical evidence. *BMC Med Res Methodol, 6*, 52.

Turner-Stokes, L., Harding, R., Sergeant, J., Lupton, C., & McPherson, K. (2006). Generating the evidence base for the National Service Framework for Long Term Conditions: a new research typology. *Clin Med (Lond), 6*(1), 91-97.

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol, 45*(6), 1776-1786.

Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters LM. (2012). *Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions.* Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. Retrieved from Rockville, MD: Agency for Healthcare Research and Quality.

Waddington, H., Aloe, A. M., Becker, B. J., Djimeu, E. W., Hombrados, J. G., Tugwell, P., . . . Reeves, B. (2017). Quasi-experimental study designs series-paper 6: risk of bias assessment. *J Clin Epidemiol*, 89, 43-52.

Weightman, A., Ellis, S., Cullum, A., Sander, L., & Turley, R. (2005). *Grading evidence and recommendations for public health interventions: developing and piloting a framework*. Retrieved 10 Dec, 2016 from http://orca.cf.ac.uk/69810/

West, S., King, V., & Carey, T. e. a. (2002). *Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47* (Prepared by the Research Triangle Institute–University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). Retrieved from Rockville, MD: Agency for Healthcare Research and Quality.

Whitlock, E. P., Williams, S. B., Gold, R., Smith, P., & Shipman, S. (2005) *Screening and Interventions for Childhood Overweight*. Rockville: MD: Agency for Healthcare Research and Quality.

Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Services., T. F. o. C. P. (2000). Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. *Am J Prev Med, 18*, 44-74.

Chapter 4. Semi-structured interviews with review authors and methodologists

Experiences and practices of using GRADE in systematic reviews of complex interventions

A paper adaptation of this chapter has been submitted to *Systematic Reviews*

Chapter Overview

This chapter describes and compares experiences and practices of review authors and GRADE methodologists in applying GRADE to systematic reviews of complex interventions. This comparison enables a separation of the "real" challenges of the GRADE methodology itself from those associated with systematic reviewing of complex interventions more broadly, such as levels of expertise needed to appropriately implement the GRADE guidelines. Semi-structured interviews were undertaken with 10 review authors and 5 methodologists from the GRADE Working Group regarding their views on GRADE use in the same sample of systematic reviews of complex interventions. Reviews were purposively chosen from the Cochrane and Campbell Libraries, and a thematic approach with cross-case analysis was employed to examine and compare the perspectives of review authors and GRADE methodologists.

Fourteen themes were identified related to participants' perceptions of complexity in systematic reviewing, familiarity with and perceived utility of GRADE, views on initial categorisation of evidence in GRADE, implementation of GRADE in systematic reviews of complex interventions, and suggestions to enhance the GRADE guidance for reviews of complex interventions. Review authors found GRADE to require additional methodological expertise in reviews of complex interventions because of broad review

questions, inclusion of heterogeneous studies and different types of study designs in these reviews. In the meantime, challenges of implementing GRADE in reviews of complex interventions were often seen by GRADE methodologists as primarily influenced by decisions at earlier review stages. While GRADE methodologists noted the necessity of considering GRADE from the beginning of the review process when framing questions on complex interventions and deciding on the analytic approaches to structure the Summary of Findings (SoFs) tables, review authors often reported considering GRADE only at the end of the review process. Specific suggestions were made for a more tailored guidance and examples of GRADE application in reviews of complex interventions, as well as for structural changes to the GRADE approach, such as incorporation of additional domains of evidence to enable upgrading evidence and to revise the initial categorisation of evidence based on study design. It was, however, also noted that these structural changes might be difficult to achieve, in particular, the approval of the GRADE Working Group, and that proposed changes would need to be substantiated with compelling examples and evidence.

Introduction

The previous chapter of this thesis reported on a systematic review of evidence rating systems from health and social policy aiming to describe how the previous systems have addressed sources of complexity of interventions when rating the certainty of evidence on intervention effectiveness. The findings demonstrate that evidence rating systems have more frequently been designed and used in healthcare than in social interventions, and that, with the exception of a few instances, systems do not attend to the sources of complexity in evidence synthesis. The review also showed that the existing systems would benefit from a more rigorous and transparent techniques for development and dissemination. In this regard, the GRADE approach was found to be the most comprehensive in its guidance, development and dissemination, and several attempts have been reported to modify aspects of GRADE for health system and public health interventions. Since the GRADE approach has had large uptake by the leading organisations in evidence synthesis and guideline development, the previous chapter concluded that it would be most appropriate to pursue advancing the existing GRADE guidance for complex interventions.

Following the strategy adopted by this thesis, which draws on the recommended techniques for developing research reporting guidelines (Moher, Schulz, Simera, & Altman, 2010), the next step in developing an official GRADE guidance for complex interventions entails generating suggestions for consideration at a face-to-face meeting of relevant stakeholders. For this, the best practices for developing reporting guidelines recommend conducting a systematic review followed by a Delphi process to seek input for a new guidance from a large number of stakeholders (Moher et al., 2010). However,

the review reported in the previous chapter did not produce many suggestions for immediate testing in a consultation survey, such as a Delphi process. Furthermore, an evidence rating system like a GRADE approach is different from a health reporting guideline in that it does not include a set of structured items, but rather provides a general guidance on how to interpret and rate different domains of evidence across examples of systematic reviews. This means generating suggestions for a new GRADE guidance would benefit from further consultation with key stakeholders before proceeding with testing these suggestions for scientific agreement in a Delphi-based process.

The need for further in-depth investigation into specific challenges of using GRADE and suggestions for advancing the GRADE methodology for complex interventions was also reinforced by the members of the GRADE Working Group at the annual group meeting in Seoul (Cochrane Colloquium, 2016). During this meeting, the DPhil candidate presented the results from Chapters 1 and 3 to the members of the GRADE Working Group seeking feedback on the need for developing a new GRADE guidance for complex interventions. While the Group supported that a new guidance for complex interventions would be appropriate, concerns were raised that many of the reported challenges of using GRADE in systematic reviews of complex interventions might stem from inadequate adherence to the existing GRADE guidelines during the process of GRADE application by review authors, such as lack of judgment in applying the GRADE domains rather than inadequacies of the GRADE guidance itself. An additional stakeholder consultation was, therefore, suggested to guard against what in intervention research has been termed as Type III error (Basch, Sliepcevich, Gold, Duncan, & Kolbe, 1985). Specifically, seeking input from professionals with experience in GRADE use would enable separation of the

"real" challenges of the GRADE methodology from those related to the lack of knowledge and inadequate implementation of the existing GRADE guidelines, and consequently help generate appropriate suggestions for a new guidance.

This chapter reports on an investigation aiming to examine issues around using GRADE in systematic reviews of complex interventions by way of collecting in-depth qualitative data on the practices and experiences of both review authors and members of the GRADE Working Group, who regularly attend Group meetings, have experience in GRADE use and are involved in developing different GRADE guidelines (these stakeholders will be referred to as "GRADE methodologists" in the rest of this chapter). Although qualitative investigation can be very diverse and include a range of methods and approaches, at a general level, it is described as a naturalistic and interpretative approach, concerned with exploring phenomena *"from the interior"* (Flick, 2009); this entails attempting to explain phenomena in terms of how they are experienced by people and the meanings that people bring to them (Denzin & Lincoln, 2011). As a guide for researchers, Patton suggests several research questions, which are suitable for a qualitative study (Patton, 2002). These include: questions about people's views and experiences; enquiry into the meanings people make of their experiences; and studying people in the context of their social environment and/or research, where it is difficult to develop a standardised instrument because of the lack of knowledge on the phenomenon (Patton, 2002). In answering these enquiries, qualitative research serves to provide outputs including a detailed description of the phenomena being researched, grounded in the perspectives and accounts of participants (Ormston, Spencer, Barnard, & Snape, 2014). Considering that approaches for addressing sources of complexity in evidence assessment are few and still in development, detailed description of users' and

methodologists' accounts of GRADE in reviews of complex interventions would be highly relevant for generating suggestions for the GRADE guidance for complex interventions.

The study reported in the present chapter uses this qualitative approach to describe and compare experiences of GRADE use in systematic reviews of complex interventions. A sample of systematic reviews of complex interventions was chosen, and the GRADE approach was applied to this sample of reviews first by review authors and then by GRADE methodologists. By comparing experiences of authors (i.e., immediate users of GRADE) and GRADE methodologists (i.e., developers of the GRADE guidelines) on the same set of reviews, this qualitative research aims to distil aspects of the GRADE methodology that are challenging to apply in reviews of complex interventions from those related to the inappropriate implementation of the existing GRADE guidance, and elucidate ways to advance the GRADE approach for complex interventions. In this way, this research can build on the findings of the investigation of review contents and feedback from the review authors reported in Chapter 1, as well as the systematic review reported in the previous chapter, and provide further insights for the GRADE guidance for complex interventions.

Methods

Data collection

First, reviews of complex interventions that have applied GRADE were purposively selected from the databases of the Cochrane and Campbell Collaborations published in 2015 and 2016. This study aimed to focus on recently published reviews to mitigate the possible recall bias associated with the use of GRADE in the review, while trying to achieve diversity in intervention types (such as health policy, health system,

occupational, environmental, behavioural, and educational interventions) and levels (such as individual, community, organisational, and population) from those available on these databases. The review authors were then contacted with up to three email reminders to explain the research aims and obtain their agreement for an interview. For pragmatic reasons, contact with review authors was ceased once authors of ten reviews agreed to participate in this study. These ten reviews were then assigned to a convenience sample of five methodologists from the GRADE Working Group, who re-applied the GRADE approach to assess the certainty of evidence in the primary outcomes of these reviews. The GRADE methodologists were chosen and contacted from the list of those who previously showed interest in the topic at the GRADE Working Group meeting held during the 24th Cochrane Colloquium in Seoul. Each GRADE methodologist who agreed to participate in this study was assigned to two of the ten included reviews, along with a list of available publications on the GRADE approach and a template of the GRADE Evidence Profile. They were asked to document each judgment they made about the certainty of evidence in these reviews and any issues they encountered when applying GRADE. This allowed for a cross-examination of the experiences of using GRADE in the same sample of reviews from both methodological (i.e., GRADE methodologist) and user (i.e., review author) perspectives.

Once the sample of reviews and participants was finalised, the intervention Complexity Assessment Tool within Systematic Reviews (iCAT_SR) was applied by the DPhil candidate to describe the level of complexity of interventions in the reviews (Lewin et al., 2017). The iCAT_SR tool is comprised of ten dimensions of complexity: (1) the number of active components in the intervention; (2) the number of behaviours of recipients to which the intervention is directed; (3) the range and number of

organisational levels targeted by the intervention; (4) the degree of tailoring intended or flexibility permitted across sites or individuals in applying or implementing the intervention; (5) the level of skill required by those delivering the intervention; (6) the level of skill required by those receiving the intervention; (7) the degree of interaction between intervention components; (8) the degree to which the effects of the intervention are context-dependent; (9) the degree to which the effects of the interventions are changed by recipient or provider factors; and (10) the nature of the causal pathway between intervention and outcome. Dimensions 1–6 are considered to be the "core" dimensions of complexity, while dimensions 7–10 are "optional" (Lewin et al., 2017). Each of these dimensions can be rated on one of three levels of complexity, thereby, allowing descriptions of interventions along a "simple-complex" continuum (Lewin et al., 2017).

Semi-structured interviews with review authors and GRADE methodologists were conducted in February and March 2017 via Skype or by telephone. Two interviews were conducted jointly by the DPhil candidate and her supervisor to validate the interview guide. The remaining interviews were conducted by the candidate alone. The interview guide covered the following organising topics: complexity in systematic reviewing, familiarity with and perceived utility of GRADE, initial categorisation of evidence in GRADE, implementation of GRADE in the review, and suggestions for enhancing the GRADE guidance on complex interventions. The questions in the guide were open-ended, and probes were used as necessary to elicit further information (see Appendix 5 for the interview guide). The duration of the interviews ranged from 24 to 80 minutes. Amazon gift vouchers were offered to all participants at the end of the interviews as a token of appreciation for their time and contribution.

Data analysis

All interviews were audio-recorded and transcribed by the candidate. A thematic analysis was employed with a mixed approach of deductive and inductive coding (Attride-Stirling, 2001). A thematic analytic approach allowed to identify the common themes of the experiences of review authors and GRADE methodologists regarding reviewing complex interventions and using the GRADE approach. Following this, cross-case analysis was used to explore differences in experiences between review authors and GRADE methodologists (Miles & Huberman, 1994). A cross-case analysis enabled to examine how the themes played out between the perspectives of review authors and GRADE methodologists by way of charting and comparing the identified themes between these two groups of participants.

The organising topics in the interview guide were used to pre-specify the initial coding categories. Transcripts were, therefore, read several times, and structural coding was first employed to assign the pre-defined high-level codes representing a topic of inquiry to a segment of data that related to that topic (Saldana, 2013). These segments of data were further coded using a more inductive approach. These codes were then collated and sorted into corresponding themes. These themes were further charted and cross-compared between review authors and GRADE methodologists to examine similarities and differences in the stakeholder perspectives within each theme. For the final step, transcripts were re-read to double check that the derived codes and themes comprehensively represent the data. Coding was conducted by the DPhil candidate using NVivo 11, and further reviewed and discussed with a co-investigator from the project on developing *GRADE guidance for Complex Interventions*. The study was approved by the

Departmental Research Ethics Committee at the University of Oxford (Ref: SPI_C1A_16_008).

Results

Characteristics of the included reviews and participants

Fourteen teams of review authors were contacted; authors of two reviews did not respond to the invitation email, and authors of two other reviews were not available because of busy schedules. Overall, ten review authors and five methodologists from the GRADE Working Group participated in the interviews (see Table 4.1). All of the participants were affiliated to academic institutions across 8 countries, including UK (n=5), USA (n=3), Australia (n=2), Canada (n=1), New Zealand (n=1), Switzerland (n=1), Netherlands (n=1), and Nepal (n=1). Review authors came from different social disciplines, including public health (n=6), social care (n=2), psychology (n=1), and economics (n=1). From GRADE methodologists, three had expertise in clinical epidemiology, one in social care, and one in public health policy. Half of the sample of review authors (n=5) reported previous experience of using GRADE in either a systematic review or a guideline development context, while the other half (n=5) were first-time users of GRADE. One of the review authors who was initially contacted for an interview had been involved with the activities of the GRADE Working Group previously and was a member of the Group, and, therefore, this author's views are presented both as a methodologist, as well as from a review author's perspective. The other five methodologists were all experienced GRADE users and current members of the GRADE Working Group.

The sample of reviews considered in this study included interventions with different levels of complexity for each of the ten dimensions outlined in iCAT_SR. As shown in Table 4.2, *behaviours of recipients*, *levels of target, flexibility of implementation,* and *skills of providers* were identified as dimensions with the highest level of complexity in the included sample of reviews. In the meantime, lack of reporting hindered adequate rating of the optional dimensions of complexity in iCAT-SR. Most of these dimensions were assessed as "unclear or unable to assess". While this study aimed to have a balanced representation of reviews from Cochrane and Campbell Libraries in the sample, the majority of those included were published on the *Cochrane Library* (see Table 4.2). Difficulties were encountered in locating recently published reviews that used GRADE in the Campbell Library. Correspondence with a Campbell review author suggested that many Campbell reviews of social interventions may not currently use GRADE due to perceptions that it unfairly penalises high-quality quasi-experimental studies, is inappropriate for meta-analyses that focus explicitly on exploring heterogeneity (rather than mean effect sizes), and is unclear about how to use tools other than the Cochrane Risk of Bias tool (Higginson et al., 2015; Wilson, Gill, Olaghere, & McClure, 2016) when assessing study limitations (e.g., the International Development Campbell Group tool).

Thematic analysis

The thematic analysis identified 14 themes broadly categorised into the five pre-defined organising topics (see Table 4.3). Selected quotations that illustrate similarities and differences in participant perspectives regarding these themes are presented below; "[…]" in the quotations indicates instances where a text has been removed for brevity or anonymity

Table 4.1. Participant characteristics

	Country	Discipline	Prior GRADE use
Author 1	Australia	Public health	Yes
Author 2	Australia	Public health	No
Author 3	Canada	Public health	No
Author 4	Nepal	Public health	No
Author 5	New Zealand	Psychology	Yes
Author 6	Switzerland	Public health	Yes
Author 7	UK	Social care	Yes
Author 8	UK	Economics	Yes
Author 9	UK	Social care	No
Author 10	USA	Public health	No
Methodologist 1	Netherlands	Clinical epidemiology	Yes
Methodologist 2	UK	Social care	Yes
Methodologist 3	UK	Public health	Yes
Methodologist 4	USA	Clinical epidemiology	Yes
Methodologist 5	USA	Clinical epidemiology	Yes

Notes: Author 8 is also a methodologist engaged with the activities of the GRADE Working Group.

Table 4.2. Review characteristics

	Publication source	**Intervention type**	**iCAT_SR dimensions of complexity**									
			Active components	Recipient behaviour	Levels of target	Flexibility / tailoring	Provider skills	Recipient skills	Component interactions	Context dependency	Individual dependency	Causal pathway
Review 1	Cochrane Public Health	Behavioural/ educational	++	+++	+++	+++	+++	+++	Unclear	Unclear	Unclear	Unclear
Review 2	Cochrane Public Health	Health policy	++	+++	+++	+++	+++	+	Unclear	Unclear	Unclear	+++
Review 3	Cochrane EPOC	Health system	++	+++	+++	+++	+++	+	+	Unclear	Unclear	Unclear
Review 4	Cochrane EPOC	Health system	+	+++	++	+++	+++	+	+	Unclear	Unclear	Unclear
Review 5	Campbell SWCG	Behavioural/ educational	Varies	+++	+	+++	+++	+	Unclear	Unclear	Unclear	++
Review 6	Cochrane Public Health	Health policy	++	+++	+++	Varies	Varies	+	Unclear	Unclear	Unclear	+++
Review 7	Cochrane Public Health	Health policy	Varies	+++	+++	Varies	Varies	+	Unclear	Unclear	++	+++
Review 8	Cochrane CCRG	Behavioural/ educational	Varies	+++	+	+++	+++	+	Unclear	Unclear	Unclear	Unclear
Review 9	Cochrane Public Health	Behavioural/ educational	+	+++	+	Varies	+	+	+	Unclear	++	+
Review 10	Cochrane Work	Occupational	Varies	Varies	+++	Varies	Varies	++	Unclear	Unclear	Unclear	Unclear

Notes: +++: “high levels of complexity”, ++: “moderate levels of complexity”; +: “low levels of complexity”. CCRG = Consumers and Communication Group; EPOC = Effective Practice and Organisation of Care; SWCG = Social Welfare Coordination Group

Table 4.3. Organising topics, themes and representative quotes

Organising topic	Theme	Quote (review author)	Quote (GRADE methodologist)
Complexity in systematic reviewing	Need to include heterogeneous evidence (PICO elements & study design)	"In a complex intervention always different people will apply the intervention [differently] depending on what's available in the environment." (Author 9)	"I was trying to get a sense of what are the essential elements of that complex intervention, because each of the included studies had different elements." (Methodologist 4)
	Need to ask broad questions (or "lumping")	"The research question for this review was deliberately broad, because [...] we were trying to capture as much evidence as possible, so that we could say something useful about what we were interested in." (Author 2)	"I think that is in these kind of complex interventions that you often have a broader question, because interventions can be different or have different elements in their complexity." (Methodologist 1)
	Need for flexibility in the choice of methods and approaches	"[…] If there could be a little bit more flexibility and scope for creativity within the Cochrane process to let the authors propose their own approach to summarising the evidence, and not necessarily to have to use GRADE or any other criteria or scale, I would tend to lean towards that […]". (Author 3)	"[…] All the studies included in the review need to be assessed for quality [...] It might be possible to have non-Cochrane tools feeding into the GRADE assessment on the study limitations criterion, but those would still need to be tools that would assess risk of bias". (Author 8 & Methodologist)
Familiarity with and perceived utility of GRADE	GRADE domains for rating the certainty in the estimates of intervention effects vs. GRADE Evidence to Decision (EtD) criteria[1]	"I don't want to imply that our data were perfect or fantastic [...] But I guess my issue is the breadth of considerations that goes into GRADE. So, they are all methodological, and I just think as soon as you start thinking about public health interventions you are stepping outside of methods and you need to be able to take into consideration things like the magnitude of the problem for a population and even the cost-effectiveness of the intervention relative to other approaches [...]" (Author 3)	"[…] GRADE absolutely assesses your certainty in the estimate of effect by statistical association, it's really time to capture that element of it. But there may be other things that would affect your understanding of that effect, really [...] And some of this may or may not be relevant to a systematic review, but certainly a key to working in a guideline world." (Methodologist 2)

[1] For further details on GRADE Evidence to Decision (EtD) criteria see Alonso-Coello et al. (2016)

	Use of different domains to rate the certainty of evidence	"Clinicians who engage with research on a more casual basis compared to academics have a tendency to look at whether or not something is outright statistically significant, or not, and so, I liked all the additional considerations [e.g., inconsistency and imprecision] GRADE brings to evaluation of [certainty] of evidence, and I like also that it goes beyond simply evaluation of risk of bias." (Author 5)	"I think the purpose of GRADE is to give a framework. You have the evidence and ok, what are your concerns about the evidence, how do I trust the evidence, how much confidence do I have? Ok, I need a structured approach. So, I look at the risk of bias, I look at inconsistency, and all the factors." (Methodologist 1)
	Provision of a structured summary of key review findings	"Everyone does not have time to go through all the literature and everything, so one can have a look at these SoFs tables and find out very important information about the review; and it's also easier for those who are not researchers, such as politicians and guideline [developers], so they can easily understand the results of the review." (Author 4)	"SoFs table really shows also how/if your [review] question is answered." (Methodologist 1)
Initial categorisation of evidence in GRADE	Inadequate differentiation of risk of bias from different NRSs	"For me, the issue with GRADE was being able to differentiate between the strengths and weaknesses of non-RCTs [...] It's like having a, picking a test and the results are you either pass or fail, but it doesn't give you any sense of the distribution of knowledge for example, if you are taking the test. It's like pass or fail, period. So, the evidence fell into that low certainty rank, and wasn't very sensitive to describe elements of the studies that could inform people that are trying to use these kind of interventions." (Author 10)	"I think first of all it makes sense for RCTs to start high, and what I think is, there is an issue with respect to quasi-experimental designs, because currently all non-RCTs start out as low, and I think there is at least the case for some non-RCTs, namely quasi-experimental study designs to start out as moderate, possibly even high." (Author 8 & Methodologist)
	Integrating ROBINS-I tool into GRADE risk of bias assessment	–	"Based on this work [recent discussions in the GRADE Working Group], what you can do, is, if you have nonrandomised evidence and you applied a ROBINS tool, which goes really deeply into

			confounding and a lot of other stuff [...] If you use the ROBINS tool you can say, ok, all evidence starts high, then in ROBINS you will see that you have high risk of bias because of lack of randomisation and then you downgrade for lack of randomisation and you downgrade two levels, unless you have very good reasons that one level downgrading is enough." (Methodologist 1)
	Required methodological expertise and time for applying ROBINS-I	"I think there are certainly some issues with the nonrandomised studies risk of bias [ROBINS-I] tool in that its very challenging to implement, very time-consuming." (Author 8 & Methodologist)	"The problem with that [ROBINS-I] is that it is not an easy to use tool, we tried to use it and it is really difficult to use and it is hard to get agreement between reviewers on how to use it, so from an operational stand point that tool has not gained traction." (Methodologist 5)
Implementing GRADE in the review	Situating GRADE in the review process	"I think one of my main concerns is that inevitably GRADE is something that you tend to use at the end of the project; in some reviews it just becomes an annoying add on you've got to do, a final step, a final hurdle you've got to go through." (Author 7)	"My view on GRADE is that it's not something that should be just thought about at the end of the process as an adjunct to the review in order to complete a SoFs table [...] It's a tool that you should always have in mind and also in terms of thinking about at the protocol stage, how to pre-specify or stratify perhaps the analyses that you are going to do in different ways, in order to than feed into those SoFs tables." (Author 8 & Methodologist)
	Use of judgment and expertise in GRADE ratings	"They [the GRADE ratings] feel almost quite arbitrary conclusions that you are coming to, it almost feels like, oh well, I think let's just put that … I didn't have a great deal of confidence in what I was doing, frankly […] I actually ended up discussing it more with the Cochrane team, because they seemed to know what to do with it. And, everything I did, they seemed to have to redo. So, I don't know if I was any good at it." (Author 7)	"I think sometimes, particularly when we are challenged with time [...] it is easy to just go, oh no, that looks different, I'll downgrade it; but actually, you should be thinking much more critically: well is that going to make a difference, even though it has just been done [in a different population]. I think we have to be better in doing that." (Methodologist 2)

	Time and resources required for adequate implementation of GRADE	"It [the GRADE guidance] is long, and I think as you noticed it's mostly for straightforward interventions or drugs ... As soon as you are into something a little bit more different, more complex, I find it harder to follow […] When I was doing reviews it was mostly me and only one other person, so we didn't have the resources of someone specialising in the methods that could do the GRADE ratings for us." (Author 6)	"I see people doing it [GRADE ratings] very quickly, at the end, without much thought, because they spent so much time on all the other phases, and some of that time that is spent earlier is really useless time. Like you spend months going through abstracts that are completely irrelevant to the topic [...] So, I think the work in a systematic review should be rebalanced, where this phase of going through thousands of abstracts, which is completely waste of time, should be done quickly and then, a lot more time spent on the actual studies that are included to look at them in a more thoughtful way." (Methodologist 5)
Suggestions for enhancing the GRADE guidance for complex interventions	Guidance and examples targeting complex interventions	"I think that the guidance could be enhanced by examples of saying, you know, in this review the authors chose to upgrade, because of the effect size, although there are no guidelines for doing this, these are the type of reasons that they chose." (Author 1)	"I think if it was something that gives people that ability to be freer in using judgment […] If there is a guidance for complex interventions and gives that confidence to people and have some good examples, then that might be very helpful." (Methodologist 2)
	Additional possibilities to upgrade evidence	"I would probably support the point that if you have nonrandomised trials that are showing consistent effects, then that would improve my confidence, and we should probably upgrade the evidence." (Author 2)	"I think there probably is need for some extra components … The other obvious area that to me seems missing, is the role of theory. So, if you have a very-very strong mechanistic reasons for expecting an effect, which are pre-specified in some way, and then you find an effect, then, should that be considered in terms of upgrading or not?" (Methodologist 3)

Notes: GRADE: Grading of Recommendations Assessment, Development and Evaluation; PICO: Population, Intervention, Comparison, Outcome; RCT: Randomised Controlled Trials; SoFs: Summary of Findings; "[…]" indicates that a t text has been removed for brevity or anonymity

Complexity in systematic reviewing

The thematic analysis identified three themes related to complexity in systematic reviewing: *the need to include heterogeneous studies in terms of PICO elements and types of study designs, the need to ask broad review questions*, and *flexibility in the choice of methods and approaches* (see Table 4.3).

Among the key sources of heterogeneity of evidence identified by review authors and GRADE methodologists were the presence of different components within interventions, the presence of multiple health and social outcomes associated with interventions, differences in the context and implementation of interventions across individual studies, and finally, the need to use diverse sources and types of evidence:

> *"[…] This review had a very complicated intervention. We were looking to test the impact of various implementation strategies, so the studies were very heterogeneous. They included randomised and nonrandomised designs […]"* (Author 2).

Participants perceived these sources of complexity as affecting the entire review process. For example, four participants highlighted the need to ask broad review questions, which is commonly known as a "lumping approach" (Caldwell & Welton, 2016), in order to collect comprehensive evidence on the topic and generalise it across settings. Eight review authors mentioned that the heterogeneity of evidence prevented them from pooling estimates of the effect in a meta-analysis and that they were obliged to synthesise the results narratively (Popay et al., 2006). Furthermore, participants frequently reported the need to use a range of methods to assess and interpret this heterogeneous evidence. In this regard, a few review authors noted that the methodological expectations imposed by the traditional Cochrane methodology in terms of requiring review authors to focus on certain types of evidence (most commonly RCTs)

and specific tools, such as the Cochrane Risk of Bias (RoB) tool, did not allow them to adequately address and interpret sources of complexity in their review:

> *"[…] If there could be a little bit more flexibility and scope for creativity within the Cochrane process to let the authors propose their own approach to summarising the evidence, and not necessarily to have to use GRADE or any other criteria or scale, I would tend to lean towards that […]".* (Author 3)

It should be noted, however, that this flexibility in the choice of methods was also seen by participants as potentially distorting the consistency of the review process. In this view, use of broader frameworks, and, specifically, the GRADE approach, was perceived by an author and GRADE methodologist to serve as an anchor for the review process, facilitating structured conceptualisation of important constructs such as the quality (certainty) of evidence:

> *"[…] All the studies included in the review need to be assessed for quality [...] It might be possible to have non-Cochrane tools feeding into the GRADE assessment on the study limitations [another term for risk of bias] criterion, but those would still need to be tools that would assess risk of bias [...] And I think, its risk of bias rather than any sort of related nebulous notions of study quality or anything like that, that needs assessing here in terms of the tools that are used."* (Author 8 & Methodologist)

Familiarity with and perceived utility of GRADE

Three further themes described participants' familiarity with and perceived utility of GRADE in reviews of complex interventions. Specifically, these themes described *awareness of the GRADE criteria for rating the certainty of evidence in the context of systematic reviews versus. GRADE Evidence to Decision (EtD) considerations used in the context of guideline development*, and regarding the utility of GRADE – *the use of the different domains of evidence for rating the certainty of evidence* and *the provision of a structured summary of key review findings.*

While GRADE methodologists defined the remit of GRADE as a system for rating the certainty in the estimates of intervention effect produced in systematic reviews, a few review authors were unaware of GRADE's application in a guideline development context—specifically, how GRADE certainty of evidence ratings are incorporated with other criteria to develop practice recommendations (see GRADE Evidence to Decision (EtD) frameworks (Alonso-Coello et al., 2016)). Part of the frustration with using GRADE in reviews of complex interventions was consequently attributed to the perceived failure of GRADE to account for important criteria relevant for a decision-making context, rather than a systematic review focusing on intervention effectiveness:

> *"I don't want to imply that our data were perfect or fantastic [...] But I guess my issue is the breadth of considerations that goes into GRADE. So, they are all methodological, and I just think as soon as you start thinking about public health interventions you are stepping outside of methods and you need to be able to take into consideration things like the magnitude of the problem for a population and even the cost-effectiveness of the intervention relative to other approaches [...]" (Author 3)*

In this view, many participants, including the GRADE methodologists made a case for separating the domains of the certainty of evidence rating in systematic reviews from those considerations that would enhance the understanding of the impact of a complex intervention more broadly. Those considerations were perceived to be important, however, beyond the scope of a single review and more relevant in a guideline development context:

> *"The way I have been trying to think about GRADE at the moment is that GRADE absolutely assesses your certainty in the estimate of effect by statistical association; it's really time to capture that element of it. But there may be other things that would affect your understanding of that effect, really […] I think it is a big question, whether GRADE is the right thing to do that at all, because we need a lot more information to fully understand complex interventions […] And I think once you've done the GRADE assessment you may then think, and this may just be*

> *from a guideline perspective […] Well, how do people feel about receiving this intervention/service? […] Is it acceptable to them, does it meet their needs? […] And some of this may or may not be relevant to a systematic review, but certainly a key to working in a guideline world." (Methodologist 2)*

Both review authors and GRADE methodologists appreciated the structure and transparency that GRADE offers for rating the certainty in the estimates of effect for separate outcomes. A few participants highlighted the value of GRADE in integrating various domains, such as inconsistency and imprecision, in addition to risk of bias when conceptualising certainty of evidence.

> *"Clinicians who engage with research on a more casual basis compared to academics have a tendency to look at whether or not something is outright statistically significant, or not, and so, I liked all the additional considerations [e.g., inconsistency and imprecision] GRADE brings to evaluation of [certainty] of evidence, and I like also that it goes beyond simply evaluation of risk of bias."* (Author 5)

Finally, while many participants appreciated the structure of the SoFs tables as a concise way to summarise and communicate key review findings, two authors raised concerns that the SoFs tables might, on the other hand, divert the readers from considering the important nuances of evidence of complex interventions that are not included in the tables:

> *"Some people only look at the [SoFs] tables, but that's a problem as well, as they miss significant information that isn't in [them]. As I was saying, I was quite shocked when I looked at the [systematic review] pdf file [...] Once I saw [the table], I wouldn't look further and would dismiss this evidence*." (Author 9)

Initial categorisation of evidence in GRADE

Three themes were identified relating to initial categorisation of evidence in GRADE. These included *inadequate differentiation of risk of bias from different NRSs,* such as from interrupted time series vs. case series, *integration of the Risk of Bias in*

Nonrandomised Studies of Interventions (ROBINS-I) tool in GRADE risk of bias assessment, and *challenges associated with the time and expertise for applying ROBINS-I*.

In general, review authors and GRADE methodologists agreed with the hierarchy of evidence approach, and concurred that RCTs ought to be regarded as the best study design in assessing intervention effectiveness when feasible. However, a few authors highlighted that, for complex interventions in particular, where RCTs might not always be feasible or ethical, authors should be using a range of evidence from different levels of the hierarchy, rather than limiting the scope of reviews to RCTs:

> *"What I find more of a problem is when we can't [do RCTs] … In the absence of that evidence, in a Cochrane review, we say, well, there is no evidence. But there is evidence, it is just not the type of evidence that we want to use."* (Author 7)

A few review authors further argued that the initial categorisation of evidence in GRADE does not allow adequate differentiation between the strengths and weaknesses of different types of non-RCT evidence. In this view, GRADE ratings were felt to produce a floor effect (Everitt, 2002):

> *"For me, the issue with GRADE was being able to differentiate between the strengths and weaknesses of non-RCTs [...] It's like having a, picking a test and the results are you either pass or fail, but it doesn't give you any sense of the distribution of knowledge for example, if you are taking the test. It's like pass or fail, period. So, the evidence fell into that low certainty rank, and wasn't very sensitive to describe elements of the studies that could inform people that are trying to use these kind of interventions."* (Author 10)

Related to this, both review authors and methodologists suggested that evidence from certain types of NRSs, such as quasi-experimental study designs (e.g., regression discontinuity) be initially rated as "moderate" or "high" certainty. Participants argued that these designs offer more rigour in the evaluation of the causal effects of interventions by using exogenous variation and, thereby, providing random variation in

the exposure of interest.

As a potential solution to the debates on the initial categorisation of evidence in GRADE, GRADE methodologists discussed integration of ROBINS-I in the GRADE assessment (Schünemann et al., 2018; Sterne et al., 2016). ROBINS-I offers a rigorous process for assessing risk of bias in nonrandomised studies, including the issues of selection bias and confounding, treating each study as an attempt to mimic a "target" trial. GRADE methodologists informed that there are ongoing discussions within the GRADE Working Group concerning whether or not the initial categorisation of evidence in GRADE based on study design should be dropped in reviews in which authors use rigorous tools to assess risk of bias in NRSs, such as the ROBINS-I tool (Schünemann et al., 2018). However, almost all of the participants who were aware of the ROBINS-I tool in this study found it very challenging and time-consuming to apply.

> *"The problem with that is that it is not an easy to use tool, we tried to use it and it is really difficult to use, and it is hard to get agreement between reviewers on how to use it, so from an operational standpoint, that tool has not gained traction."* (Methodologist 5)

Furthermore, concerns were raised that application of this tool in the GRADE approach still results in all types of non-RCT studies being downgraded to a "low" certainty level for risk of bias. A few GRADE methodologists suggested that the current version of ROBINS-I is designed for cohort studies and lacks signalling questions needed to adequately assess risk of bias in quasi-experimental study designs, such as interrupted time series and regression discontinuity.

> *"With regards to the ROBINS-I tool, I think quasi-experimental designs do fit into that framework and can be assessed, but I think what's lacking currently is the signalling questions that apply to those specific study designs within the ROBINS tool. I think the designers of this tool have acknowledged that they are aware of that shortcoming and some signalling questions were under development, but I*

> *am not sure what phase that work has reached. But it's a known limitation of that tool that it doesn't speak as well to quasi-experimental designs as to certain forms of nonrandomised nonexperimental studies."* (Author 8 & Methodologist)

Implementation of GRADE in the review

Three themes were identified regarding how participants implemented GRADE in reviews of complex interventions. These included how participants *situated GRADE in the review process*, *the use of judgment and expertise in GRADE ratings*, and *the time and resources* required for adequate implementation of GRADE.

In terms of situating GRADE in the review process, a few review authors reported implementing GRADE at the end of the review process, when evidence has already been synthesised. In this view, implementation of GRADE ratings at the end of the review was perceived to be tedious and untimely:

> *"I think one of my main concerns is that inevitably GRADE is something that you tend to use at the end of the project; in some reviews it just becomes an annoying add on you've got to do, a final step, a final hurdle you've got to go through."* (Author 7)

In the meantime, GRADE was conceived by GRADE methodologists as an approach that requires consideration from the beginning of the process to help frame the review questions and translate those questions into the eligibility criteria. From this perspective, use of GRADE from the outset of the review process was thought to be fundamental to improving the quality of systematic reviews:

> *"My view on GRADE is that it's not something that should be just thought about at the end of the process as an adjunct to the review in order to complete a SoFs table [...] It's a tool that you should always have in mind and also in terms of thinking about at the protocol stage, how to pre-specify or stratify perhaps the analyses that you are going to do in different ways, in order to than feed into those SoFs tables."* (Author 8 & Methodologist)

In relation to this, domains of the GRADE approach that were perceived by review authors as particularly challenging to implement in reviews of complex interventions (mostly inconsistency and indirectness) were often seen by GRADE methodologists as influenced by earlier review stages (such as how review questions are framed and important elements are specified in the first place to further inform the structure of the SoFs tables):

> *"I really felt like GRADE was not the right one for the type of interventions that we were looking at […] I don't think I have a clear solution, but it definitely had to embrace, well, just picking on that one domain of inconsistency, not to reduce this idea of inconsistency to a methodological problem [to downgrade certainty of evidence], but rather, a real and genuine part of certain type of intervention."* (Author 3)

> *"I think for complex interventions how you connect the question that you have, a problem in clinical practice or public health, then you make your PICO, and you formulate your question, and then the intervention you are trying to assess, the different elements. I think it is more challenging to present that in a structured way and that you have to decide beforehand, maybe already in your protocol, if you want to do a more overall question [lumping], and you are not interested in specific [intervention] elements. But if you are interested in specific elements or groups of elements that is equal in all studies, then your SoFs table should reflect that […] And when it comes to the certainty of evidence, I think that your challenges are maybe in the inconsistency/heterogeneity, but it is part of how you structure your table and what you put together."* (Methodologist 1)

Further challenges in implementing GRADE were reported by review authors related to the required levels of methodological expertise and the need to use subjective judgement:

> *"I am not sure really on what basis I judge […] Like you really have to consider many things, which I find difficult when you are not in GRADE and in doing every day."* (Author 6)

Many GRADE methodologists, on the other hand, highlighted that appropriate implementation of GRADE should always involve a careful judgment. In relation to this, a

concern was raised by a GRADE methodologist on the often-observed low ratings of evidence as a result of a narrow understanding, "algorithmic" implementation of the GRADE domains of evidence, and a lack of critical judgment on part of review authors:

> *"One tendency that I see in people evaluating is that they think about bias and then they [easily] rate down […] You really don't need to rate it down unless there is something major, obvious that clearly would reverse the observed association […] Obviously, all is judgment, and I think the GRADE guidance says that it is judgment and actually, all the GRADE papers tell you that you cannot use ... E.g., a statistical test to rate down for publication bias, you cannot use a statistical test just to rate down for inconsistency. So, they always reinforce this issue about not doing these 'algorithmic' decisions, and it is a judgement."* (Methodologist 5)

Several review authors felt that they didn't have enough capacity within the review team to allow sufficient time and resources for using GRADE. Regarding existing guidance, one author mentioned lack of guidance on using GRADE, while one other author found the many publications on GRADE overwhelming to follow. A concern was also raised that much of the available guidance is not applicable to reviews of complex interventions:

> *"A lot of the material on GRADE, and I would say also for Cochrane as a whole, there are so many rules first of all, but then, if you are doing observational complex interventions, so many of the rules actually don't apply, or are moot, and so sifting through all of that constitutes its own project in and of itself. So, at the end of the day, I really learned the bare minimum that I needed to know about GRADE to try to make the best case for the value of our literature."* (Author 3)

A suggestion was made by a GRADE methodologist to rebalance the work in the review process so that review authors spend more time on implementation of GRADE as opposed to other phases of review, such as screening and data extraction:

> *"I see people doing it [GRADE ratings] very quickly, at the end, without much thought, because they spent so much time on all the other phases, and some of that time that is spent earlier is really useless time. Like you spend months going through abstracts that are completely irrelevant to the topic [...] So, I think the work in a systematic review should be rebalanced, where this phase of going*

> *through thousands of abstracts, which is completely waste of time, should be done quickly and then, a lot more time spent on the actual studies that are included to look at them in a more thoughtful way."* (Methodologist 5)

Suggestions for enhancing the GRADE guidance for complex interventions

Participants came up with different suggestions on how to enhance the GRADE approach for reviews of complex interventions. These suggestions can be classified into those related to the use of a more tailored *guidance and examples targeting complex interventions*, and those requesting *additional possibilities to upgrade evidence* in the GRADE approach.

A few methodologists highlighted that many of the reported challenges of using GRADE in reviews of complex interventions (Movsisyan, Melendez-Torres, & Montgomery, 2016; Rehfuess & Akl, 2013) can be ameliorated by a more integrated and critical use of GRADE from the beginning of the review process, at the stages of formulation of review questions, description of intervention components, and a priori specification of effect modifiers. These steps were perceived to be key in further informing the structure of the SoFs table:

> *"I think part of our guidance is that we should request the reviewers to [describe intervention components by included studies], because I don't know how can I make good judgments without looking at this. And it might give you reasons for different sensitivity analyses to determine, which are the essential components. And then, the essential components would drive the [GRADE] 'directness' assessment."* (Methodologist 4)

In this view, both review authors and GRADE methodologists felt that the GRADE guidance could be further enhanced through examples focusing on complex interventions and more explicit description of how judgments are made with regard to

each distinct domain of GRADE. Review authors also noted that the guidance could be enhanced by further examples of GRADE ratings for a body of evidence comprised of both randomised and nonrandomised designs (often referred to as "mixed evidence"). Finally, a few participants suggested the need for examples on how different methods for analysing evidence on complex interventions can inform the GRADE ratings. For example, one participant highlighted the potential utility of the Qualitative Comparative Analysis (QCA) in exploring sources of heterogeneity in reviews of complex interventions and, thereby, informing judgements on inconsistency in GRADE. It is worth noting that QCA aims to identify configurations of participant, intervention, and contextual factors that may explain the observed effects by way of cross-tabulation of evidence in comparison to quantitative sub-group analyses (Thomas, O'Mara-Eves, & Brunton, 2014).

> *"What there is a need for the guidance on is methods and systems for implementing these analyses [such as QCAs], which often are feasible but not done, the types of analyses that we've done in our review may be feasible in a lot more reviews than there have been tried in."* (Author 8 & Methodologist)

Many participants requested extended possibilities within the GRADE framework for upgrading the body of evidence, such as when consistent results are found across different study designs and settings. Reservations were made by GRADE methodologists, however, that the GRADE Working Group might be resistant to these major changes to the GRADE structure, and that to make progress, these proposals needed to be substantiated by convincing examples and evidence:

> *"It would be interesting to give us an example, where you feel consistency warrants more certainty […] Again, some people talk about if there is an RCT and observational studies, and they are all telling you the same thing, but usually in this case you will not downgrade, you know. Usually in this case, when they are all telling you the same thing, the overall quality [certainty] of evidence it might be based on the higher one. I don't see why you would increase your certainty."* (Methodologist 4)

Participants also noted the importance of considering the role of theory in reviews of complex interventions. However, there were disagreements and uncertainties on whether theoretical considerations should be directly incorporated into the GRADE approach, for example, as a domain to upgrade evidence in the presence of strong pre-specified mechanistic reasons for an effect. While a few participants leaned towards the latter argument, arguments were also raised that considerations of theory of change and use of logic models may be very informative in reviews of complex interventions, but should not be directly addressed within the GRADE framework. Rather, they should be addressed by guidance on good practice in designing systematic reviews of complex interventions more broadly:

> *"The other obvious area that to me seems missing, is the role of theory. So, if you have a very-very strong mechanistic reasons for expecting an effect, which are pre-specified in some way, and then you find an effect, then, should that be considered in terms of upgrading or not?"* (Methodologist 3)
>
> *"I am not so sure that [theory] needs to be part of GRADE; I think that theory of change and logic models, etc. are very useful in these types of reviews and we used one in this review. It is important to specify so far as possible and operationalise in these types of tools in systematic reviews of complex interventions, but I don't know if we need GRADE to address it, and I think that should be addressed by guidance on good practice in designing the systematic reviews."* (Author 8 & Methodologist)

Discussion

This qualitative study expands on the investigation into using GRADE in systematic reviews of complex interventions (see Chapter 1), provides a more detailed and nuanced picture of the challenges of applying GRADE in these interventions, and elucidates further suggestions for developing a GRADE guidance for complex interventions.

Overall findings

The findings of this study confirm that systematic review authors encounter a range of challenges when reviewing complex interventions in general, and when using GRADE more specifically (Movsisyan et al., 2016; Rehfuess & Akl, 2013). This is consistent with past qualitative studies that explore authors' understanding and management of complexity in systematic reviews (Anderson et al., 2013; Lorenc et al., 2016; Shepherd, 2013). Review authors frequently report the need to take a more inclusive approach in reviews of complex interventions by way of asking broader questions, considering evidence from a range of study designs, and using different approaches to synthesise evidence, such as narrative synthesis beyond "conventional" meta-analysis (Melendez-Torres et al., 2017; Popay et al., 2006). In their interview-based study of researchers' practices with complexity, Lorenc et al. (2016) describe that reviewers of complex interventions use the existing methods and guidance *"in a form of bricolage":* that is to say, reviewers need to apply a pragmatic approach to decide on the relevance of the methods to their review question or data from a range of available approaches (Lorenc et al., 2016). Constraining the possibilities of reviewers in terms of imposing specific review inclusion criteria, as well as tools and guidance to use to appraise and synthesise evidence, might, therefore, create genuine barriers to conducting and interpreting the results of systematic reviews of complex interventions (Petticrew et al., 2017).

A few important discrepancies were noted in the views on GRADE use between review authors and GRADE methodologists, which are worth discussing. First, review authors and GRADE methodologists disagreed on the "positioning" of the GRADE approach in the review process. Guidance on the conduct of systematic reviews commonly describes systematic reviewing as a linear process, where the preceding

stages of reviewing inform the following stages (Gough, Oliver, & Thomas, 2012; "A guide to conducting systematic reviews," 2016). The review process starts with the formulation of a question followed by defining the inclusion and exclusion criteria, searching and selection of studies, data extraction, risk of bias assessment of individual studies included in the review, evidence synthesis, and discussion of findings. Following this logic, review authors reported applying GRADE at the end of the review process, specifically, after evidence synthesis and when summarising data and writing discussions. Meanwhile, GRADE methodologists supported an integrated use of GRADE from the beginning of the review process, when deciding on the scope of the review, specifying the intervention and the potential effect modifiers to inform further subgroup analyses. GRADE methodologists found it critical for appropriate application of GRADE to think about evidence assessment and the structure of the SoFs tables at an earlier stage.

This view is in line with the arguments against portraying the systematic review process as a progressive and linear trajectory of action, as discussed by Moreira et al. (2007) in their ethnographic study of the processes of knowledge making in systematic reviews (Moreira, 2007). Specifically, authors found that knowledge production in systematic reviews follows a parallel process of evidence *"disentanglement"* (i.e., extraction of data from their original context) and evidence *"qualification"* (i.e., transformation of data into new knowledge). Furthermore, fundamental to the generation of new knowledge in reviews is the malleability of the process, that is, the ability of reviewers to adjust and link the various stages of the review during the course of the research. This appears to be even more relevant for reviews of complex interventions, which often lack common definitions and implementation procedures requiring additional efforts from reviewers in formulating questions, locating and making

sense of the heterogeneous evidence (Kelly et al., 2017; Rohwer et al., 2017; Squires, Valentine, & Grimshaw, 2013).

In contrast to the views of GRADE methodologists, the need to exercise critical judgment was found challenging by review authors. In general, the developers and proponents of the GRADE approach are against using the GRADE framework as a "checklist" approach and argue for the need to use critical judgment in GRADE ratings based on the context and purpose of the review. By way of illustration, the guideline for judging the GRADE domain of inconsistency discusses the decision to downgrade evidence in a systematic review of an intervention, which shows wide variation in point estimates, appreciable non-overlap of confidence intervals, a statistically significant heterogeneity, and high I^2 (65.1%). Furthermore, the a priori hypotheses failed to explain the inconsistency:

> *"Despite the observed inconsistency, the decision to rate down is not straightforward. All studies, with one exception, favour treatment. The inconsistency is, therefore, almost completely between studies that show moderate, large and very large effects. Thus, although there is large inconsistency, the importance of the inconsistency for decision making is uncertain. Whether to rate down quality [certainty] is, therefore, a matter of judgment"* [Guyatt et al., p. 1297] .

Evidence assessment necessarily involves interpretation and subjective judgment based on tacit knowledge (Chan, Macdonald, Carnevale, Steele, & Shrier, 2017; Polanyi, 1966). In this view, Toye et al. (2013) compare quality interpretation in systematic reviews to the experience of *"trying to pin down jelly"*, highlighting the interpretative aspect of the process (Toye et al., 2013). In many instances, however, review authors feel unconfident in using subjective judgment, as might be expected by GRADE developers and methodologists, and report relying on external advice (such as from their Cochrane

review group) in making those decisions. Difficulties in exercising judgment in systematic reviews have also been previously noted by Shepherd (2013), specifically, among novice review authors, while those with more training in research methods are generally found to be better at developing quality appraisal skills. While this suggests that systematic reviewing might require substantive training and possibly a higher education qualification, little research has been conducted to date on the necessary prerequisites for conducting systematic reviews, including the required time and resources, expertise and training of review authors, and size and composition of review teams (Uttley & Montgomery, 2017).

Implications for the GRADE guidance for complex interventions

Findings from this qualitative investigation have important implications for developing a new GRADE guidance for complex interventions and for the practice of systematic reviewing of complex interventions more broadly. As noted above, despite the expectations of the GRADE developers and methodologists, review authors are frequently challenged by the need to exercise subjective judgment and sift through many GRADE guidelines and might not always have sufficient expertise and resources to use GRADE "appropriately". With the increase of the demand for systematic reviews, there are currently more opportunities for training in skills of systematic reviewing, such as workshops organised by the Cochrane Collaboration and the webinars and online resources provided by the GRADE Working Group. As shown in this study, it is important that a more tailored training and guidance is available for systematic reviewers who investigate complex interventions. Since the complexity perspective has implications for the whole review process, it is essential that the training and guidance target the entire

machinery of systematic reviewing, including journal editors and peer-reviewers who serve as important promoters of research standards. Lack of such guidance, on the other hand, may hinder consistent use of GRADE across communities of reviewers posing further challenges for harmonious decision-making.

The findings show that lack of capacity within the review teams, in terms of scarce resources, time constraints, and small teams, can be another potential barrier to producing methodologically sound reviews. More investment into funding of systematic reviews may likely resolve some of the reported challenges by way of encouraging collaboration and larger review teams with engagement of experienced reviewers and methodologists (such as through communities or "crowds" of reviewers) (Thomas et al., 2017). Independent methodological support and team science are also suggested to enhance the transparency and reproducibility of research results (Munafo et al., 2017). In the meantime, it is important that designers of new methods and guidance consider the needs and circumstances of the immediate users in making them feasible for application. As shown in this study, review authors may often feel overwhelmed by many guidelines to follow during the review process; in this view, it is important that future guidelines, including the GRADE guidance for complex interventions are produced through collaborative efforts building on and consolidating the previous efforts, so that review authors only consult the most recent versions to save time and resources.

Although this study mainly focused on the use of the GRADE approach in systematic reviews of complex interventions, the new GRADE guidance for complex interventions may need to take a more inclusive approach in terms of highlighting how decisions at different stages of the review process can inform development of the SoFs tables and the GRADE ratings. Specifically, contrary to the current practice, where review

authors learn about GRADE and apply it at the end of the review process, an integrated use of GRADE from the beginning of the review process may ameliorate several reported challenges in GRADE use, including judgments on inconsistency and indirectness of evidence. As discussed in Chapter 1, review authors often felt uncertain on how to use the SoFs tables and the GRADE approach in general to accommodate the "inherent" heterogeneity and indirectness of complex interventions (Movsisyan et al., 2016). Considering GRADE at the outset of the review process will, therefore, inform whether the review has been appropriately framed to accommodate the structure of the SoFs tables and whether there is a need for further specification of intervention components and/or effect modifiers.

Results of this study show that there are ongoing methodological developments both within the GRADE Working Group and in the broader field of evidence synthesis, which might be highly relevant for the new GRADE guidance for complex interventions. As discussed in detail in the preceding chapters, one of the most frequently reported challenges in using GRADE in complex interventions is associated with GRADE's evidence hierarchy approach, that is, the initial categorisation of a body of evidence into "high" and "low" certainty of evidence categories based on the design of included studies (Movsisyan et al., 2016; Rehfuess & Akl, 2013). This initial categorisation of evidence is also perceived as an important barrier to using GRADE in wider social interventions, such as in Campbell reviews, as it is viewed to "unfairly penalise" high quality quasi-experimental studies. From interviews with the GRADE methodologists, however, it was found that there are ongoing discussions in the GRADE Working Group to drop this categorisation of evidence based on study design so that all bodies of evidence can be initially rated at "high" certainty. These discussions have mainly been held in light of the

recent development of the ROBINS-I tool as a rigorous approach to assessing risk of bias in nonrandomised studies of interventions (Sterne et al., 2016).

If this new approach to initial rating of a body of evidence is implemented, it may mitigate the challenges of the evidence hierarchy of the GRADE approach. However, it is worth noting that participants in this study found ROBINS-I very time-consuming and complicated to use. It might, therefore, be worth testing further the feasibility of this new approach, as well as any relevant alternatives, when developing the GRADE guidance for complex interventions. The new GRADE guidance should also follow the recent developments in the methods of reviewing complex interventions more broadly. Participants of this study informed about relevant ongoing initiatives for advancing the methods of synthesising complex interventions, including the Meta-analysis, Complexity and Heterogeneity (MACH) project (2017) and the WHO project on strengthening the process and methods for retrieval, synthesis and assessment of evidence on complex multidisciplinary interventions (2017). As these initiatives are yet under way, it is important to note that although timely, development of a new GRADE guidance for complex interventions will be provisional at this stage and subject for further refinements as the understanding of the methods for synthesising complex interventions increases.

Strengths and limitations

As an exploratory investigation, this study was able to provide a comparison of experiences of GRADE application among immediate users of GRADE (i.e., review authors) and developers of the GRADE guidance (i.e., GRADE methodologists). This cross-examination, therefore, helps to further specify the remit of the new GRADE guidance for

complex interventions by way of separating challenges of the GRADE methodology itself from those related to the procedures of systematic reviewing more broadly. Review authors were purposively sampled in this study based on their systematic reviews including interventions of different types and levels of target. While this study aimed to have a balanced representation of reviews undertaken both within Cochrane and Campbell Collaborations, the sample of reviews included in this study predominantly came from the Cochrane Library. As noted above, there were difficulties in locating Campbell reviews that used the GRADE approach; lack of uptake of GRADE in Campbell reviews can be partly explained by perceived difficulties in using GRADE in reviews of social interventions. It is, therefore, possible that the views and experiences of the considered sample of review authors may not wholly reflect those of the wider community of systematic reviewers of complex interventions, especially reviewers in social disciplines, where adoption of GRADE has not been as widespread as in the health sector.

Furthermore, despite the best efforts to ensure representation from low- and middle-income countries, the sample predominantly includes reviewers from high-income countries. In general, the scope of this investigation had to be limited, so that the DPhil candidate could analyse and summarise the data in time to inform the following online expert panel and the face-to-face expert meeting (see Chapters 5 and 6). This further hinders full assessment of data saturation, and it is possible that recruitment of additional review authors could generate new relevant themes. The sample of the GRADE methodologists was also limited by convenience and time constraints. It is likely that another group of methodologists could have a different view on complexity in systematic reviewing, considering existing different perspectives to defining complexity

(such as “complex interventions and “complex systems” (Shiell, Hawe, & Gold, 2008)). In this study, complexity was primarily viewed by participants based on the dimensions described in the 2008 UK Medical Research Council (MRC) framework (Craig et al., 2008).

There are concerns that telephone interviews might create challenges for establishing appropriate rapport with participants and capturing contextual data. These challenges, however, were largely mitigated in the present study, as it did not involve discussion of sensitive data, and participants themselves were keen to share their experiences with GRADE and talk about their area of research (Novick, 2008). Further evidence is available showing that telephone interviews do not reduce quality of data as compared to face-to-face interviewing (Sturges & Hanrahan, 2004). Finally, it should be noted that the coding of the data was conducted by the DPhil candidate alone, and the study did not consider measures of participant validation as recommended in the best practices of qualitative research, including seeking participant feedback on the accuracy of the data analysis and interpretation (Creswell & Miller, 2000). Nevertheless, versions of the themes were read and commented by two other co-investigators from the project on developing *GRADE Guidance for Complex Interventions* and revisions were made according to their feedback. In general, this study was reported following the Consolidated Criteria for Reporting Qualitative Studies (COREQ) checklist to enable explicit and comprehensive reporting of the research (see Appendix 6) (Tong, Sainsbury, & Craig, 2007).

Conclusions

To sum up, this qualitative research has augmented understanding of specific challenges of GRADE use in systematic reviews of complex interventions and delineated

suggestions for a new GRADE guidance for complex interventions. Interviews discussed in this chapter were also informative in terms of highlighting relevant initiatives and stakeholder groups in the field, who will be important to consult in subsequent phases of this thesis research. Results are consistent with the content discussed in the preceding chapters in supporting the need for a new GRADE guidance for complex interventions. Specifically, review authors will benefit from a single, integrated guidance, which consolidates relevant aspects of the existing GRADE guidance with tailored examples for complex interventions, as well as incorporates most up-to-date thinking on the methods and approaches to synthesising evidence of complex interventions. The following chapters of this thesis aim to further specify the content of this new guidance.

References

Alonso-Coello, P., Schunemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., . . . GRADE Working Group (2016). GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ, 353*, i2016.

Anderson, L. M., Petticrew, M., Chandler, J., Grimshaw, J., Tugwell, P., O'Neill, J., . . . Shemilt, I. (2013). Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol, 66*(11), 1205-1208.

Attride-Stirling, J. (2001). Thematic networks: an analytic tool for qualitative research. *Qual Res, I*(3), 385-405.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F., & Kolbe, L. J. (1985). Avoiding type III errors in health education program evaluations: a case study. *Health Educ Q, 12*(4), 315-331.

Caldwell, D. M., & Welton, N. J. (2016). Approaches for synthesising complex mental health interventions in meta-analysis. *Evid Based Ment Health, 19*(1), 16-21.

Chan, L., Macdonald, M. E., Carnevale, F. A., Steele, R. J., & Shrier, I. (2018). Reconciling disparate data to determine the right answer: A grounded theory of meta analysts' reasoning in meta-analysis. *Res Synth Methods*, *9*(1), 25-40.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, P. (2008). Developing and evaluating complex interventions: new guidance. Retrieved 8 Feb, 2018 from https://www.mrc.ac.uk/documents/pdf/developing-and-evaluating-complex-interventions/

Creswell, J., & Miller, D. (2000). Determining validity in qualitative inquiry. *Theory Pract, 39*(3), 124-130.

Denzin, N., K., & Lincoln, Y., S. (2011). *The Sage handbook of qualitative research* (4th ed.). London: Sage Publications Ltd.

Everitt, B. (2002). *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press.

Flick, U. (2009). *An introduction to qualitative research* (4th ed.). London: Sage.

Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. London, UK: SAGE Publications Ltd.

A guide to conducting systematic reviews. (2016). Cornell University Library. Retrieved 12 Oct, 2017 from http://guides.library.cornell.edu/c.php?g=459012&p=3137889/

Guyatt, G., H., Oxman, A., D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., . . . GRADE Working Group. (2011). GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol, 64*(12), 1294-1302.

Higginson, A., Benier, K., Shenderovich, Y., Bedford, L., Mazerolle, L., & Murray, J. (2015). Preventive interventions to reduce youth involvement in gangs and gang crime in low-and middle-income countries: a systematic review. *Campbell Syst Rev, 18*.

Kelly, M. P., Noyes, J., Kane, R. L., Chang, C., Uhl, S., Robinson, K. A., . . . Guise, J. M. (2017). AHRQ series on complex intervention systematic reviews-paper 2: defining complexity, formulating scope, and questions. *J Clin Epidemiol, 90*, 11-18.

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., . . . Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

Lorenc, T., Felix, L., Petticrew, M., Melendez-Torres, G. J., Thomas, J., Thomas, S., . . . Richardson, M. (2016). Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Syst Rev, 5*(1), 192.

Melendez-Torres, G. J., O'Mara-Eves, A., Thomas, J., Brunton, G., Caird, J., & Petticrew, M. (2017). Interpretive analysis of 85 systematic reviews suggests that narrative syntheses and meta-analyses are incommensurate in argumentation. *Res Synth Methods, 8*(1), 109-118.

Miles, B. M., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2), e1000217.

Moreira, T. (2007). Entangled evidence: knowledge making in systematic reviews in healthcare. *Sociol Health Illn, 29*(2), 180-197.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

Munafo, M., R., Nosek, B., A., Bishop, D., V., M., Button, K., S., Chambers, C., S., Percie du Sert, N., . . . Ioannidis, P., A. (2017). A manifesto for reproducible science. *Nat. Hum. Behav., 1*(0021).

Novick, G. (2008). Is there a bias against telephone interviews in qualitative research? *Res Nurs Health, 31*(4), 391-398.

Ormston, R., Spencer, L., Barnard, M., & Snape, D. (2014). *The foundations of qualitative research.* In Ritchie J., Lewis J., Mcnaughton-Nicholls C., Ormston R. (Eds). Qualitative research practice. London: Sage Publications Ltd.

Patton, M. (2002). *Qualitative evaluation methods*. CA: Sage Publications Ltd.

Petticrew, M., Shemilt, I., Lorenc, T., Marteau, T. M., Melendez-Torres, G. J., O'Mara-Eves, A., . . . Thomas, J. (2017). Alcohol advertising and public health: systems perspectives versus narrow perspectives. *J Epidemiol Community Health, 71*(3), 308-312.

Polanyi, M. (1966). *The tacit dimension*. London: Routledge.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rogers, M., . . . Duffy, S. (2006). Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme. Retrieved 17 Oct, 2017 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.3100&rep=rep1&type=pdf

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rohwer, A., Pfadenhauer, L., Burns, J., Brereton, L., Gerhardus, A., Booth, A., . . . Rehfuess, E. (2017). Series: Clinical Epidemiology in South Africa. Paper 3: Logic models help make sense of complexity in systematic reviews and health technology assessments. *J Clin Epidemiol, 83*, 37-47.

Saldana, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Los Angeles, Calif.; London: SAGE publications.

Schünemann, H. J., Cuello, C., Akl, E. A., Mustafa, R. A., Meerpohl, J., Thayer, K., . . . GRADE Working Group. (2018). GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomised studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol*. In Press.

Shepherd, J. (2013). Judgment, resources, and complexity: a qualitative study of the experiences of systematic reviewers of health promotion. *Eval Health Prof, 36*(247-67).

Shiell, A., Hawe, P., & Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *BMJ, 336*(7656), 1281-1283.

Squires, J. E., Valentine, J. C., & Grimshaw, J. M. (2013). Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol, 66*(11), 1215-1222.

Sterne, J. A., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., . . . Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919.

Sturges, J. E., & Hanrahan, K. J. (2004). Comparing telephone and face-to-dace qualitative interviewing: a research note. *Qual Res, 4*, 107-118.

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., . . . Living Systematic Review, N. (2017). Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol, 91*, 31-37.

Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Syst Rev, 3*, 67.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care, 19*(6), 349-357.

Toye, F., Seers, K., Allcock, N., Briggs, M., Carr, E., Andrews, J., & Barker, K. (2013). 'Trying to pin down jelly' - exploring intuitive processes in quality assessment for meta-ethnography. *BMC Med Res Methodol, 13*, 46.

Uttley, L., & Montgomery, P. (2017). The influence of the team in conducting a systematic review. *Syst Rev, 6*(1), 149.

Wilson, D., B., Gill, C., Olaghere, A., & McClure, D. (2016). Juveline curfew effects on criminal behavior and victimization. *Campbell Syst Rev, 3*.

Chapter 5. An online expert panel

Rating the certainty of evidence in reviews of complex interventions

A paper adaption of this chapter has been submitted to *Implementation Science*

Chapter overview

This chapter describes an online expert panel to explore areas of agreement and disagreement around the content of the GRADE guidance for complex interventions. An online-modified Delphi method called ExpertLens^{TM} was used, and a multi-stage strategy was employed to recruit participants meeting pre-defined eligibility criteria based on the targeted stakeholder groups. Participants reviewed, rated and discussed 50 criteria for inclusion in the guidance. These criteria were identified from the systematic review and the qualitative interviews reported in the previous chapters. Quantitative ratings were analysed using the RAND/UCLA Appropriateness Method for determining agreement, and the reasons for high and low ratings were examined based on participants' free-text comments using a cross-case thematic analytic approach. The pre-analysis plan for this chapter is available on the Open Science Framework (https://osf.io/eb9ec/).

Of 468 stakeholders invited, 114 participated in at least one round of the panel. No significant disagreement was identified among participants on any criterion. Most of the criteria (32 of 50) were rated as *critically important* for consideration in the guidance, while 17 criteria as *important, but not critical*. Only 1 criterion was rated as of *limited importance,* specifically, the suggestion to drop the initial categorisation of evidence based on study design. Participants' comments helped to elucidate the reasons behind the quantitative ratings.

Introduction

There is a growing body of literature discussing the challenges of, and aiming to provide additional guidance for addressing complexity in systematic reviews (Guise, Chang, Butler, Viswanathan, & Tugwell, 2017; Lorenc et al., 2016; Petticrew, Anderson, et al., 2013). Most frequently, complexity in reviews has been described by referring to the multiple interacting components of interventions, differences in the implementation across contexts and settings, as well as difficulties in the behaviours required by those providing and receiving the intervention (Craig et al., 2008; Lewin et al., 2017). The more recent discourse, however, highlights sources of complexity, which are informed by complex systems thinking and are viewed to further compound the challenges of systematic reviewing, such as nonlinear dynamic relationships between the intervention and the outcome, emergent properties and multiple interactions among the intervention, its implementation and contextual factors (Petticrew et al., 2017; Pfadenhauer et al., 2017). These sources of complexity may pose challenges for systematic reviews in relation to (1) scoping reviews and defining interventions (Kelly et al., 2017; Squires, Valentine, & Grimshaw, 2013); (2) locating evidence (Guise et al., 2014); (3) synthesising evidence (Petticrew, Rehfuess, et al., 2013; Viswanathan et al., 2017); and finally, (4) assessing and rating the certainty of evidence (Movsisyan, Melendez-Torres, & Montgomery, 2016; Rehfuess & Akl, 2013).

Standard frameworks and methods for systematic reviewing initially developed and validated in biomedicine often do not adequately address these sources of complexity (Kelly et al., 2017). Discussed in detail in the previous chapters of this thesis, review authors encounter challenges with regard to using the established domains and

criteria of the GRADE approach for assessing the certainty of evidence produced in systematic reviews of complex interventions (Movsisyan et al., 2016). These challenges are specifically discussed in relation to the need to employ a range of study designs, beyond randomised controlled trials (RCTs) to evaluate complex interventions, as well as uncertainties around how to assess indirectness and inconsistency of evidence, when intervention effects are highly contingent on implementation and contextual factors. As demonstrated in the previous chapter, uncertainties associated with the need to manage this heterogeneous and often "messy" evidence hinders review authors in their judgements creating barriers to implementing and interpreting the GRADE ratings. It this view, it was argued that a tailored GRADE guidance, which builds on the knowledge from the existing GRADE guidelines, meanwhile broadening the scope of the conceptual and analytic approaches that may be used in systematic reviews of complex interventions can help alleviate these challenges, better articulate the value of, and provide more confidence to reviewers in using GRADE.

According to the recommended techniques for developing research reporting guidelines (Moher, Schulz, Simera, & Altman, 2010), this chapter employs an online expert panel methodology (Khodyakov et al., 2011) to explore stakeholders' opinions about the content of this new GRADE guidance for complex interventions. Using the evidence generated from the previous phases of the thesis work (see Chapters 1, 3 and 4), this study asked a diverse and purposively selected group of relevant stakeholders with experience in systematic reviewing and/or interest in complexity to assess the importance of considering different criteria when rating the certainty of evidence in reviews of complex interventions. Expert panels are an established method to examine group consensus in clinical and health services research (Fink, Kosecoff, Chassin, & Brook,

1991; Jones & Hunter, 1995; Khodyakov, Grant, et al., 2017). They follow a modified Delphi method, which involves a structured process of obtaining information from a group of experts by employing a series of questionnaires (Fitch et al., 2001). After each round, group responses are summarised and fed back to participants. In addition, expert panels also include a discussion round, where participants have an opportunity to interact regarding their ratings (Fitch et al., 2001; Khodyakov et al., 2011).

The aforementioned features of the expert panel method provide a number of advantages over focus groups and traditional Delphi panels (Fitch et al., 2001). First, the anonymity of responses mitigates socio-cognitive biases, which may lead to conformity in open group discussions. Second, by using a series of questionnaires and a controlled feedback, participants have an opportunity to appreciate the group opinions in a more relaxed environment as compared to a focus group format. Third, the discussion round allows participants to discuss their responses with each other, and, thereby, to modify their views in light of collective opinions. Moreover, conducted online, expert panels can further enhance the efficiency of the process in allowing to engage a more diverse and geographically dispersed group of stakeholders from different countries to contribute to the process at time convenient to participants, and finally, to make online discussions anonymous, and thus, reduce possible biases based on participant status and personal characteristics (Khodyakov, Grant, et al., 2017; Khodyakov et al., 2011).

The methodology adopted for this thesis, which largely extends on the best practices for developing research reporting guidelines, recommends conducting a Delphi process after performing preliminary activities of literature review and consultation to synthesise the latest available scientific evidence on the new guidance (Moher et al., 2010). Delphi-based expert panels are increasingly employed in development and

adaptation of research reporting guidelines (Guise, Butler, et al., 2017; Montgomery et al., 2013), and have also been used in a few projects of the GRADE Working Group (2017). Correspondingly, this chapter reports on an exploratory online expert panel to identify areas of agreement and disagreement among stakeholders on the domains and criteria for rating the certainty of evidence in systematic reviews of complex interventions. Evidence presented and discussed in the previous two chapters of this thesis has been used to design this expert panel.

Methods

Recruitment

Targeted stakeholders for the expert panel involved the following stakeholders who work in the area of complex interventions across a range of social disciplines: (1) researchers conducting systematic reviews; (2) methodologists; (3) developers of guidelines who provide practice recommendations; (4) practitioners who deliver interventions in real-world contexts; and (5) policy-makers who use evidence for decision-making. To ensure that a wide range of opinions are included in the panel, a decision was made to consider both those stakeholders who had prior experience with using GRADE and those who didn't. Two separate panels were, therefore, convened to tailor the questionnaire in accordance with the stakeholders' familiarity with GRADE.

A multi-stage strategy was employed to recruit panel participants. First, the initial list of potential participants was constructed by the DPhil candidate over the course of 5 months through extensive searches on the websites of key stakeholder organisations, as well as author lists of relevant publications, including the systematic review reported in Chapter 3. Conference proceedings and member lists of pertinent research societies and

collaborations, such as that of the GRADE Working Group were also searched, while presentations at international conferences and workshops conducted before the launch of the panel further helped to identify relevant stakeholders (see Dissemination of the thesis work). This initial list was then circulated to the co-investigators of the project on developing *GRADE Guidance for Complex Intervention* for a check and further nomination of participants. Finally, a snowball sampling approach was used to extend the professional network of the immediate project team, which involved requesting the identified experts to nominate further participants (Marshall, 1996). On the basis of the previous research on online expert panels, the intended sample size for each panel was 30-40 participants (Khodyakov et al., 2011). To reach this goal, an invitation email was sent to 468 stakeholders with a request to register for the three-round online expert panel by providing their demographic information. Only those who registered and consented to participate were sent the link to the Round One questionnaire.

Design

The recruitment of participants was conducted between 4 March and 7 April 2017 using SelectSurvey. This survey collected demographic information on country of residence, stakeholder group membership, discipline/policy area, years of experience in the area of complex interventions, and familiarity with the GRADE approach.

The online-modified Delphi expert panel ran from 22 March to 16 May 2017 using ExpertLens^{TM} – a mixed-method online system, which combines two rounds of structured questionnaires with a feedback, and one round of asynchronous and anonymous discussion (Khodyakov et al., 2011). Participants were given anonymous usernames, and two concurrent panels were conducted: one with stakeholders who reported previous

use of the GRADE approach (Panel A), and the other with stakeholders who reported no previous use of the GRADE approach (Panel B). ExpertLens, developed by the RAND Corporation, was chosen for this study as it allows for efficient engagement of a large and geographically distributed number of stakeholders, as well as explicit integration of *quantitative* (ratings) and *qualitative* input (comments and online discussion) in understanding agreement and disagreement within the panel. It has been successfully used to run different panels, including for development of national suicide prevention research goals (Claassen et al., 2014), identification of defining features of continuous quality improvement in healthcare (Rubenstein et al., 2014), development of quality and performance indicators for patients with arthritis (Barber et al., 2015), and examination of patient engagement roles in planning and designing outpatient care improvements at the Veterans Administration Health-care System (Khodyakov, Stockdale, et al., 2017).

In Round One, stakeholders reviewed and rated 50 distinct criteria for rating the certainty of evidence in reviews of complex interventions. These criteria were generated based on the findings from the systematic review of evidence rating systems, as well as semi-structured interviews with review authors and GRADE methodologists presented in Chapters 3 and 4 of this thesis, respectively. These criteria were organised in accordance with the domains of the GRADE approach, specifically: initial categorisation of a body of evidence based on study design, study limitations or risk of bias, inconsistency, indirectness, imprecision, publication bias and the domain for upgrading certainty of evidence. The instructions to complete the questionnaire were modified for Panel B to accommodate stakeholders' unfamiliarity with the GRADE terminology and process. Each of the questionnaires were piloted with two researchers: the questionnaire for Panel A was piloted with two researchers from the project team, and the questionnaire for Panel

B was tested with two other researchers from the network of the DPhil candidate who had not used the GRADE approach. The final versions of the questionnaires incorporated the received feedback (see Appendices 7 and 8 for Round One questionnaires for Panels A and B, respectively).

Participants used a 9-point Likert scale to rate the importance of considering each criterion when rating the certainty of evidence in reviews of complex interventions. Participants were informed that their ratings would be interpreted as follows:

- Scores of 7 to 9 indicate that it is critically important to consider the criterion;
- Scores of 4 to 6 indicate that it is important, but not critical to consider the criterion;
- Scores of 1 to 3 indicate that it is of limited importance to consider the criterion

Participants were given an opportunity to provide a rationale for their ratings in free-text boxes beneath each Likert scale. Participants were also asked a few open-ended at the end of Round One questionnaire. These related to the additional rating criteria, features of complexity that pose challenges to rating the certainty of evidence, challenges associated with reviewing mixed bodies of evidence comprised of different study designs, personal requirements for rating the certainty of evidence, such as training, and finally, suggestions for disseminating and implementing the new guidance.

In Round Two, participants reviewed bar charts showing each participant their own response in relation to the distribution of responses of other participants. Participants were also provided with information on the group's agreement or disagreement on the importance of considering each criterion (see data analysis below for more information on the approach for determining the group disagreement). An

example of this is presented in Figure 5.1. Participants were also able to view the comments made for each criterion, and were invited to interact with each other in an anonymous and asynchronous online discussion forum. The DPhil candidate and a project co-investigator served as moderators of these discussions and encouraged a dialogue by way of asking participants to share their views on key themes from Round One responses.

In Round Three, participants were given an opportunity to revise their responses in consideration of Round One results and Round Two feedback and discussions (see Appendices 9 and 10 for distribution of Round One answers and Round Three questionnaires for Panels A and B, respectively). No major changes were made to the questionnaires in Round Three.

The approval to run the panel was obtained from the Departmental Research Ethics Committee at the University of Oxford (Ref: SPI_C1A_16_009).

Data analysis

To explore areas of agreement and disagreement on the criteria for rating the certainty of evidence, the two-step analysis technique described in the RAND/UCLA Appropriateness Method (RAM) User's Manual was applied to Round Three data (Basger, Chen, & Moles, 2012; Fitch et al., 2001). The first step in RAM's technique, includes determination of disagreement among participants by way of comparing the inter-percentile range (IPR) for each criterion, that is, a range of responses between the 70th and 30th percentiles to the inter-percentile range adjusted for symmetry (IPRAS; see Box 5.1 on how to calculate IPRAS). The latter provides a measure of dispersion accounting for asymmetric distributions. Typically, disagreement is considered when the

Disagreement Index (DI), calculated as the ratio of the IPR to the IPRAS, is greater than 1 (Fitch et al., 2001). This indicates an uncertain group decision regarding the importance of the criterion. However, if DI is less than 1, the analysis continues with the second step in RAM's technique, specifically, calculation of the median score to assign a group decision. In online expert panels, group decisions may be positive, negative or uncertain provided there is no group disagreement (Fitch et al., 2001; Khodyakov, Stockdale, et al., 2017). Correspondingly, to address the aims of this study, group decisions for each criterion were determined: a median score of 7 to 9 indicated that a criterion is critically important to consider when rating the certainty of evidence in reviews of complex interventions; 4-6 indicated that a criterion is important but not critical; and finally, 1-3 indicated that a criterion is of limited importance. Data from Panels A and B were combined to calculate the final group decisions.

All qualitative data from text boxes and Round Two discussions were analysed by the DPhil candidate using a cross-case thematic analytic approach to examine reasons for high and low ratings for each criterion (Miles & Huberman, 1994). First, participant comments and explanations were grouped based on the criterion and the domain they described and the specific category of scores they referred to (i.e., of limited importance, important but not critical, and critically important). The DPhil candidate then coded all the text inductively, line-by-line, to elucidate the common and the most prominent reasons for the scores of the criteria (see Appendix 11 for the full list of codes). This sequential analysis of the qualitative data from the participants' comments helped to explain and interpret the quantitative ratings (Pluye & Hong, 2014). Data were analysed using Microsoft Excel, Stata 13 and NVivo 11.

Domain 1. Initial certainty rating: study design

Criterion 4. Same rating across all studies

An initial rating of "high" for a body of evidence consisting of any type of study design

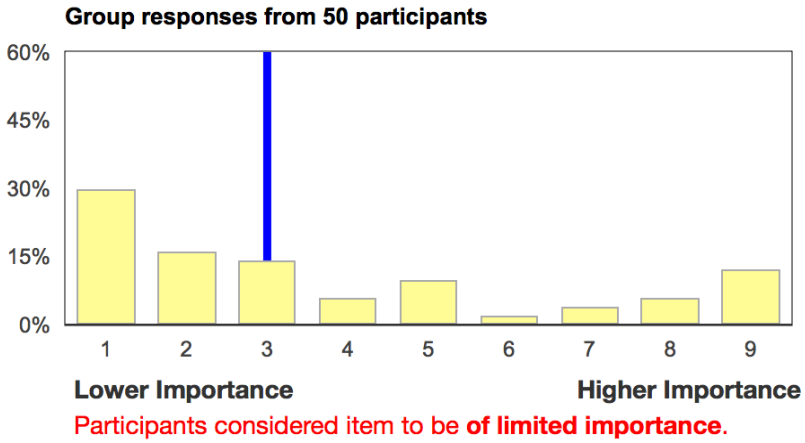


Figure 5.1. Distribution of Round One responses presented to Panel A participants in Round Two.

Notes: the height of the yellow bars indicates the number of participants choosing a particular response category; the blue line indicates the group median; in the original panel the graphs also had a red dot representing each participant's response.

Box 5.1. Calculation of group disagreement for each criterion (Fitch et al., 2001)

Lower Limit IPR = 30th percentile of the series of ratings
Upper Limit IPR = 70th percentile of the series of ratings
IPR = (Upper Limit IPR) – (Lower Limit IPR)
IPRCP (Central point of IPR) = Average of Upper Limit IPR and Lower Limit IPR
Asymmetry Index = (5) – (IPRCP)
IPRAS = 2.35 + (1.5*Asymmetry Index)
Disagreement Index (DI) = IPR/IPRAS

Results

Participants

Of 468 stakeholders invited, 156 (33%) responded in the recruitment survey that they were interested to participate in the online expert panel. Of these 156 stakeholders,

114 (73%) participated in at least one round of the panel: 102 of 114 (90%) participating stakeholders provided ratings in Round One, and 70 (61%) provided ratings in Round Three. Of 63 participants who logged into Round Two, 31 (49%) posted at least one comment. Overall, 229 comments were posted for discussion. Table 5.1 describes participants' characteristics.

The panel included a balanced number of female and male stakeholders predominantly from Europe (62%) and working in the field of public health (44%). Most frequently, stakeholders identified themselves as methodologists in the area of evidence synthesis (82%) and active in the area of complex interventions for more than 10 years (67%). Only one participant reported to be unfamiliar with GRADE: most of the participants reported using GRADE before (63%). Based on the Chi-square test (χ^2), no statistically significant differences were found on any of the participant characteristics outlined in Table 5.1 between (1) recruitment survey respondents who did and did not participate in the online panel, and (2) those participants who did and did not complete the Round Three. Participants in Panel A and Panel B were also similar in these characteristics for the exception of familiarity with GRADE and discipline: namely, participants in Panel A who reported using GRADE before were more likely to work in the field of clinical epidemiology. Rating results presented below are based on the analysis of the input from those participants who provided Round Three ratings (n=70; 61% of 114 participating stakeholders).

Ratings of criteria

The final ratings for each criterion are presented in Table 5.2. Within each domain, criteria are listed in order from highest to lowest median (and least to most

dispersion for criteria with the same median). No significant disagreement was identified among participants on any criterion (see Appendix 12). “Indirectness of the evidence base” (this was phrased as “inapplicability of the evidence base” for participants in Panel B) was the only domain for which all criteria were rated as critically important (i.e., median of 7 to 9), while “initial certainty rating: study design” was the only domain containing a criterion with a negative decision, that is, of limited importance for considering in reviews of complex interventions (i.e., median of 1 to 3). Among remaining domains, the majority of the criteria were rated as critically important in “limitation of included studies” (both for randomised and nonrandomised trials), “upgrading the initial certainty rating”, and “imprecision of effect estimates”. By contrast, the majority of the criteria were rated as important but not critical (i.e., median of 4 to 6) in “inconsistency of effects in the evidence base” and “publication bias” domains.

Table 5.1. Participant characteristics (N = 114)

Characteristics	N	%
Sex		
Female	59	52
Geographic region		
Europe	71	62
North America	33	29
Australia	4	3
Asia	3	3
Africa	2	2
South America	1	1
Discipline		
Public health	50	44
Clinical epidemiology	22	19
International development	6	5
Social work	4	3
Psychology	3	3
Criminology	2	2
Education	1	1
Sociology	1	1
Political science	1	1
Other	24	21
Stakeholder group[1]		
Methodologist	93	82
Researcher	78	68
Guideline developer	43	38
Journal editor	24	21
Policy-maker	12	11
Practitioner	12	11
Other	3	3
Activity in area		
More than 10 years	76	67
6-10 years	24	21
2-5 years	13	11
GRADE use		
Used GRADE	72	63
Familiar but no use	41	36
Not familiar	1	1

[1] Categories are not mutually exclusive.

Notes: categories within each variable are in order of most to least common response.

Table 5.2. Rating results

Domain of evidence		Criterion	Median	IQR	Decision
Initial certainty rating: study design					
1.	Randomised controlled trials	An initial certainty rating of “high” when the body of evidence consists of randomised controlled trials (RCTs)	8	6-9	+
2.	Non-experimental observational studies	An initial certainty rating of "low" when the body of evidence consists of non-experimental observational studies (e.g., cohort design, case-control study)	7	4.5-8	+
3.	Nonrandomised experimental studies	An initial certainty rating of "moderate" when the body of evidence consists of nonrandomised experimental study designs (e.g., natural experiments, quasi-experimental studies)	6	3-7	±
4.	Same rating across all study designs	An initial rating of "high" for a body of evidence consisting of any type of study design	3	1-4	−
Limitations of included RCTs					
5.	Participant attrition	Whether there is a significant amount of participant loss to follow-up and inadequacy of analytic methods for dealing with participant loss to follow-up	8	8-9	+
6.	Selective outcome reporting	Whether there is incomplete or absent reporting of some outcomes and not others on the basis of the results	8	8-9	+
7.	Quality of outcome measures	Whether studies used outcome measures with low validity and/or reliability to assess intervention effects	8	7-8	+
8.	Allocation concealment	Whether those enrolling participants are aware of the group (or period in a crossover trial) to which the next enrolled participant will be allocated (e.g. allocation by day of week, birth date, chart number, etc.)	8	7-9	+
9.	Blinding outcome measures	Whether those assessing outcomes are aware of the arm to which participants have been allocated	8	7-9	+
10.	Blinding data analysts	Whether those analysing data are aware of the arm to which participants have been allocated	7	5-8	+
11.	Stopping study early for benefit	Whether studies have been ended early when beneficial results were found in interim analyses	6	5-7	±

12.	Fidelity of intervention implementation	Whether there were deviations in intervention implementation from what was intended	6	5-7.5	±
13.	Blinding intervention recipients	Whether recipients of the intervention are aware of the arm to which they have been allocated	5	5-7	±
14.	Blinding intervention providers	Whether providers of the intervention are aware of the arm to which participants have been allocated	5	4.5-7	±
Limitations of included nonrandomised studies					
15.	Confounding	15. Whether the study used measures to adequately control for confounding	9	8-9	+
16.	Appropriate comparison group	16. Whether the study developed and applied an appropriate comparison group	9	8-9	+
17.	Measurement of outcomes	17. Whether the study appropriately measured outcomes in all groups	8	8-9	+
18.	Missing data	18. Whether the study used appropriate analytic methods for dealing with participant missing data	8	7-9	+
19.	Selection of the reported results	19. Whether there is selected reporting of effect estimates based on multiple measurements and analyses	8	7-9	+
20.	Classification of interventions	20. Whether intervention groups were clearly defined	8	7-9	+
21.	Selection of participants into the study	21. Whether the study used procedures to appropriately select participants into the study or into the analysis	8	7-9	+
22.	Follow up	22. Whether the study had adequate follow-up of participants	7	6-8	+
23.	Deviations from intended interventions	23. Whether there were deviations from the intended intervention beyond what would be expected in usual practice	6	5-7	±
Inconsistency of effects in the evidence base					
24.	Quantitative analyses exploring heterogeneity	Results of pre-specified quantitative analyses exploring moderators or methodological features that help explain heterogeneity (e.g., sub-group analyses, sensitivity analyses, meta-regressions)	7	5-8	+
25.	Qualitative analyses exploring heterogeneity	Results of qualitative analyses of evidence exploring varying effects of interventions that help explain heterogeneity (e.g., qualitative comparative analysis)	6	5-7	±

26.	Overlap of confidence intervals	The degree to which confidence intervals overlap across individual studies	6	4-7	±
27.	Variability in point estimates	The degree to which point estimates vary across individual studies	5	5-7	±
28.	Magnitude of statistical heterogeneity	The magnitude of the I^2 value, which indicates the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error	5	3-6	±
29.	Statistical test for heterogeneity	The magnitude of the P-value for a statistical test of the null hypothesis that all studies have the same underlying magnitude of effect	4	2-5	±
Indirectness of the evidence base					
30.	Indirectness of study populations	Degree to which the participants in included studies compare to the population of interest	8	7-8	+
31.	Indirectness of study interventions	Degree to which the interventions in included studies compare to the intervention of interest	8	6.75-8.25	+
32.	Indirectness of outcomes	Degree to which the outcomes considered in included studies compare to the outcomes of interest	8	7-9	+
33.	Indirectness of comparisons	Degree to which effect estimates are from comparison groups of interest	7	6-8	+
34.	Indirectness of follow-up timing	Degree to which the timing of outcome assessments in included studies compares to the follow-up time-points of interest	7	6-8	+
Imprecision of the effect estimates					
35.	Width of confidence intervals	35. Whether the confidence interval includes estimates of important benefit and harm	8	7-9	+
36.	Overlap of confidence intervals with line of no effect	36. Whether the confidence interval for the overall estimate includes both benefit and harm	7	3.5-7.75	+
37.	Optimal information size	37. Whether the total number of participants in the review meets a conventional sample size for a single adequately powered trial	6	5-7	±
Publication bias					
38.	Indexed literature search	The comprehensiveness of the reviewer authors' search of indexed literature to identify eligible studies	8	7-9	+
39.	Study sponsorship	Whether developers and purveyors of the intervention had influence on studies included in the review	8	6.5-9	+

40.	Discrepancies between published and unpublished studies	Results from any approaches to assess discrepancies in findings between published and unpublished studies	7	6.5-8	+
41.	Grey literature	The comprehensiveness of the reviewer authors' search of grey literature to identify eligible studies	6	5-7	±
42.	Funnel plot asymmetry	Whether there was evidence of funnel plot asymmetry	5	4-6.5	±
43.	Number of small studies	Degree to which the body of evidence consists of studies with small sample sizes	5	3.25-6	±
44.	Language of included studies	Whether authors applied restrictions to study selection on the basis of language	5	3-6	±
Upgrading the initial certainty rating					
45.	Dose-response relationship	Rating up certainty of evidence from nonrandomised studies when there is presence of a dose-response gradient	7	6-8	+
46.	Effect of plausible residual confounding	Rating up certainty of evidence from nonrandomised studies when all plausible residual confounding from nonrandomised studies are likely to reduce the demonstrated effect or increase the effect if no effect was observed	7	5-7	+
47.	Coherence of the evidence for the causal pathway	Rating up certainty of evidence when there is coherence of results in individual links in the causal pathway between intervention and distal outcomes	7	5-8	+
48.	Consistency across diverse contexts	Rating up certainty of evidence when there is consistent evidence on the effects of interventions across diverse contexts (e.g., various settings, geographical locations, study designs, outcome measures, research teams)	7	5-8	+
49.	Large magnitude of an effect	Rating up certainty of evidence from nonrandomised studies that yield large or very large estimates of the magnitude of an intervention effect	7	3-8	+
50.	Analogous evidence	Rating up certainty of evidence when there is supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanism(s)	5	2.75-6	±

Notes: +: “critically important”; ±: “important but not critical”; −: “of limited importance”. IQR: interquartile range.

Thematic analysis of participants' comments and discussions

Challenges of complexity

While participants frequently used the phrase "complex interventions" in their discussions, different views were expressed in the panel on how best to conceptualise complexity in systematic reviews. One participant offered to think of complexity in systematic reviews through key dimensions of complexity:

> *"I think it is time to stop talking about 'complex interventions' as this concept is too broad to be useful and no one agrees on what it means […] I think we will get further by looking at how specific dimensions of complexity impact on intervention implementation, fidelity, effects, etc." (Panel A, Participant 82)*

Similarly, another participant noted that complexity should be viewed as a perspective that review authors may adopt in each specific case regardless whether they are assessing a clinical intervention or a public health policy. In this view, a suggestion was made to think of complexity as an attribute of the review question, rather than as an attribute of the intervention itself. For example, if a "simple review question" examines whether an intervention is effective in a specific population and setting, a "complex review question" asks what components are effective for which populations and outcomes. Three other participants commented on conceptualising complexity from a systems' perspectives; however, this perspective was perceived to be premature in its methodological solutions to inform a guidance on an evidence rating system, such as the GRADE approach.

> *"GRADE is really geared towards looking at single/simple/linear links between an intervention and an outcome. It was not developed for the idea of 'complex interventions in complex systems', where we often expect multiple unmeasurably small effects across multiple aspects, system reactions (often dampening effects but sometimes leading to emergent properties such as sudden jumps). The problem is that we have not really found good methodological solutions to dealing*

with this at the primary research and systematic review level, hence, it is even more difficult to think of GRADE adaptations to address these challenges." (Panel A, Participant 49)

Participants identified several sources of "complexity", which were perceived to pose challenges to rating the certainty of evidence in reviews of complex interventions (see Table 5.3). These included variations in the evidence-base with regard to the design, contexts and implementation of interventions in individual studies, use of multiple intervention components and a range of outcomes, and the need to integrate evidence from diverse study designs. This variation was perceived to challenge the use of the traditional meta-analytic approaches, thereby, often creating the need to synthesise evidence without quantitatively pooling the estimates of the effect, and to additionally consider and integrate qualitative evidence. In this view, participants expressed the need for further guidance and criteria on how to rate the certainty of evidence in narrative synthesis (i.e., when estimates of the effect are not pooled in a meta-analysis) and in mixed-methods systematic reviews (i.e., when using quantitative evidence together with qualitative). Participants also felt that addressing aspects of complexity may require increased judgment in contrast to a "checklist-based" use of the rating criteria. Finally, the overall lack of adequate reporting and specification of key review elements in primary studies, such as the intervention, its implementation and contextual factors was perceived to further complicate the process of rating the certainty of evidence.

Table 5.3. Sources of complexity posing challenges to rating the certainty of evidence in reviews of complex interventions

Organising theme[2]	Basic theme	Example quotation
Heterogeneity in the evidence-base	Variation in intervention design	"When performing reviews of complex interventions, the packages of the interventions in the individual studies are not identical most of the time". (Panel A, Participant 99)
	Variety of interacting components	"Different components in a complex intervention can interact synergistically, antagonistically or independently - it is difficult to know which scenario is most likely in each circumstance". (Panel B, Participant 41)
	Variation in intervention implementation	"Often a need for flexibility and tailoring in complex interventions. Protocols need to allow for this and deviation within prescribed limits not considered a methodological weakness". (Panel A, Participant 21)
	Interaction of effects with context	"Complex interventions often interact with the actual research environment, which can be the political environment, the social environment, the economic environment etc. In most circumstances, this interaction cannot be determined". (Panel A, Participant 99)
	Variety of outcomes	"One level of complexity related to the number and range of outcomes that should be affected by the intervention". (Panel B, Participant 24)
	Use of diverse study designs	"The need to synthesise evidence from diverse nonrandomised and quasi-experimental studies, including less familiar designs such as natural experiments". (Panel A, Participant 11)
Need to use different (non-traditional) approaches to synthesise evidence	Integration of qualitative and quantitative evidence	"Literature on complex interventions often includes qualitative studies how do we account for that in GRADE?" (Panel A, Participant 39)

[2] These themes were extracted following the technique for conducting thematic analysis described by Attride-Stirling (2001)

	Lack of meta-analyses	"I anticipate that it will be difficult (or not appropriate) to meta-analyse evidence in complex interventions. I think it is essential to acquire the skills to synthesise the evidence in a narrative way and be able to integrate this information in the SoF tables. Please note that to, my knowledge, this training is difficult to find". (Panel A, Participant 42)
Need for increased judgement	Flexibility in judging the appropriateness of the criteria in each specific review	“I am concerned about comparison of questions according to body-of-evidence criteria when the applicability of those criteria is not uniform across question types. As a simple example a question should not be downgraded in any comparative analysis because intervention recipients are necessarily not blinded to the intervention”. (Panel B, Participant 48)
	Against using rating criteria as a “checklist”	“[Challenging] only if criteria are used as a tick-box. They should be used as a heuristic device, and then 'complexity' - when carefully described in publications - is not a challenge, nor is assessing the quality of evidence”. (Panel B, Participant 34)
Inadequate reporting of key elements of evaluation in primary studies	Lack of specification of intervention and its implementation	"Poor pre-specification of all features of an intervention. Usually intervention characterisation is in broad categorical sense with many shades in between". (Panel A, Participant 45)
	Lack of reporting on context	"Unless contextual influences on outcomes are identified by authors of primary studies, it will be difficult for reviewers to make these discriminations". (Panel B, Participant 23)

Implementation and dissemination of the guidance

Participants frequently noted that application of the rating criteria considered in the Delphi process, including those of the GRADE approach, may require substantive methodological expertise, such as a post-graduate qualification. As one observed, *"it is beyond the skill set of most systematic reviewers who are not usually methodologists. In particular, reviews of social interventions and policy interventions have little epidemiological knowledge which these methods assume."* In addition to training in research methods, and, specifically, in systematic reviewing and evidence synthesis, participants often emphasised the importance of content expertise. Related to this, a few participants highlighted the importance of working in review teams, where different roles may be taken by different members according to their area of knowledge:

> *"A thorough methodological background of a team member, in addition to the thorough clinical understanding of another review team member, who together do the GRADE-ing. You need both. Fine if less experienced persons (student, specialist in training etc.) do a first draft, but the senior should check and correct." (Panel A, Participant 36)*

Several activities were noted by participants on how best to disseminate the guidance for rating the certainty of evidence in reviews of complex interventions, including provision of training resources and opportunities like workshops and webinars, publication of the guidance in prominent journals, such as the *Journal of Clinical Epidemiology*, which publishes the official GRADE guidelines. The roles of journal editors were mentioned in promoting the use of the guidance, as well as partnering with the major organisations in evidence synthesis, such as AHRQ, Cochrane and Campbell Collaborations, the Guidelines International Network (G-I-N), the GRADE Working Group,

and WHO. Collaboration with the GRADE Working Group was particularly noted as a key factor in widespread dissemination of the guidance:

> *"Based on experience of CERQual [adaptation of the GRADE guidance for qualitative evidence synthesis], there is considerable advantage to be had from allying oneself to GRADE initiatives". (Panel B, Participant 44)*

Domains and criteria

Initial certainty rating: study design

As evidenced by the discussion comments, participants almost unanimously agreed that randomised controlled trials serve as the *"most appropriate study design"* or the *"gold standard"* for making a causal inference and assessing intervention effectiveness. Randomisation was generally perceived to provide the best control for all known and unknown confounders and was deemed equally important for simple and complex interventions despite frequent comments on the ethical and practical challenges of using RCTs to evaluate population-based interventions. The high scores for criterion 1 (see Table 5.2), therefore, were often accompanied by comments on RCTs being less prone to bias and providing evidence of higher internal validity as opposed to nonrandomised studies, which were perceived to rely on more *"stringent and often non-verifiable"* assumptions. In the majority of cases, however, participants highlighted that the internal validity of a study should not be linked with the design in a strictly linear way and highlighted the importance of additionally considering the nuances related to the study implementation and analysis before making judgement about study quality:

> *"In general, RCTs have a better chance of being able to deliver robust causal evidence, but whether they in fact do, depends largely on issues, such as sampling, allocation, blinding, intervention fidelity and control isolation etc." (Score 6, Criterion 1).*

This argument was commonly used by participants to justify the lower scores on criterion 1 and, conversely, the higher scores on criterion 4. Specifically, a few participants argued that a risk of bias tool should be used to determine the quality of evidence, rather than a two-step process that involves an initial rating based on study design followed by risk of bias assessment. Furthermore, dropping of the initial categorisation of a body of evidence so that an initial rating of "high" is given for a body of evidence consisting of any type of study design (see criterion 4 in Table 5.2), was perceived as a more consistent approach for synthesising and rating a mixed body of evidence comprised of different study designs (e.g., RCTs and NRSs). Many participants, however, expressed reservations with regard to dropping the initial categorisation of evidence. There was a strong conviction among participants that study designs differ in terms of their potential for bias, and, therefore, many argued that initially treating strong and weak designs for making causal claims, such as RCTs and case series as equal would be misleading. A few participants argued that this approach may hinder the use of evidence upgrading criteria in the GRADE approach. Finally, a concern was raised that this approach would only work efficiently when reviewers use a rigorous approach to assess risk of bias in both randomised and nonrandomised studies, such as the Cochrane risk of bias tools. In this view, several participants highlighted that the two-step approach, involving an initial sorting of evidence based on study design and a further fine-tuning of the certainty rating by way of using general risk of bias criteria may be easier to use by less experienced reviewers and may guard against reviewers making too strong claims based on weak evidence:

> *"This [refers to Criterion 4] would be a suitable alternative way of starting off, provided that this is followed by a detailed assessment of the study design-specific risks of bias for each body of evidence. This process is, however, likely to be a lot*

> *less efficient compared to the 'quick-and-dirty' sorting of bodies of evidence according to study design. Different designs bring different quality of evidence; therefore, they should not be treated equally." (Score 2, Criterion 4)*

In general, participants thought positively of the current GRADE approach, which initially categorises a body of evidence into high and low categories based on studies using a randomisation to assess the effects of an intervention (see Criterion 2 in Table 5.2). While many participants argued that certain nonrandomised studies, such as those involving an exogenous intervention (e.g., a regression discontinuity design) may provide evidence of higher internal validity than other epidemiological studies (e.g., a case-control study), and, therefore, they should be initially rated as "moderate" certainty (see Criterion 3 in Table 5.2), a few reservations were raised about this suggestion. Specifically, participants noted the general confusion and inconsistencies in the field with regard to labelling and defining nonrandomised studies, such as existing multiple definitions on what constitutes a "natural experiment" (Craig, Katikireddi, Leyland, & Popham, 2017; Geldsetzer & Fawzi, 2017). In consequence, participants highlighted the potential challenge of distinguishing and classifying those study designs, which could be initially rated as "moderate" certainty:

> *"I think we should make the distinction between randomised and nonrandomised, and not introduce another category [such as non-experimental observational and nonrandomised experimental studies]. Although it might be reasonable that some nonrandomised designs better protect against bias than others, it will be difficult to operationalise." (Score 1, Criterion 3)*

Limitations of included studies (RCTs and NRSs)

Criteria that were rated to be *critically important* in this domain (see Table 5.2) were frequently perceived by participants as fundamental for judging the extent of bias in the body of evidence. Participants often referred to supporting empirical evidence to

justify the need to consider these criteria when rating the certainty of evidence in systematic reviews. In many instances, participants' comments highlighted the applicability of these criteria to all types of reviews regardless the level of complexity:

> *"Empirical evidence demonstrates the potential for bias when allocation is not concealed. This probably applies equally to trials of complex interventions." (Score 9, Criterion 8)*

> *"This would be standard and needs to be considered – I cannot see why this might be of lesser importance for complex interventions..." (Score 8, Criterion 15)*

For the remaining criteria in this domain, which were rated as important, but not critical, participants expressed uncertainties about their relevance and importance in systematic reviews of complex interventions. By way of illustration, a large number of comments were made on the impossibility to blind recipients or providers of complex interventions (criteria 13 and 14 in Table 5.2). In this view, uncertainties were raised regarding the need to downgrade evidence for lack of blinding:

> *"If the comparison is with a business as usual scenario, blinding of providers and recipients are difficult to achieve in many complex interventions, so frankly, I am not sure I think it applicable. Or maybe this is a reason why RCTs should not automatically be rated as high quality evidence." (Score 3, Criterion 13)*

> *"If we talk about complex public health interventions (prevention, health promotion), the most time it will not be possible [to blind]. Does it make sense to use criteria which cannot be reached?" (Score 4, Criterion 14)*

In several instances, participants noted that decisions to downgrade evidence should be made in consideration of the multiple criteria, such as lack of blinding of intervention recipients combined with use of subjective outcome measures. Similarly, participants highlighted that decisions to downgrade evidence for lack of fidelity to intervention implementation (criteria 12 and 23 in Table 5.2) should be contingent upon

considerations, such as when tailoring or flexibility of intervention implementation is a desirable and expected feature and is adequately described in the review:

> *"With many complex interventions one expects a certain degree of tailoring/adaptation across different contexts. As such, this does not provide a reason for down-rating the quality of evidence, but which adaptations/deviations are implemented, and how these were decided on should be clearly reported." (Score 2, Criterion 12).*

> *"Again, not sure this should be used to downgrade evidence, given that complex interventions may vary in terms of the level of permissible tailoring." (Score 3, Criterion 23)*

Inconsistency of effects in the evidence base

Participants frequently observed that reviews of complex interventions are likely to yield highly heterogeneous evidence primarily because of differences in how the interventions are implemented across different contexts and settings. In this view, many participants argued for a need to thoroughly investigate the sources of inconsistency in systematic reviews of complex interventions:

> *"It is not the presence of heterogeneity in reviews of complex interventions that is of importance (of course there is heterogeneity!), it's whether the authors appropriately examine and attempt to minimise the heterogeneity." (Score 7, Criterion 24)*

Quantitative analytical methods, such as sub-group analyses and meta-regressions were generally preferred by participants to less conventional methods for investigating heterogeneity, such as qualitative comparative analysis (QCA; see Criterion 25 in Table 5.2). The latter was perceived as having a potential value in informing judgments on inconsistency; however, participants felt that the use of these "non-quantitative" methods may require further research and guidance:

> *"In my view, QCA (and N.B. QCA is not a purely 'qualitative analysis') and other 'configuring' qualitative research synthesis methods frameworks are useful*

complementary methods for exploring heterogeneity; ideally implemented as an adjunct to 'quantitative analyses exploring heterogeneity'. We probably need further guidance on how the results of such analyses should inform judgements regarding the extent of concern about inconsistency." (Score 6, Criterion 25)

The remaining criteria for rating inconsistency in the body of evidence were scored by the group as important, but not critical (see Table 5.2), and many participants made the case for considering these criteria in tandem and in light of the attempts to explore variation in the evidence base. Participants argued against downgrading of evidence based on information regarding 1 criterion only (this was often referred to as "automatic" downgrading), such as when evidence suggests high levels of I^2 or large variation in point estimates. In fact, many participants highlighted the importance of considering direction of the effect in reviews of complex interventions over the observed variation in point estimates:

"I would expect much variability in point estimates across studies of complex interventions and, as such, this is not a reason for down-rating confidence in the body of evidence. Where the direction of effect varies, this is a reason of concern although, if explainable through moderator/predictor analyses, not necessarily a reason for down-rating." (Score 5, Criterion 27)

Particular doubts were expressed about the usefulness of the statistical test in assessing heterogeneity in reviews of complex interventions:

"This is almost useless regardless of whether the test has a low or high p-value. One expects many sources of non-random variance in the effects of complex interventions, so any adequately powered test should demonstrate such heterogeneity." (Score 1, Criterion 29)

"Avoid this test at all costs - it has low power in most cases and is less important than variation in the direction of effect." (Score 1, Criterion 29)

Indirectness of the evidence base

Most frequently participants referred to the existing evidence demonstrating how extrapolation from a different body of evidence may yield biased results to justify the high scores on the criteria for rating indirectness of a body of evidence:

> *"Extrapolation from one population to another has been shown to be a serious source of bias, for example, interventions found to be beneficial in one population show no effect, or harmful effects in other populations." (Score 9, Criterion 30)*

Participants, however, also often highlighted that the judgements of evidence indirectness depend on the nature of outcomes, the specific intervention, and the degree of differences in the body of evidence compared to the evidence of interest. This was often provided as a reason for lower scores (still in the range of *critically important* and *important but not critical):*

> *"It is important to consider this aspect, but the importance of the length of follow-up really depends on the intervention in question." (Score 5, Criterion 34)*

While one participant made a comment that the criteria of indirectness should not be different in reviews of "complex interventions", a few others highlighted the expected variation in reviews of complex interventions across different settings. In this view, a few participants assessed the indirectness of interventions as of limited importance, while a few others, in contrast, highlighted the critical importance of assessing indirectness of interventions in reviews of complex interventions:

> *"Complex interventions are highly likely to provide indirect evidence only as implementation will likely vary greatly from one setting to the next." (Score 1, Criterion 31)*

> *"This is especially important for complex interventions, where there is more room for diversity in the definition of an intervention as compared with non-complex." (Score 9, Criterion 31)*

Finally, one participant raised the point that a causal chain approach might be a more appropriate way to assess outcomes in reviews of complex interventions. The proposed approach towards rating indirectness in outcomes primarily following the GRADE approach, however, was not perceived to accommodate a causal chain approach:

> *"Outcomes and their appropriate measurements are critical but, unlike for clinical interventions, it may not always be necessary or even feasible to have 'hard health outcomes' (e.g. lung cancer deaths); often intermediate outcomes (e.g. smoking cessation rates) may be equally valid. This raises the important issue of the causal chain leading from an intervention to its short-term and long-term outcomes, and how one might be able to piece together evidence across this causal chain, something that the current version of GRADE does not cater to." (Score 7, Criterion 32)*

Imprecision of the effect estimates

Many of the comments for the imprecision domain highlighted that the width of the confidence intervals is a more important criterion for imprecision assessment as compared to the overlap of the confidence intervals with the line of no effect. As one participant summarised, *"it raises the possibility that the effect could be one of serious harm"* (Score 7, Criterion 35). Participants, however, thought of the overlap of the confidence intervals with the null effect as also useful for assessing the balance of benefits and harms and the precision of the results:

> *"This is very useful information and also links to the earlier point about variations in the direction of effect." (Score 8, Criterion 36)*

Two participants further suggested to combine the assessment of the width of the confidence intervals and the overlap with the line of no effect (criteria 35 and 36 in Table 5.2), as they were perceived to relate to the same process of assessing whether the confidence intervals are large enough to include/exclude *"meaningful effects".* In this view, a few participants, highlighted the challenges of defining the *"clinically meaningful*

ranges" of benefit and harm, as they were deemed to vary widely across fields of practice and outcomes of interest. Finally, one participant suggested to look at the width of the prediction intervals in addition to confidence intervals in reviews estimating random-effects means.

Uncertainties were expressed in the panel regarding the importance of the optimal information size (OIS) criterion. Those who gave high scores on the OIS criterion argued that it provides an objective approach towards defining whether the review is *"sufficiently powered", and* encouraged to make a better use of power analyses in systematic reviews in general, and in reviews of complex interventions more specifically*:*

> *"Systematic reviews of complex interventions tend to have more potential confounders/moderators than a single trial, so should arguably have even greater total sample sizes. Also, power analyses are typically underused in meta-analyses." (Score 7, Criterion 37)*

Others, however, thought of the power of the individual studies, as well as the number of studies included in the review as more important considerations for assessing precision of review results:

> *"Sample size is an important criterion at the study level, not so much at the meta-analysis level. At the meta-analysis level, the N of studies is more important." (Score 2, Criterion 37)*

Further doubts were expressed about the importance of the OIS criterion, as it was perceived to ignore the role of context and allocation of higher level units (such as in population-level interventions); its importance was also argued to depend on the study designs and the number of outcomes considered in the review.

Publication bias

Similar to the criteria in the foregoing domains, participants often referred to the existing supporting empirical evidence to justify their high scores on the criteria for rating publication bias. Conversely, the lack of supporting evidence was stated by participants as a common reason for providing lower scores on the criteria, such as the grey literature and the language of included studies:

> *"For my area (public health), there is strong evidence that a range of databases are needed to find all the relevant studies (Medline only not enough)." (Score 9, Criterion 38)*

> *"See work done by Petrosino and Soydan on developers' involvements' influence on results." (Score 9, Criterion 39)*

> *"Role and impact of grey literature is insufficiently examined." (Score 3, Criterion 41)*

While most of the participants agreed that comprehensive searches in indexed databases, as well as searches in grey literature are important for locating the relevant evidence, many comments were made that these considerations should be used as quality indicators of the systematic review more broadly, rather than as criteria for rating the certainty of evidence. Searches in grey literature were generally perceived to be complementary to searches of indexed literature for locating evidence of complex interventions, especially those implemented outside of health sector. Many comments, therefore, highlighted that the importance of grey literature searches should be contingent upon the field of interest and the review topic. Similarly, the importance of the inclusion of multi-lingual studies (see Criterion 44 in Table 5.2) was perceived to be dependent on the review scope. As one participant explained, *"the more one is taking a global perspective, the more important including non-English manuscripts becomes"*

(Score 6, Criterion 44). Participants, however, frequently discussed that consideration of multi-lingual studies in a systematic review may not always be feasible:

> *"Authors should have made an effort to conduct searches and include studies published in languages other than English, but one also has to be realistic that this is not always feasible." (Score 5, Criterion 44).*

Participants felt unsure regarding the importance of the criteria 42 and 43 in Table 5.2, namely the funnel plot asymmetry and the number of small studies. With regard to the latter, several comments highlighted that the criterion would be *"captured by the optimal information size"* consideration (Score 4, Criterion 43), and that there might be no further need to consider the number of small studies when assessing publication bias. For the funnel plot asymmetry criterion, participants felt that it may provide a good indicator of publication bias; however, the use of funnel plots was often perceived to be hindered in systematic reviews by the lack of sufficient numbers of studies. A few participants also observed that funnel plots may not be highly reliable indicators of publication bias, because of the different possible reasons for the observed asymmetry:

> *"An asymmetrical funnel plot may be due to many possible factors other than publication bias; and the existence of publication bias cannot be safely ruled out, even if the funnel plot is symmetrical. Therefore, the results of visual inspections of funnel plots and/or tests of funnel plot asymmetry are unlikely to be decisive in forming a judgement regarding the extent of concern about publication bias within 'the body of evidence'." (Score 3, Criterion 42)*

Upgrading the initial certainty rating

Participants providing high scores on the criteria for upgrading the initial certainty rating commonly referred to the Bradford Hill's viewpoints for causation to justify their rating:

"Classic and strong Bradford Hill criterion for causality." (Score 7, Criterion 45)

"Going back to Bradford-Hill viewpoints, this is an important aspect." (Score 8, Criterion 48)

Participants generally thought positively of the established criteria for upgrading evidence in the GRADE approach (criteria 45, 46 and 49 in Table 5.2). Several comments, however, highlighted challenges and specific considerations with regard to using them in reviews of complex interventions. For example, a few participants raised the question on how best to define the "dose" in multi-component interventions. Further challenges were raised with regard to the criterion on the effects of plausible residual confounding: many participants observed that this criterion is hard to understand and is rarely (or almost never) applied in reviews of complex interventions. Another participant noted that this criterion should be integrated with risk of bias (limitations of included studies) assessment instead. Likewise, many participants observed that the large magnitude of an effect should not be used as an indicator of certainty of evidence, as the observed large effect size may be because of existing biases in the evidence base. In this view, a few participants highlighted that the large magnitude of an effect may be used as a criterion to upgrade the initial certainty rating only when the evidence is shown to have low risk of bias. Two other participants observed challenges in appropriately interpreting this criterion:

"This can be confusing as when there is large magnitude of an effect we are increasing our certainty about having an effect, not about the actual effect estimate itself." (Score 5, Criterion 49)

Participants almost uniformly agreed that assessment of the coherence of the causal pathways is an important consideration in reviews of complex interventions, which are often described to have long pathways linking the intervention inputs with the

outcomes. Participants noted that this consideration is consistent with the need to develop better methods for integrating complexity and systems perspectives into evidence synthesis, and may help to address "black-box evaluations", as well as identify evidential gaps in intervention pathways. A few reservations, however, were also discussed on how best to operationalise this criterion to avoid overuse and arbitrary decisions, as in most cases, causal pathways were thought to be speculated, rather than evidence-based. Similarly, while many participants agreed that upgrading of evidence may be warranted if the body of evidence across diverse contexts doesn't have major flaws and shows consistent results (see criterion 48 in Table 5.2), several noted that it may be illogical to use the same domain of evidence to downgrade and upgrade evidence (i.e., use "inconsistency" to downgrade and "consistency" to upgrade). Finally, participants seemed to be more sceptical regarding the criterion of analogous evidence: many commented that it may be difficult to operationalise and may need to use a different body of evidence, thereby, introducing speculation into systematic reviewing:

> *"Not sure how you could reliably define similar or analogous interventions for these purposes; it seems like this could introduce a great deal of subjectivity and room for bias (from the reviewer)." (Score 3, Criterion 50)*

One participant observed that analogous evidence can be used to explain the causal mechanisms within the body of evidence, rather than as a criterion to upgrade certainty in the estimate of effect.

Discussion

Main findings

This online expert panel provides helpful and rich information for developing the agenda for the following expert meeting and for the write-up of the GRADE guidance for

complex interventions. No disagreements were identified among participants on any criterion, and all of the criteria but one, were rated as important to consider in reviews of complex interventions. This provides evidential support for considering these criteria in the GRADE guidance for complex interventions. The rich comments from panel participants further build on the qualitative interviews summarised in the previous chapter and help to filter out the important and most persistent issues, including the relatively controversial topics, which should be prioritised at the expert meeting (such as, the initial categorisation of evidence in GRADE and operationalisation of additional domains to upgrade evidence).

It is worth noting that panel participants agreed on the importance of the domains and criteria from the existing GRADE guidelines. This demonstrates that the GRADE approach is a valuable framework for rating the certainty of evidence in reviews of biomedical as well as wider social and public health interventions, and that, as argued in the previous chapters, extending the work of the GRADE Working Group through a further integrated guidance can efficiently address the reported challenges. The only criterion that was rated *of limited importance* related to dropping of the initial categorisation of evidence based on the design of studies included in the body of evidence (see Criterion 4 in Table 5.2). This finding seems inconsistent with the challenges reported in the previous phases of the thesis research, specifically, that the initial categorisation of evidence in GRADE based on study design "unfairly penalises" evidence from rigorous quasi-experimental studies. The free-text comments from participants, however, elucidate that the low scores on this criterion may be due to inappropriate phrasing of the criterion in the Delphi questionnaire, and participants not fully understanding the process of its implementation; specifically, the criterion did not

detail that the initial rating of “high” certainty for any body of evidence is to be followed by a rigorous assessment of evidence for selection and confounding biases, such as by using the ROBINS-I tool. In this view, participants often commented about the need to differentiate between strong and weak study designs as they were perceived to have different levels of susceptibility to bias. It should, however, be noted that the online expert panel does not provide the appropriate structure and method to convey these implementation details (Fink et al., 1991). It is, therefore, important that the issue is further revisited and discussed more in-depth at the face-to-face expert meeting, particularly considering the findings from the qualitative interviews reported in the previous chapter suggesting that the GRADE Working Group is currently discussing to revise the initial categorisation of evidence.

A few criteria currently not described in the GRADE guidelines were rated highly for inclusion in the GRADE guidance for complex interventions, such as those related to upgrading (e.g., coherence of evidence for the causal pathway and consistency across diverse contexts) and the initial categorisation of evidence (e.g., initially rating certain nonrandomised study designs as “moderate” certainty). However, the comments provided by participants, as well as the qualitative interviews from the previous chapter reveal concerns with regard to adequately operationalising these criteria. It is, therefore, necessary to re-examine the importance of these criteria at the face-to-face expert meeting and try to build consensus on how these may best be operationalised in the guidance. Finally, as noted above, a few participants also raised issues regarding the definition of complexity in the new guidance. Different views were expressed in the panel on conceptualising complexity as an attribute of an intervention, that of a review question or following the systems perspective (Shiell, Hawe, & Gold, 2008). The

conceptual issues related to the remit of the GRADE guidance for complex interventions, including how to define complexity in the guidance, therefore, would also be important to discuss and resolve at the expert meeting.

Strengths and limitations

As is typical for expert panels, this study has several strengths and limitations worth discussing (Jones & Hunter, 1995). First, as a further exploratory investigation into the areas of agreement and disagreement on the criteria to consider when rating the certainty of evidence in reviews of complex interventions, this study included a large sample of purposively chosen experts from diverse disciplinary traditions and methodological expertise, including those who did and did not have previous experience of using GRADE. Compared to the sample size of 9 used in traditional expert panels (Fitch et al., 2001), the achieved size and diversity of the sample is a major strength of this panel. It is also worth noting that despite this diversity, the panel held similar views and did not exhibit significant disagreement on any criterion. The engagement with these targeted stakeholder groups is further evidenced by a large number of comments provided in Round Two and the high retention rate of over 60% in Round Three. Through feedback from a more diverse and larger group of stakeholders, the qualitative data from participants' comments further serve to triangulate and validate the findings from the qualitative interviews reported in the previous chapter.

A potential limitation of this study is the relatively high proportion of experts in the panel from Europe and US and with a background in public health. However, public health itself represents a multi-disciplinary field of research, and public health researchers produce disproportionally larger number of systematic reviews. This is also

the case of communities of researchers in Europe and US, which provides a rationale to sample more participants with these demographic characteristics. Nevertheless, it should be noted, that the panel included a balanced number of male and female stakeholders.

Another factor that may potentially limit the representativeness of the findings is the lower recruitment rate at the inclusion stage (33%). While this may possibly be due to the large number of researchers approached, who may have not perceived themselves as sufficiently experienced in systematic reviewing of complex interventions, this suggests the possibility that those entering the online expert panel were in support of GRADE or the practice of evidence rating more broadly. The findings, therefore, may not be representative of wider communities of researchers in the area, who may have different views on the considered criteria.

It should also be noted that the panel yielded a large number of criteria for inclusion in the GRADE guidance for complex interventions. Consistent with the GRADE approach, however, and as also frequently observed by panel participants, these criteria should not be used as items in a "checklist" approach, but rather should be considered as important heuristics to guide the rating process. In this view, participants often reported that some of the criteria may not be as relevant in certain types of reviews of complex interventions, or that judgement should be exercised on how different criteria may interact and inform each other in a specific review. Considering the format of the online expert panel, however, participants were more prone to rate these criteria as important rather than suggesting to exclude them from the guidance. In general, the ability to assess the relative importance of criteria to each other, as well as how they may play out in different scenarios is limited in the Delphi-based panels by sequential presentations of the criteria in the questionnaire format. Furthermore, while participants may provide

free-text comments, the opportunities to ask questions and clarify uncertainties is also limited. For example, at times some participants noted that they didn't fully understand a criterion. It is, therefore, important that the criteria are discussed in more detail at an expert meeting, where participants can have the opportunity to interact face-to-face and ask for clarifications. This will also provide a more appropriate format to make decisions on how the criteria may be presented and operationalised in the guidance and across different examples of reviews complex interventions.

Finally, it is worth mentioning that for the purposes of this thesis work, the comments and discussions from panel participants were coded and thematically analysed by the DPhil candidate alone. To comply with the recommended practices in qualitative methods (Tong, Sainsbury, & Craig, 2007), the codes and analyses should further be checked and validated by an independent researcher when adapting this manuscript for publication.

Conclusions

This online expert panel examined areas of agreement and disagreement about the criteria for inclusion in the GRADE guidance for complex interventions. The rich qualitative feedback from the panel participants were also highly informative in highlighting issues and topics that should be prioritised at the subsequent expert meeting. Overall, results support the initiative for developing a new guidance for complex interventions, which largely builds off the existing GRADE guidelines. Findings from this study, along with results from the systematic review and interviews presented in previous chapters provide substantive evidence for setting the agenda for the expert meeting, where the content of the GRADE guidance for complex interventions is further

discussed and finalised. The process and the key themes of the discussions at the face-to-face expert meeting is reported in the next chapter.

References

Attride-Stirling, J. (2001). Thematic networks: an analytic tool for qualitative research. *Qual Res, I*(3), 385-405.

Barber, C. E., Marshall, D. A., Alvarez, N., Mancini, G. B., Lacaille, D., Keeling, S., . . . Quality Indicator International, P. (2015). Development of Cardiovascular Quality Indicators for Rheumatoid Arthritis: Results from an International Expert Panel Using a Novel Online Process. *J Rheumatol, 42*(9), 1548-1555.

Basger, B. J., Chen, T. F., & Moles, R. J. (2012). Validation of prescribing appropriateness criteria for older Australians using the RAND/UCLA appropriateness method. *BMJ Open, 2*(5).

Claassen, C. A., Pearson, J. L., Khodyakov, D., Satow, P. M., Gebbia, R., Berman, A. L., . . . Insel, T. R. (2014). Reducing the burden of suicide in the U.S.: the aspirational research goals of the National Action Alliance for Suicide Prevention Research Prioritization Task Force. *Am J Prev Med, 47*(3), 309-314.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, P. (2008). Developing and evaluating complex interventions: new guidance. Retrieved 8 Feb, 2018 from https://www.mrc.ac.uk/documents/pdf/developing-and-evaluating-complex-interventions/

Craig, P., Katikireddi, S. V., Leyland, A., & Popham, F. (2017). Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annu Rev Public Health, 38*, 20.1-20.18.

Fink, A., Kosecoff, J., B., Chassin, M., R., & Brook, R., H. (1991). *Consensus methods: Characteristics and guidelines for use*. Santa Monica, CA.

Fitch, K., Bernstein, S., J., Aguilar, M., D., Burnard, B., LaCalle, J., R., Lazaro, P., . . . Kahan, J., P. (2001). *RAND/UCLA Appropriateness Method User's Manual.* Santa Monica, CA: RAND Corporation. Contact No.: MR-1269-DG-XII/RE.

Geldsetzer, P., & Fawzi, W. (2017). Quasi-experimental study designs seris - paper 2: complementary approaches to advancing global health knowledge. *J Clin Epidemiol, 89*, 12-16.

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. (2017). Retrieved October 18, 2017 from http://gradeworkinggroup.org/

Guise, J. M., Butler, M. E., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017). AHRQ series on complex intervention systematic reviews-paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol, 90*,43-50.

Guise, J. M., Chang, C., Butler, M., Viswanathan, M., & Tugwell, P. (2017). AHRQ series on complex intervention systematic reviews-paper 1: an introduction to a series of articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol*, *90*,6-10.

Guise, J. M., Chang, C., Viswanathan, M., Glick, S., Treadwell, J., Umscheid, C. A., . . . Trikalinos, T. (2014). Agency for Healthcare Research and Quality Evidence-based Practice Center methods for systematically reviewing complex multicomponent health care interventions. *J Clin Epidemiol, 67*(11), 1181-1191.

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ, 311*(7001), 376-380.

Kelly, M. P., Noyes, J., Kane, R. L., Chang, C., Uhl, S., Robinson, K. A., . . . Guise, J. M. (2017). AHRQ series on complex intervention systematic reviews-paper 2: defining complexity, formulating scope, and questions. *J Clin Epidemiol*, *90*,11-18.

Khodyakov, D., Grant, S., Barber, C. E., Marshall, D. A., Esdaile, J. M., & Lacaille, D. (2017). Acceptability of an online modified Delphi panel approach for developing health services performance measures: results from 3 panels on arthritis research. *J Eval Clin Pract, 23*(2), 354-360.

Khodyakov, D., Hempel, S., Rubenstein, L., Shekelle, P., Foy, R., Salem-Schatz, S., . . . Dalal, S. (2011). Conducting online expert panels: a feasibility and experimental replicability study. *BMC Med Res Methodol, 11*, 174.

Khodyakov, D., Stockdale, S. E., Smith, N., Booth, M., Altman, L., & Rubenstein, L. V. (2017). Patient engagement in the process of planning and designing outpatient care improvements at the Veterans Administration Health-care System: findings from an online expert panel. *Health Expect, 20*(1), 130-145.

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

Lorenc, T., Felix, L., Petticrew, M., Melendez-Torres, G. J., Thomas, J., Thomas, S., . . . Richardson, M. (2016). Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Syst Rev, 5*(1), 192.

Marshall, M. N. (1996). Sampling for qualitative research. *Fam Pract, 13*(6), 522-525.

Miles, B. M., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publication ltd.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2).

Montgomery, P., Grant, S., Hopewell, S., Macdonald, G., Moher, D., Michie, S., & Mayo-Wilson, E. (2013). Protocol for CONSORT-SPI: an extension for social and psychological interventions. *Implement Sci, 8*, 99.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

Petticrew, M., Anderson, L., Elder, R., Grimshaw, J., Hopkins, D., Hahn, R., . . . Welch, V. (2013). Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol, 66*(11), 1209-1214.

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P., Mayhew, A., Pantoja, T., . . . Sowden, A. (2013). Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol, 66*(11), 1230-1243.

Petticrew, M., Shemilt, I., Lorenc, T., Marteau, T. M., Melendez-Torres, G. J., O'Mara-Eves, A., . . . Thomas, J. (2017). Alcohol advertising and public health: systems perspectives versus narrow perspectives. *J Epidemiol Community Health, 71*(3), 308-312.

Pfadenhauer, L. M., Gerhardus, A., Mozygemba, K., Lysdahl, K. B., Booth, A., Hofmann, B., . . . Rehfuess, E. (2017). Making sense of complexity in context and implementation: The Context and Implementation of Complex Interventions (CICI) framework. *Implement Sci, 12*(1), 21.

Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers. *Annu Rev Public Health, 35*, 29-45.

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rubenstein, L., Khodyakov, D., Hempel, S., Danz, M., Salem-Schatz, S., Foy, R., . . . Shekelle, P. (2014). How can we recognize continuous quality improvement? *Int J Qual Health Care, 26*(1), 6-15.

Shiell, A., Hawe, P., & Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *BMJ, 336*(7656), 1281-1283.

Squires, J. E., Valentine, J. C., & Grimshaw, J. M. (2013). Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol, 66*(11), 1215-1222.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care, 19*(6), 349-357.

Viswanathan, M., McPheeters, M. L., Murad, M. H., Butler, M. E., Devine, E. E. B., Dyson, M. P., . . . Morton, S. C. (2017). AHRQ series on complex intervention systematic reviews-paper 4: selecting analytic approaches. *J Clin Epidemiol*, *90*,28-36.

Chapter 6. An expert meeting

Finalising the content of the GRADE guidance for complex interventions

A paper adaption of this chapter has been submitted to in *PLOS ONE*

Chapter overview

This chapter reports on the discussions of the face-to-face expert meeting on finalising the content of the GRADE guidance for complex interventions. A group of 28 stakeholders including guideline developers, systematic reviewers, methodologists of complex interventions, journal editors, and a research funder met in Oxford in May 2017 for a three-day meeting. The meeting was held to discuss and make decisions on the content and dissemination of the GRADE guidance for complex interventions in light of the evidence from the previous phases of the thesis research.

Five overarching themes emerged from the meeting discussions, which describe decisions regarding the content and dissemination of the new guidance. These include (1) suggestions on how to conceptualise complexity, (2) modifications to the existing GRADE guidance that are perceived as considerable changes to the existing framework (referred to as "thinking outside of the GRADE box"), (3) modifications to the existing GRADE guidance requiring further detail and elaboration through worked examples (referred to as "fine-tuning"), (4) considerations to enhance the usability and uptake of the new guidance, and (5) areas for future research. The new guidance should facilitate explicit decisions on rating the certainty of evidence in reviews of complex interventions. These ratings should adequately communicate the level of certainty associated with an estimate of effect of a complex intervention to decision-makes.

Introduction

A face-to-face consensus meeting is considered the most vital phase in the recommended techniques for developing health research reporting guidelines (Moher, Schulz, Simera, & Altman, 2010). The aim of this meeting is to bring together a group of key stakeholders in the area over the course of a few days to discuss and reach agreement regarding the content of a new guidance. A number of advantages have been discussed for a face-to-face consensus development meeting: a wide range of knowledge and experience is brought to bear; the interaction between participants stimulates consideration of a wide range of options and debate that challenges received ideas and stimulates in-depth discussions; idiosyncrasies are filtered out; and most importantly, participants are given an opportunity to share reasons for held opinions and seek further clarification, which are hard to achieve without in person interactions (Jones & Hunter, 1995; Moher et al., 2010; Murphy et al., 1998). In this view, a systematic review of literature and consultation with different stakeholders, including a Delphi-based process, are seen as important *pre-meeting* activities to inform the agenda and help prioritise the topics of the *face-to-face meeting*.

According to the best practices for developing research reporting guidelines, a face-to-face consensus meeting should preferably last between two to three days to give an opportunity for participants to thoroughly engage with the discussion topics and share their views in a relaxed environment, as well as revisit key points and decisions on different days (Moher et al., 2010). The discussions should be grounded in evidence. Furthermore, the most detailed and structured discussions should focus on information content and not precise wording of the guidance. While it is expected that the views of

the meeting participants will eventually converge to a consensus, it may occasionally be necessary to vote on some issues. This meeting is also viewed as a good opportunity for participants to discuss a plan for producing guidance documents and options for knowledge translation (Moher et al., 2010). Following this strategy, a three-day consensus development meeting was held to decide on the content and dissemination of the GRADE guidance for complex interventions allowing meeting participants to discuss the results from the previous phases of the thesis research in a greater detail.

This thesis chapter reports on the process and key themes of the discussions at the face-to-face consensus meeting. By providing a detailed and transparent account of the meeting proceedings it aims to provide a foundation for the content of the new GRADE guidance for complex interventions; in the meantime, it can also serve as an important output of a collective intellectual process in itself, to inform future methodological work on systematic reviewing of complex interventions more broadly.

Methods

Recruitment and setting

Meeting participants were purposively selected from the online-modified Delphi process and through consultation with the co-investigators of the project on developing *GRADE Guidance for Complex Interventions*. Participants were drawn to represent key stakeholder groups across different disciplines, including methodologists, systematic reviewers, guideline developers, journal editors and funders in public health, psychology, international development, education, social care and criminology (Moher et al., 2010). At the outset of the recruitment process a decision was made to have a group of 20 to 30 participants to optimise the diversity of viewpoints with a balanced number of

representatives from different stakeholder groups and disciplines. In the meantime, it was meant to keep the number of participants relatively small to maximise the efficiency of discussions and the likelihood of achieving consensus (Jones & Hunter, 1995).

Once the dates of the meeting were set, an initial list of invitees was compiled by the DPhil candidate and project co-investigators, and individualised invitation emails were sent to this list in January and February 2017 with up to two email reminders. Invitation emails outlined the meeting aims, procedures and setting and were predominantly sent by the DPhil candidate; however, if a project co-investigator knew the invitee, they sent the letter personally to encourage participation. The DPhil candidate kept log of these emails and reported back to the project co-investigators regarding the responses received from the invitees. The project co-investigators and the DPhil candidate met regularly from January to May 2017 to review the list and nominate new participants in case invitees could not attend the meeting or refused after an initial positive response.

The meeting took place from 24 to 26 of May 2017 at the Oxford Spires Hotel in Oxford, UK. This provided a relatively quiet location in a short distance from the city centre. All meeting discussions were held in one group and in the same large room. Tables were arranged in a U-shape around a projector, which was used for presentations. In this set-up participants could see and hear each other speaking at any time. All participants, for the exception of those who lived in Oxford, stayed on site. Travel, accommodation and meals for each participant were provided or reimbursed by the project grant (ES/N012267/1). Ethical approval for using the meeting discussion data for this thesis work was obtained from the Departmental Research Ethics Committee at the University of Oxford (Ref: SPI_C1A_16_009).

Procedures

The meeting agenda was first drafted by the DPhil candidate and later revised in light of the feedback from project co-investigators (see Appendix 13). The agenda largely followed a structure based on GRADE to ensure that all essential constructs and domains of evidence could be discussed in separate sessions.

Pre-meeting reading materials were also developed in advance and sent to participants via a Dropbox link one week before the meeting. These materials were organised in different Dropbox folders including a folder with practical information, such as the meeting agenda, the biographies of all participants attending the meeting, and further information on who to contact in case of emergency and how to travel around in Oxford. Another folder included recommended readings for the meeting, such as publications and reports from the previous phases of the research (these materials are covered in Chapters 1, 3, 4 and 5), and an information sheet outlining the details of the meeting for participants to read before they were asked to consent to audio-recording of the meeting discussions. Finally, a separate folder included supplementary readings on the GRADE approach more broadly, such as the official GRADE guidelines published in the *Journal of Clinical Epidemiology*. All of these materials, for the exception of supplementary readings, were also provided to participants in a printed pack at the meeting.

Closer to the date of the meeting, Skype calls were conducted by the DPhil candidate and two other project co-investigators with each meeting participant separately. These calls aimed to clarify logistical issues, check whether participants had any concerns or questions, discuss their expectations from the meeting and identify important items to add to the agenda.

Meeting agenda

Following the established format of conducting face-to-face consensus meetings for development of research reporting guidelines, all discussions of the meeting were grounded in empirical evidence (Boutron et al., 2008; Moher et al., 2010). Specifically, the three-day meeting was divided into sessions, and discussions in each session began by a review of relevant literature and results from the previous project phases. The DPhil candidate and project co-investigators took turns as moderators of separate sessions to first, present the literature review and relevant evidence, and then, lead discussions that followed. While moderators aimed to adhere to the agenda, flexibility was allowed in session time and length to accommodate the needs of the group. The decision to vote for any considerations related to the content of the new guidance was also left to the discretion of the group. Each day of the meeting began with an overview of the objectives for that day and discussion of key issues from the previous day. A Research Assistant (RA) was hired for the duration of the meeting to take minutes of the discussions and assist in tasks for smooth running of the sessions (e.g., setting the projector screen and audio-recorders).

Day 1 agenda was broadly designed to welcome participants and provoke a free-flowing conversation around key concepts and perspectives regarding the new guidance. The first few presentations of the day aimed to introduce participants, provide an overview of the meeting objectives and familiarise participants with the activities of the GRADE Working Group and procedures for producing official GRADE guidance. Further discussions of the day were aimed to revolve around the main principles of the GRADE approach, the project's conceptualisation of complex interventions and results from the previous project phases. While participants could raise questions and engage in

discussions at any time, the agenda had a structured Discussion at the end of the day to have participants reflect on the specific features of complexity, which challenge the application of the GRADE approach. This discussion also aimed to encourage participants to express their general views on the needed modifications to the existing GRADE guidance. By the end of Day 1, participants were intended to have a shared understanding of the main concepts and the scope of the new guidance, as well as clarity on the expectations from the rest of the meeting and their specific contribution.

Day 2 agenda was designed to enable more in-depth discussions and resolutions on the content of the GRADE guidance for complex interventions. The review of the aims for the day was followed on the agenda by a discussion of a recent conceptual paper published by the GRADE Working Group regarding different approaches to conceptualising the GRADE construct of "certainty of evidence" (Hultcrantz et al., 2017). It was found important by project co-investigators to have a separate session on this paper as it provides important updates to the existing GRADE guidance and, therefore, was deemed highly relevant in relation to this meeting. An open discussion session then followed for participants to further share their views regarding the content of the new guidance.

For the remaining sessions of the day, structured discussions were planned for each domain of the GRADE approach, such as risk of bias, inconsistency, indirectness, imprecision and publication bias as they relate to complex interventions. Participants were intended to discuss which criteria and considerations should be used to operationalise these domains in the GRADE guidance for complex interventions and how these were different from what is already described in the existing GRADE guidance. The last discussion of the day was meant for participants to agree on any additional domains

of evidence, which should be included in the new guidance, but which are currently not described in the existing GRADE guidance. As outlined above, the specific items for these discussions were informed by the results from the previous phases of the project, including the evidence from the systematic review, interviews with key stakeholders and the online-modified Delphi process. After the meeting, the minutes of the discussions were to be reviewed by the DPhil candidate and project co-investigators to highlight the key issues that required revisiting on Day 3.

Day 3 agenda included two lengthy sessions, namely, *Finalising the GRADE Guidance for Complex Interventions* and *Write-up and Implementation*. After reviewing the aims of the day and summarising the key issues from Day 2, a structured discussion on the agenda aimed at finalising the content for the new guidance allowing for participants to revisit controversial points and address remaining questions. The final session of the meeting included discussions on how to proceed with drafting and piloting of the guidance, as well as planning for the strategy to best implement and disseminate it. Stakeholders from key organisations present at the meeting, including the World Health Organization (WHO), the Cochrane and Campbell Collaborations and the GRADE Working Group were to be specifically asked for their views on how to ensure uptake of the new guidance by these organisations. After this, the meeting was planned to close.

Data analysis

All meeting discussions were audio-recorded and transcribed verbatim by a project RA. This transcription was further checked by the DPhil candidate by way of listening to the recordings and making corresponding revisions to the initial transcript. Discussions were analysed using a thematic approach (see below) (Attride-Stirling, 2001).

As the meeting predominantly involved conceptual discussions around methodological topics in evidence synthesis, such as how to conceptualise the construct of "certainty of evidence" in reviews of complex interventions, a thematic analytic approach was deemed suitable to identify the key themes underpinning the decisions and recommendations made at the meeting. In the meantime, in order to describe the process wherein different arguments unravelled throughout the course of the meeting, the discussion minutes are additionally summarised and reported below. This serves to provide further transparency and context elucidating how and where in the chronological narrative of the meeting the key themes were developed, maintained and/or discarded. It is worth noting, that the reported discussion minutes have already been circulated among the meeting participants and incorporate their feedback and suggestions.

Thematic analysis offers a flexible method for identifying and reporting patterns (themes) within data, which can be used across a range of epistemologies and research questions, including studies with pre-specified topics (Boyatzis, 1998). It differs from other analytic techniques that seek to describe patterns in qualitative data, such as the interpretative phenomenological analysis (IPA) and grounded theory. While both of the latter seek patterns in data, they subscribe to specific epistemological positions and are theoretically bounded. For example, IPA is attached to a phenomenological epistemology, which gives primacy to experiences and aims to understand phenomena through people's experiences (Smith & Osborn, 2003). Although there are different perspectives to viewing thematic analysis as a method in its own right (Braun & Clarke, 2006), or simply as a tool that can be used across different methods (Boyatzis, 1998; Ryan & Bernard, 2000), thematic analysis is not wedded to any epistemic paradigm. As the study reported in this chapter primarily aims to describe how selected participants

justified their arguments and came to decisions regarding the content of the GRADE guidance for complex interventions, the thematic analytic approach following an essentialist/realist approach was deemed appropriate. Such an approach assumes a straightforward and largely unidirectional relationship between meaning and experience and language (Aronson, 1994). Consequently, all presented themes are described and organised *from what participants (experts in the field) have said at the meeting* to show patterns in the semantic content. This is different from a constructionist perspective, which seeks to theorise the important contexts and structural conditions that enable individual meanings and accounts (Burr, 1995). It is, however, worth noting, that the analysis and interpretation of the meeting discussions has been largely driven by the DPhil candidate's theoretical and analytic interest in the area in light of her engagement in the research phases leading up to this meeting (see Chapters 1, 3, 4 and 5). She was a participant at the meeting herself and was actively involved in designing the meeting agenda and prioritising discussion topics in different sessions.

The method of thematic analysis described by Attride-Stirling (2001) was employed in this chapter. This method supports thematic analyses with thematic networks defined as *"web-like illustrations that summarise the main themes constituting a piece of text"* (Attride-Stirling, p. 386). It is argued that presentation of thematic analyses through thematic networks aids the organisation of the analysis, enables methodological systematisation of textual data, and allows a sensitive and rich exploration of a text's overt structures and underlying patterns. Thematic network analysis is developed starting from the *Basic Themes*, which are considered the lowest-order premises in the text. According to the underlying story, these basic themes are further categorised into *Organising Themes* to summarise more abstract principles.

Lastly, organising themes are brought together as *Global Themes* to illustrate a single conclusion or a super-ordinate theme. It should be noted that this structure of themes in the thematic network analysis has been developed based on the principles of argumentation theory (i.e., progression from *data* through a *warrant* to a *claim*) (Toulmin, 1958). The final output is the web-like thematic map depicting the salient themes at each of the three levels and illustrating relationships between them.

Thematic network analysis of the meeting discussions was conducted following the 6 steps outlined by Attride-Stirling (2001):

Step 1. Coding the material: the DPhil candidate read and re-read the transcription, summarised the meeting minutes and marked initial ideas for coding. A mixed approach of deductive and inductive coding was used. An initial list of codes was developed based on the structure of the meeting agenda and the evidence from the previous phases of the research. This structure aimed to ensure that the analysis progressed in line with the overarching objective of the meeting itself, that is, to make decisions on the content and dissemination of the GRADE guidance for complex interventions. This initial list was further updated in light of newly developed codes (see Appendix 14 for the full list of codes). Structural coding (also referred to as "utilitarian coding" to highlight its categorisation function) was first used to dissect the transcript into manageable pieces (Saldana, 2013). For this, large segments of data related to a specific topic (e.g., discussions across different sessions of the meeting related to conceptualising complexity for the new guidance) were labelled with a corresponding content-based phrase and collated together for more detailed coding. Data were then coded descriptively, paragraph by paragraph (using both

descriptive and in vivo codes), to portray the dynamics of interactions among participants, meanwhile taking notice of time as discussions progressed through different sessions and days of the meeting (Miles & Huberman, 1994; Reed & Payton, 1997). Coding and analyses were conducted using NVivo 11.

Step 2. Identifying themes: After the first cycle coding, the identified codes were collated, reviewed and sorted into corresponding themes. This was done by re-reading the text segments in each code and extracting salient themes.

Step 3. Constructing the network: The themes identified in Step 2 (pre-specified or newly emerged) were arranged according to the structure described above. Basic themes were selected and categorised into organising themes based on similarities in content.

Step 4: Describe and explore the thematic network: Transcripts were re-read to validate the thematic map and check whether any important themes have been missed. This also included examining and discussing the initial thematic map by two investigators, namely the DPhil candidate and the research assistant, who assisted with the transcription of data. The initial versions of the map were additionally reviewed by the candidate's supervisor and revised in accordance with the feedback received.

Step 5. Summarise the thematic network: Organising themes along with their basic themes were summarised narratively in the text of the chapter using illustrative quotations.

Step 6. Interpret patterns: A written summary was generated discussing and contextualising the thematic network in light of the thesis aims.

The study and its reporting was informed by procedural guidance on quality and trustworthiness in qualitative research (Lincoln & Guba, 2000; Tong, Sainsbury, & Craig, 2007).

Results

Participants

Overall, 28 people attended the meeting comprising a gender-balanced group with 16 female and 12 male participants (see Figure 6.1 for the participant flow diagram). The group included six co-investigators of the project and the DPhil candidate herself, two methodologists from the GRADE Guidance Group (details about this group are presented below), four editors from the Cochrane and Campbell Collaborations, seven guideline developers across key stakeholder organisations, such as World Health Organization (WHO), Centers for Disease Control and Prevention (CDC) and Agency for Healthcare Research and Quality (AHRQ), further six methodologists and systematic reviewers with an interest in complex interventions and one representative from a research funding organisation. Participants predominantly came from high-income regions, UK (n=15), USA (n=6), Europe (n=5), Canada (n=1); 1 participant was from Asia. Participants represented a range of disciplines, including clinical epidemiology, public and environmental health, education, psychology, criminology, social work and social care. Table 6.1 provides further information on participants' relevant affiliations and professional roles.

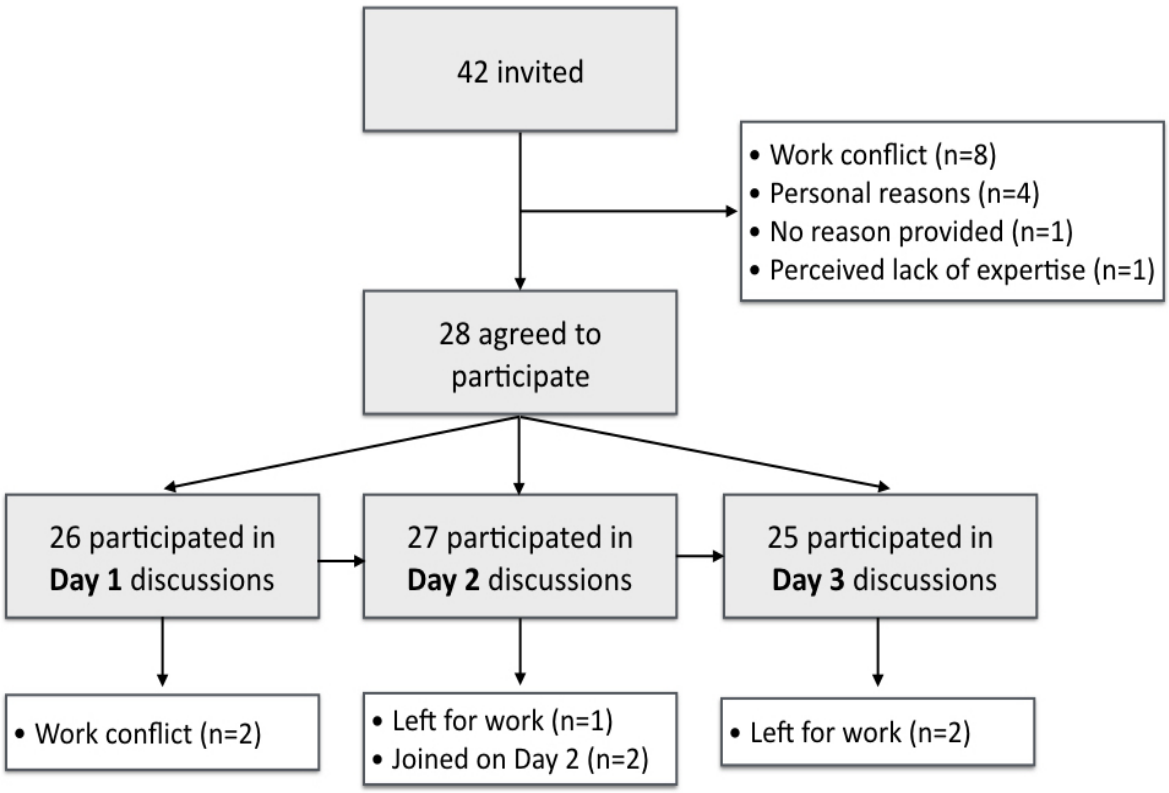


Figure 6.1. Meeting participant flow diagram

Meeting minutes

The meeting sessions generally evolved as per the agenda. A few minor logistical changes were made to the agenda, specifically, discussions on the domains for upgrading evidence in GRADE and additional considerations for the new guidance originally planned to happen in two separate sessions were combined in one session at the end of Day 2; similarly, discussions on the write-up, piloting and dissemination activities were merged in one session at the end of Day 3. Given the relatedness of the topics, the group felt that combining these sessions would provide an opportunity for more in-depth discussions. The rest of the sessions were held as described above.

Table 6.1. Meeting participants' characteristics

Participant ID	Region	Relevant Affiliation*	Professional Role
Participant 1	UK	SURE	Systematic Reviewer & Methodologist
Participant 2	Australia	Campbell Collaboration	Journal Editor & Systematic Reviewer
Participant 3	UK	GRADE Working Group	Systematic Reviewer
Participant 4	UK	NICE; GRADE Working Group	Guideline Developer
Participant 5	UK	Campbell Collaboration; 3ie	Systematic Reviewer & Methodologist
Participant 6	US	Institute for Public Health and Medicine	Trialist in Behavioural Medicine
Participant 7	US	AHRQ	Systematic Reviewer & Methodologist
Participant 8	UK	DECIPHer	Systematic Reviewer & Methodologist
Participant 9	Asia	GRADE Guidance Group	Guideline Developer & Methodologist
Participant 10	Europe	Cochrane Collaboration; WHO	Guideline Developer & Methodologist
Participant 11	UK	Cochrane Collaboration	Journal Editor & Systematic Reviewer
Participant 12	UK	DECIPHer	Systematic Reviewer & Methodologist
Participant 13	UK	Cochrane Collaboration	Journal Editor & Systematic Reviewer
Participant 14	US	AHRQ	Systematic Reviewer & Methodologist
Participant 15	UK	DFID	Research Funder
Participant 16	UK	LSHTM	Guideline Developer
Participant 17	UK	What Works Centre for Wellbeing (WWW)	Methodologist
Participant 18	UK	Cochrane Editor	Systematic Reviewer & Methodologist
Participant 19	UK	MRC/CSO Social and Public Health Sciences Unit	Systematic Reviewer & Methodologist
Participant 20	US	CDC	Guideline Developer & Methodologist
Participant 21	Europe	GRADE Guidance Group	Guideline Developer & Methodologist
Participant 22	US	Campbell Collaboration	Journal Editor & Systematic Reviewer
Participant 23	US	Campbell Editor	Systematic Reviewer & Methodologist
Participant 24	Europe	WHO; GRADE Working Group	Guideline Developer
Participant 25	Europe	GRADE Working Group	Guideline Developer & Methodologist
Participant 26	UK	Cochrane Collaboration	Journal Editor & Systematic Reviewer
Participant 27	UK	MRC/CSO Social and Public Health Sciences Unit; GRADE Working Group	Systematic Reviewer & Methodologist
Participant 28	Canada	GRADE Working Group	Systematic Reviewer & Methodologist

*Because of identifiability concerns, this includes only partial information on the participants' affiliations. Those affiliations are listed, which were deemed most relevant for the purposes of this meeting.

Notes: CDC: AHRQ: Agency for Healthcare Research & Quality; Centers for Disease Control and Prevention; DECIPHer: The Centre for Development and Evaluation of Complex Interventions; DFID: Department for International Development; ECDC: European Centre for Prevention and Disease Control; GRADE: Grading of Recommendations Assessment, Development and Evaluation; LSHTM: London School of Hygiene & Tropical Medicine; NICE: National Institute for Health and Care Excellence; SURE: Specialist Unit for Review Evidence; WHO: World Health Organization

Day 1 Minutes

Session 1.1. Welcome & introductions (P18)

The meeting began at 3pm on May 24 by the session moderator outlining the broad aims of the three-day consensus meeting followed by introduction of all experts attending the meeting. The remit of the project was briefly outlined to the participants to stimulate focused thinking, and participants were asked whether they had any questions for clarification at this point (no questions were raised). Several housekeeping issues were also explained to the participants, including the plans for the group dinner that evening and a request to read the information sheets and sign the consent forms to audio record the meeting discussions.

Session 1.2. Overview of objectives (P23)

The session moderator first presented the specific objectives of the project on developing *GRADE Guidance for Complex Interventions* and then outlined the goals of the meeting itself by highlighting three overarching questions that the meeting discussions were expected to address. Those included discussions and decisions on any terminological changes that might be needed for the new GRADE guidance for complex interventions, as well as whether the new guidance needs to clarify aspects of the existing GRADE guidance by providing further detailed examples for complex interventions, and whether there is a need to make conceptual and structural changes to the GRADE approach, such as revisions of GRADE definitions and additions of new domains of evidence. The moderator referred to the recently published GRADE equity guidelines (Welch et al., 2017), as providing a structure that may also inform the format of the GRADE guidance for complex interventions.

The moderator then presented on the strategy adopted by the project team to develop the guidance largely following the recommended techniques for developing research reporting guidelines (Moher et al., 2010), including the CONSORT and the PRISMA guidelines. The moderator highlighted the importance of the previous project phases in setting the meeting agenda:

> *P23: We have adopted a sort of funnelling approach for developing the GRADE guidance for complex interventions [see Figure 6.2 for the diagram presented at the meeting] … Starting off with a universe of considerations that we have identified in previous systems for rating the quality of a body of evidence to put on the table, feed those through to in-depth interviews with review authors and methodologists in this area, and then the Delphi process … which I think most, if not all of you have participated in … Thank you for that. And it is from this process that we have tried to prioritise the discussion topics for this meeting. (Systematic reviewer & methodologist)*

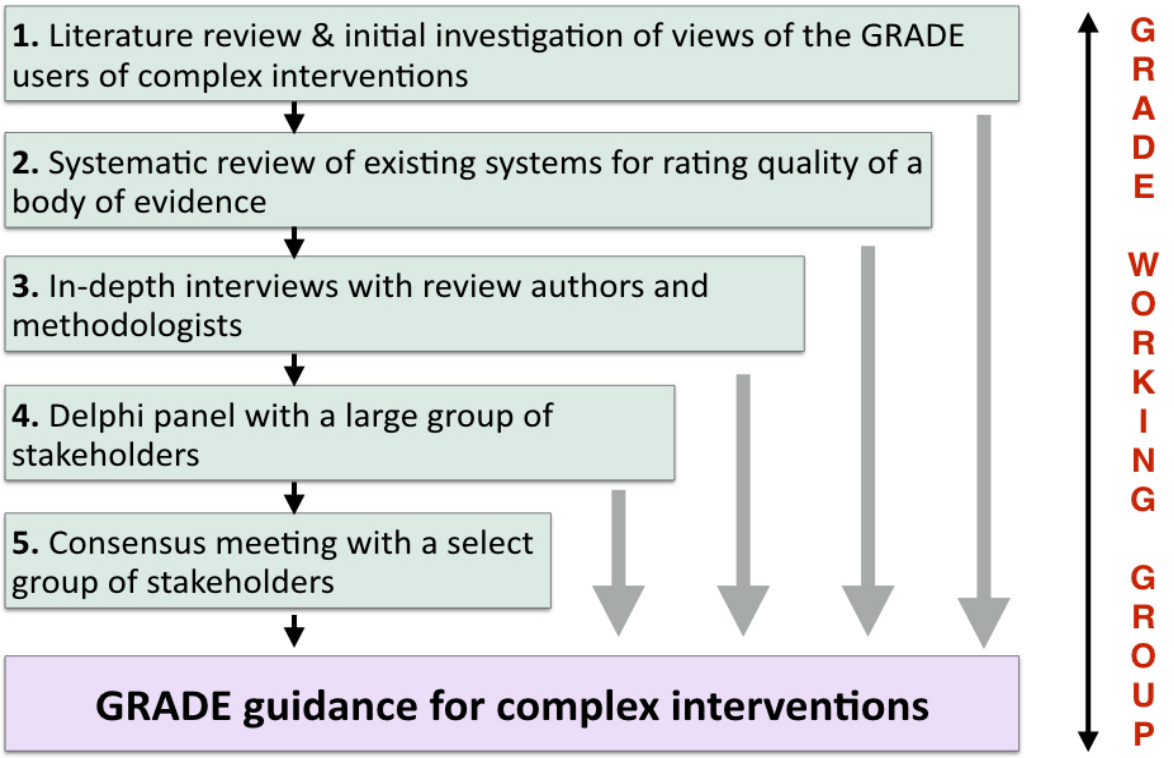


Figure 6.2. Phases for developing the GRADE guidance for complex interventions

The moderator also noted that the project team is working officially with the GRADE Working Group and once the write-up of the guidance is complete, it needs to be further approved by the GRADE Working Group (this process of official approval of the guidance was discussed in detail in session 1.3 summarised below). The agenda for Day 1 was then outlined to the participants.

Session 1.3. Process for securing GRADE approval of the guidance (P9)

Broadly, this session informed the participants about the activities of the GRADE Working Group (GWG), and, more specifically, how the work produced by this project could be approved as an official GRADE guidance. The moderator – a member of GWG and the GRADE Guidance Group – provided a background information about the activities of GWG. The moderator also notified that these activities are steered by the GRADE Guidance Group (GGG), which is comprised of 10 members serving as the "executive body" on a rotating basis.

The process of producing official GRADE guidance was then discussed. The moderator explained that GWG operates in smaller project sub-groups, which aim to advance the GRADE methodology for different subject areas. These sub-groups are formed voluntarily by interested members of GWG, which work together to submit a *Terms of Reference* specifying the objectives, timeline, plan and deliverables of their project to GGG. The sub-group members then work together to develop a paper, and when ready, the sub-group leader circulates it to the wider GWG for review and feedback. Although this process of review and feedback is often perceived as long and tedious, the moderator noted that it ensures that views of all interested members of GWG are addressed.

The sub-group is then expected to present the paper for further feedback and discussions in one of the wider GWG meetings, which are usually held twice per year. If serious concerns or controversies are raised at the meeting, the sub-group takes the paper back and refines it for the next GWG meeting. For the final approval, the sub-group needs to submit the paper to GGG, and the paper becomes GRADE guidance only when it is approved by GGG as such. To assure the transparency of the decisions, GWG follows several rules at the meetings. First, the sub-group is expected to submit a paper to GWG at least one week before the Group meeting; at the GWG meeting, members are invited to vote anonymously for ("yes) or against ("no") approving the paper produced by the sub-group and are also given an option to "abstain". For the paper to be approved at this meeting, 80% or more of those voting need to vote "yes" for approval (those who abstain are not counted in the denominator, as they are suspected to have limited familiarity with the topic). The moderator also noted that most of the papers, which are approved by GWG and GGG, are published in the Journal of Clinical Epidemiology (JCE) and are labelled as GRADE guidance, which makes them more visible and influential. In cases, when papers are not approved by GWG, the authors of the paper may still proceed with publishing their methodological work, however, it will not qualify as GRADE guidance.

Discussions of this session revolved around the peer review process of GWG publications in JCE. Several participants highlighted that as opposed to other manuscripts, the journal review process for these papers can be fast-tracked and that submitted manuscripts are unlikely to be rejected. To justify this, members from GWG present at the meeting argued that all the procedures of review and feedback within the Group provide sufficient confidence to JCE reviewers to make the review process within

the journal much faster. A few concerns were further raised among participants on whether the 80% threshold for the votes in GWG is too restrictive. To this, the moderator responded that the voting approach has been tested at the last three meetings of GWG and no concerns have been raised so far. The moderator, however, also noted that the Group would be flexible to change the 80% threshold in case if any issues are encountered at later meetings.

Session 1.4. Background: rating the certainty of evidence in GRADE (P3)

This session aimed to bring about consensus regarding the key definitions and concepts by providing a general summary of the GRADE approach and the domains for rating the certainty of evidence as currently described in the official GRADE guidance (Guyatt et al., 2011). This information is discussed in detail elsewhere in this thesis (see Chapter 1). A few participants made comments in this session regarding the important revisions to the GRADE approach that GWG has recently approved, such as dropping the initial categorisation of evidence based on study design; it was, however, agreed that these updates and their relevance for the GRADE guidance for complex interventions will be discussed in greater detail in dedicated sessions on Day 2. Further discussions of this session aimed to clarify the rationale for the GRADE criteria to upgrade evidence, as well as the Evidence to Decision (EtD) frameworks, which have been recently developed by the members of GWG (Alonso-Coello et al., 2016). It was explained that the EtD frameworks specify criteria that guideline developers should consider when making recommendations for practice, including health effects of an intervention and the GRADE ratings of the certainty in the estimates of those effects, as well as non-health effects of an intervention. The latter are referred to as contextual evidence in the EtD frameworks

and may include evidence on intervention feasibility, acceptability, equity and cost-effectiveness. The frameworks aim to encourage guideline developers to report transparently on the judgments regarding these considerations, including the rating of the certainty of evidence.

Session 1.5. Defining complex interventions in complex systems (P10)

This session included discussions on different perspectives to conceptualising complexity in intervention research, such as *"complex interventions"* and *"complex systems"* (see thematic analysis below). Participants shared their knowledge regarding recent and ongoing projects aiming to develop methods for addressing complexity in systematic reviews, including the iCAT_SR tool and the PRISMA extension for complex interventions (Guise et al., 2017; Lewin et al., 2017), as well as discussed how these projects could contribute to operationalising complexity in the new GRADE guidance. Questions on the added value of incorporating a complexity perspective in systematic reviews were also raised and addressed at the meeting. While different participants expressed their views on the most suitable perspective to conceptualising complexity for the new GRADE guidance, the group agreed to revisit the issue as the discussion unfolded in subsequent sessions.

Session 1.6. Focus of the new guidance (P3 & P23)

This session involved presentations summarising the results from the previous project phases and setting the stage for more focused and informed discussions on Day 2. Data for these presentations were taken from different chapters of this thesis work (see Chapters 1, 3, 4 and 5).

Session 1.7. Day 1 discussions (P18)

At the end of Day 1, a thirty-minute discussion was held, where participants were invited to respond to the presented results from the previous phases of the research and share their views on any aspect of the GRADE guidance for complex interventions. Different topics and issues were discussed in this session, including the recent revisions of the GRADE guidance by GWG regarding the initial categorisation of evidence based on study design. Participants agreed that this revision would be relevant for the GRADE guidance for complex interventions as well, however, contrasting views were expressed on the matter (see thematic analysis below). Furthermore, a request was made in the group to have a brief presentation on the new tool for assessing Risk of Bias in Nonrandomised Studies of Interventions (ROBINS-I) on Day 2, as the development of this tool has served as an important incentive for GWG to revise the existing GRADE guidance to initially rate evidence from all types of study designs as "high" certainty. Participants also discussed issues related to the remit of the GRADE guidance for complex interventions questioning the extent of changes to the existing GRADE guidance that the group would be willing to support:

> *P10: The last question I have is, um, um, are we trying to stay within the "GRADE box" [in terms of thinking about the content of the new guidance] or are we really trying to think out of the GRADE box? So, within the GRADE box, I think there's a lot that we can do. So, relating to some of this offering better guidance on how to deal with indirectness, offering better guidance on how to conceptualise complexity and so on. If we think outside the GRADE box, then we'll come to some of the issues that will be raised tomorrow, such as, what do we mean by certainty of evidence? Um, and then I think, you know, if we are ready to walk down that path, then there might be some bigger changes on the horizon. (Guideline developer & methodologist; Cochrane)*

On a related note, participants discussed issues regarding the audience of the new guidance and mentioned the importance of its widespread dissemination and

uptake. It was noted that an official approval by GWG will enhance the buy-in of the guidance by relevant stakeholder groups. By corollary, members of GWG present at the meeting highlighted the importance of collaborating across different sub-groups within GWG so that the new guidance speaks to the recent developments and publications of the Group. Finally, the added value of the new guidance was questioned: participants outlined aspects of the existing guidance that would benefit from a more nuanced description, including a better description of the construct of "certainty of evidence"; the upstream effects of the new guidance were also highlighted:

> *P11: One of the things we find our review authors struggle with, I know it's almost impossible to get them to do well is to articulate how the intervention went well. That's, that's where some of this trouble starts … When people come to talk about conclusions and how to interpret them, they're not really answering a very interesting question. (Journal editor & systematic reviewer; Cochrane)*
>
> *P26: Is GRADE the best for that? (Journal editor & systematic reviewer; Cochrane)*
>
> *P11: Well it might be … Good guidance on GRADE, and we might start thinking about some of the upstream effects of that. (Journal editor & systematic reviewer; Cochrane)*

Day 2 Minutes

Session 2.1. Review of Day 1 and aims of Day 2 (P18)

Key points from Day 1 were reviewed. These included the need to agree upon the suitable definition of complexity for the new guidance, further decide on the required changes to the existing GRADE guidance ("structural changes" vs. "fine-tuning"), as well as have a small presentation about the ROBINS-I tool (participants with more expertise in ROBINS-I volunteered to do this). The agenda for Day 2 discussions was then outlined.

Session 2.2. Clarifying the conceptual framework of the GRADE ratings: update on the recent publication of the GRADE Working Group (P3)

As mentioned above, it was found important by project co-investigators to have a dedicated discussion on the recent paper published by GWG regarding the construct of "certainty of evidence" (Hultcrantz et al., 2017). The presentation made at this session, thereby, aimed to summarise the content of this paper to encourage further discussions on the appropriate definition of the construct in the group in relation to the GRADE guidance for complex interventions. In this new paper, GWG suggests a revised conceptualisation of the construct as *"confidence that the true effect lies on one side of a specified threshold or within a chosen range"* (Hultcrantz et al., 2017). It is worth remembering, that in the current guidance certainty of evidence is defined as *"the extent of confidence that an estimate of effect is correct"* (in a systematic review context), or *"the extent of confidence that an estimate of effect is adequate to support a recommendation"* (in the context of a review informing a guideline development) (Balshem et al., 2011).

Session 2.3. Conceptualising "certainty of evidence" for complex interventions: response to the publication presented above (P10)

Discussions of this session began by clarifying the content of the GWG paper summarised above, specifically, the difference between the three approaches to rating the certainty of evidence (see Chapter 1 and 4 for more details on these approaches). Participants then expressed their views on these approaches; concerns were raised for the fully contextualised approach being very complex in light of difficulties of setting thresholds, and many participants supported the noncontextualised approach as a more

feasible way of defining certainty of evidence in reviews of complex interventions (see thematic analysis below). During this session, a member of GWG highlighted the difference between two types of papers published by GWG: conceptual and guidance papers. While GRADE guidance papers aim to provide users with *feasible* methods and tools, GRADE conceptual papers are written "*to push the boundaries of thinking"* within GWG to develop and test *new* approaches. An example of the GRADE conceptual paper is summarised in the previous chapters discussing the three approaches to defining certainty of evidence (Hultcrantz et al., 2017). In this view, a suggestion was made to publish the output of this meeting and the project more broadly, as a GRADE conceptual paper, in case if the meeting participants decide to adopt changes, which are *"outside of the GRADE box"* (see thematic analysis below). Participants also discussed issues related to the content of the new guidance, including the need for illustrative examples of interventions beyond clinical practice and recommendations, which can be used consistently by reviewers within the same discipline.

Session 2.4. Discussion: GRADE guidance for complex interventions (P18)

Different topics were discussed at this session ranging from challenges of GRADE use in reviews of complex interventions to discussions around the added value and dissemination of the new guidance. The perceived lack of uptake of the GRADE approach in wider social domains of practice beyond clinical medicine were attributed to both the lack of tailored guidance, such as scarcity of worked examples of GRADE application in social interventions, as well as broader structural factors associated with the lack of RCTs for these interventions:

> *P2: In some of the criminology areas we have RCTs, but we're still going to be primarily working from quasi-experiments, where some are nice quality and*

> *control for confounding quite nicely, but they're still going to be downgraded, and I think that most, most of the reviewers I've spoken to about, you know, the tools they use, and myself personally, haven't used GRADE on the basis that our primary evidence source just is sitting "low", and not that much possibility of grading upwards. (Journal editor & systematic reviewer; Campbell)*

Participants noted the importance of clearly demonstrating the value of using the GRADE approach and the need for further methodological training of reviewers as part of the dissemination activities for the new guidance. WHO was viewed as an exemplar of progress over the last decade in endorsing the GRADE approach and the principles of the evidence-based practice more broadly, when developing global health recommendations. The rest of the discussions of this session revolved around the details of the ROBINS-I tool and how it might affect the initial categorisation of evidence based on study design in the GRADE approach (see thematic analysis below).

Session 3.1. Study design (initial categorisation of evidence; P3)

The session began by a presentation summarising the main concerns associated with the current GRADE approach, which initially categorises evidence from RCTs as "high" certainty and all other study designs as "low" certainty (see the previous thesis chapters). Two modifications to this approach were discussed at this session: initially rating evidence from all types of study designs as "high" certainty provided an appropriate tool is used to assess risk of bias in nonrandomised studies, such as ROBINS-I; and introducing an additional category of "moderate" certainty to the existing GRADE guidance to initially distinguish quasi-experimental designs, such as interrupted times series, which are believed to be stronger in making causal inferences as compared to observational studies, such as cohort studies. Participants expressed different opinions

on these approaches, discussing pros and cons for both of the options and providing arguments from methodological and practical perspectives (see thematic analysis below).

Voting with regard to these options was conducted at the beginning and at the end of this session. It is worth noting, that the group did not feel ready for voting on the issue at the beginning of the session, and it took rounds of discussions to agree on the aims of the vote, that is, *"to get a flavour of where people's thinking"* was regarding the options for initial categorisation of evidence in the new GRADE guidance, rather than to derive definite decisions. In both of the votes, the vast majority of the group supported the option of initially rating evidence from all study designs as "high" certainty and using rigorous tools to assess risk of bias in the body of evidence: at the end of the session, 71% of participants voted for this option, while 13% voted for integrating an additional category of "moderate" certainty for initial categorisation of evidence from quasi-experimental studies; 17% of votes favoured the current GRADE approach of initially categorising evidence as "high" (evidence from RCTs) and "low" certainty (evidence from all other study designs). Since the voting process consumed a lot of the session time, participants agreed that for the remaining of the sessions it might be appropriate to proceed without formal voting.

Session 3.2. Risk of bias (P23)

Discussions of this session mainly revolved around assessment of performance bias in reviews of complex interventions in light of acknowledged concerns of lack of blinding in these interventions. Participants agreed that the lack of participant and personnel blinding is not a reason for *"automatically"* downgrading evidence in GRADE, and highlighted the importance of using judgment to assess whether lack of blinding

leads to bias in each specific review (see thematic analysis below). Participants suggested different criteria to assist in those judgments. It was noted, that risk of bias domain in GRADE largely follows the Cochrane Collaboration's tool for assessing risk of bias (RoB). The implications of the revised Cochrane tool for assessing risk of bias in randomised trials (RoB 2.0) and the ROBINS-I tool for rating risk of bias of a body of evidence in GRADE were also discussed during this session.

Session 3.3. Inconsistency (P3)

This session included discussions on how to assess inconsistency in reviews of interventions, which exhibit large variation in implementation and contextual dependency of effects. The group agreed that the key criteria for rating inconsistency are already described in the existing GRADE guidance, however, noted that further explication of the role of logic models in outlining potential sources of heterogeneity with examples of social interventions would be helpful. The group also discussed the importance of specifying the purpose of the review (such as, "evidence for summary" vs. "evidence for decision-making", as well as "lumping" vs. "splitting" of review topics) and the approach for rating the certainty of evidence (such as, rating the certainty in the nonnull effect vs. rating the certainty in the magnitude of effect) to inform judgments of inconsistency (see thematic analysis below). Finally, flexibility of the approach for judging inconsistency was highlighted:

> *P28: The emphasis here is that when we think about inconsistency not to be rigid in thinking there has to be a sub-group analysis that explains it … It just seems that we've got to use all the tools in our toolbox, and uh, and try to explain what's going on. And if you can, I don't think you should downgrade for inconsistency. (Systematic reviewer & methodologist; Campbell)*

Session 3.4. Imprecision (P18)

In general, the group felt that imprecision assessment should not depend on the complexity perspective adopted by reviewers. The group discussed that sample size requirements for reviews in different disciplines, such as public health and psychology might vary. However, this should not be a reason to ignore imprecision assessment. More importantly, participants highlighted that high levels of observed imprecision for the estimate of effect might often stem from variability of data across included studies (i.e., high levels of heterogeneity). In this view, the value of further guidance was mentioned in differentiating between imprecision of the estimate of the effect that is a result of heterogeneity of evidence across included studies versus imprecision that occurs as a result of variability within included studies. A few participants also mentioned that the optimal information size (OIS) criterion might not be adequate for population-level interventions, where a single study might include large numbers of participants. However, no alternative or modification to this criterion was discussed.

Session 3.5. Indirectness (P23)

Participants discussed options on how best to operationalise the domain of indirectness for the GRADE guidance for complex interventions. Overall, participants felt that *"fine-tuning"* of the existing GRADE guidance can be helpful in providing further clarification on how to categorise outcomes in reviews as intermediate versus distal, as well as examples on how to assess indirectness in reviews, which use a lumping or a splitting approach. A relatively larger amount of time in the session was devoted to the discussion of assessing linked bodies of evidence along the causal chain of an intervention (see Figure 6.3). While participants unanimously agreed that the reviews of

complex interventions should better describe the causal chain of an intervention, participants were not certain whether this should be implemented as part of the GRADE ratings: a few participants made the case for incorporating the "causal chain" thinking into assessment of indirectness in GRADE, others argued to include "coherence of the causal pathway" assessment as a separate consideration (outside of the GRADE framework) in reviews of complex interventions (see thematic analysis below).

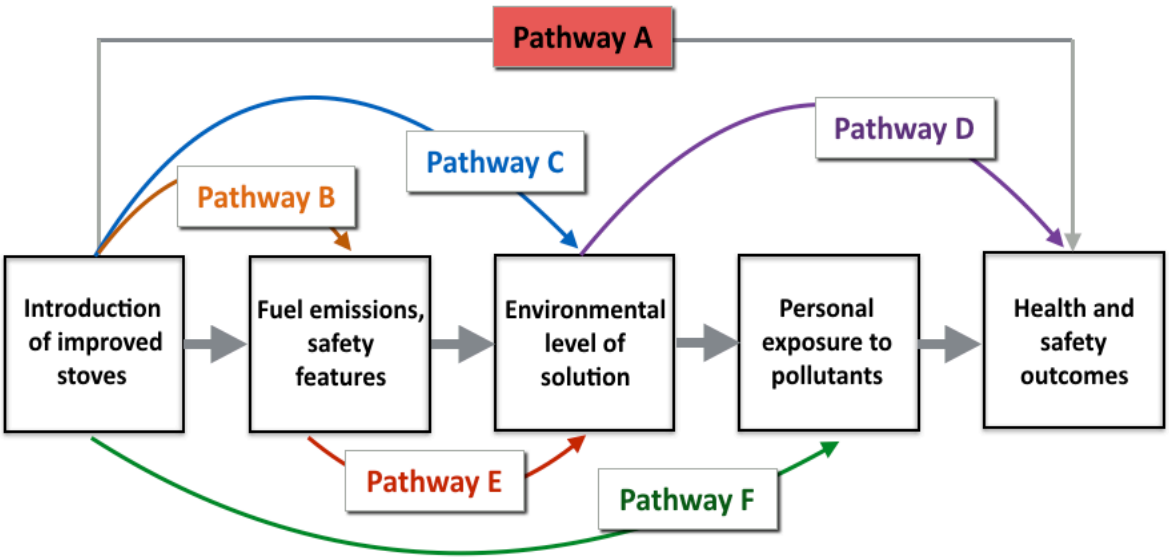


Figure 6.3. Causal chain relating household energy technology, fuel and other interventions to health and safety outcomes via intermediate links, adapted from Bruce et al., 2014.

Session 3.6. Publication bias (P18)

This was a relatively brief session, where participants discussed considerations for assessing publication bias in reviews of social interventions. Issues related to the industry sponsorship of social interventions, as well as developer bias (also referred to as allegiance bias) and the need to collect information about the research funders beyond what is reported in the publications were highlighted as important criteria to assist in

judging publication bias in these interventions. The group agreed that comprehensive multilingual literature searches are key to locating relevant studies for social and public health interventions:

> *P1: I also make a strong plea for grey literature searches and go even further than that. About supplementary search methods, citation tracking and reference list follow-up. We have done quite a bit of work, which shows that's really important for some public health topics, in particular. Um, I think the slum upgrading review was one of those reviews … 60% of the studies in that review came from the non-standard literature or from non-medical databases. (Systematic reviewer & methodologist; SURE)*

Session 3.7. Upgrading domains and additional considerations (P23)

The first half of this session included discussions on the domains for upgrading evidence in the GRADE approach. Participants asked and answered clarification questions regarding the conditions in the GRADE approach allowing to upgrade a body evidence. It was clarified that upgrading in the GRADE approach mostly applies to those circumstances, when evidence from non-RCT studies is initially rated as "low" certainty and is not further downgraded for any of the GRADE domains. In this light, the group was not sure how these upgrading criteria would be applied in the revised GRADE guidance, which suggests to initially rate evidence from all types of study designs as "high" certainty. Representatives of GGG were not sure about this issue either, and noted that GWG still needs to think about this and make further clarifications in the revised guidance:

> *P10: I always thought it had been a soft rule that you don't upgrade after downgrading. I mean, would that still apply? Because if so, then that jeopardises nonrandomised evidence even more than before. Because we assume that every nonrandomised study will somehow go down from high to moderate, low, or very low [even, when it starts off as high]. Um, so they would never have an opportunity to be upgraded. So, if that rule doesn't change, then I would be very worried. (Guideline developer & methodologist; Cochrane)*

In the second half of the session, participants returned to discussions on incorporating the "causal chain" approach into the GRADE ratings (see thematic analysis below). One of the participants noted that the GRADE guidance for network meta-analysis (Puhan et al., 2014; Salanti, Del Giovane, Chaimani, Caldwell, & Higgins, 2014) might provide insights on how to operationalise the assessment of intervention's causal chain in the GRADE approach (e.g., as part of the indirectness assessment). Lastly, a few participants raised the "floor effect" concern that the current categories of evidence in the GRADE approach (i.e., the four levels of "high", "moderate", "low" and "very low") do not provide sufficient granularity to adequately describe the distribution of the certainty of evidence of social interventions (see results from Chapter 4). Opinions in the group diverged on whether the scale of the GRADE ratings should be extended to allow for further levels of evidence or whether more effort should be put into how the existing levels of evidence are interpreted and communicated to decision-makers (see thematic analysis below).

Day 3 Minutes

Session 4.1. Aims of Day 3 (P18)

A project co-investigator outlined the agenda for Day 3, which included reviewing key issues from Day 2, a dedicated session to revisit selected topics from the previous discussions with the aim to reach resolutions, and discussions on drafting and disseminating the new guidance.

Session 4.2. Key Issues from Day 2 (P3)

Key points from Day 2 were summarised. Overall, the moderator noted that the group agreed on operationalising the complexity perspective in the GRADE guidance by

way of outlining specific sources of complexity that may be relevant in reviews depending on their aims and purpose. In terms of the added value of the new guidance, it was discussed that the new GRADE guidance, which incorporates a complexity perspective may promote the knowledge and use of GRADE in non-healthcare sectors. However, in doing this, it needs to address the structural barriers to adopting GRADE in those sectors, such as the specifics of the evidence base (e.g., more frequent use of nonrandomised study designs). The opinions also converged in the group regarding the five domains of GRADE for downgrading evidence as important constructs to consider when rating the certainty of evidence; participants agreed that further clarification with worked examples on how to interpret these domains from a complexity perspective would be beneficial for reviewers across different disciplines. This was referred to in the group as *"fine-tuning"* of the existing GRADE guidance.

Disagreements in the group primarily related to issues that participants framed as *"thinking outside of the GRADE box"*. These primarily included decisions on how best to define the construct of "certainty of evidence" from a complexity perspective (e.g., certainty in the nonnull effect vs. certainty in a specific magnitude effect), how to initially categorise evidence based on study design and apply criteria to upgrade evidence, how to incorporate a causal chain approach into the GRADE ratings, and finally, whether or not to extend the categories of the GRADE ratings. A decision was made to further discuss these issues on Day 3.

Session 4.3. Finalising the GRADE guidance for complex interventions (P23)

Participants revisited the most controversial issues from Day 2 with the aim to reach resolutions for the new GRADE guidance. Specifically, further discussions were held

regarding the initial categorisation of evidence based on study design: while the majority of the participants supported the recent revisions of GWG to initially rate all study designs as “high” certainty and use a rigorous risk of bias tool for nonrandomised studies, an opportunity was given to participants in this session to openly raise their concerns with this approach and make the case for an alternative categorisation of evidence (see thematic analysis below). Similarly, participants discussed the different approaches to defining the certainty of evidence, when using a complexity perspective, and agreed on the threshold-based approach, specifically, rating the certainty in the nonnull effect as the most suitable for the new GRADE guidance. Flexibility in providing different options for different audiences of GRADE users was highlighted in the discussions, as well as the need to work through examples from different disciplines.

Participants also revisited the issue of incorporating the causal chain approach into the GRADE ratings; however, a clear resolution on how to operationalise this was not reached and opinions diverged on whether this should be applied as part of the GRADE ratings or separately, as a methodological tool in systematic reviews more broadly. Opinions also diverged on the need to extend the categories of evidence in the GRADE approach, and the group decided to take the issue forward to GWG for a wider group discussion. The need to liaise with ongoing relevant projects within GWG was highlighted, and areas for future methodological work were outlined (see thematic analysis below).

Session 4.4. Drafting & disseminating the GRADE guidance for complex interventions (P18)

This last session included discussions on how to proceed with drafting of the new GRADE guidance, including specifying the remit, the format and the timeline for submitting the guidance to GWG for official approval (see thematic analysis below). Representatives from different organisations and disciplines present at the meeting made suggestions on how they could contribute to drafting of the new GRADE guidance, such as by way of providing examples from their areas of practice. The need to liaise with relevant project groups, including the GRADE for Public Health Interventions, was once again highlighted at this session.

Participants further discussed issues related to the dissemination strategy of the new guidance. The need to develop additional outputs for non-academic audiences was highlighted, such as producing short learning webinars. Demonstration of the added value of the new guidance through provision of feasible recommendations and worked examples was seen as key for wide uptake of the new guidance across relevant stakeholder organisations, including the Cochrane and Campbell Collaborations, 3ie, CDC and AHRQ. Participants were then thanked for their time and contribution, and the meeting concluded.

Thematic network analysis

The thematic analysis identified five overarching (organising) themes, which describe important arguments and decisions regarding the content and dissemination of the GRADE guidance for complex interventions (the global theme). Figure 6.4 depicts the thematic network generated using the method by Attride-Stirling (2001). The basic themes within each organising theme are further described below with illustrative quotations.

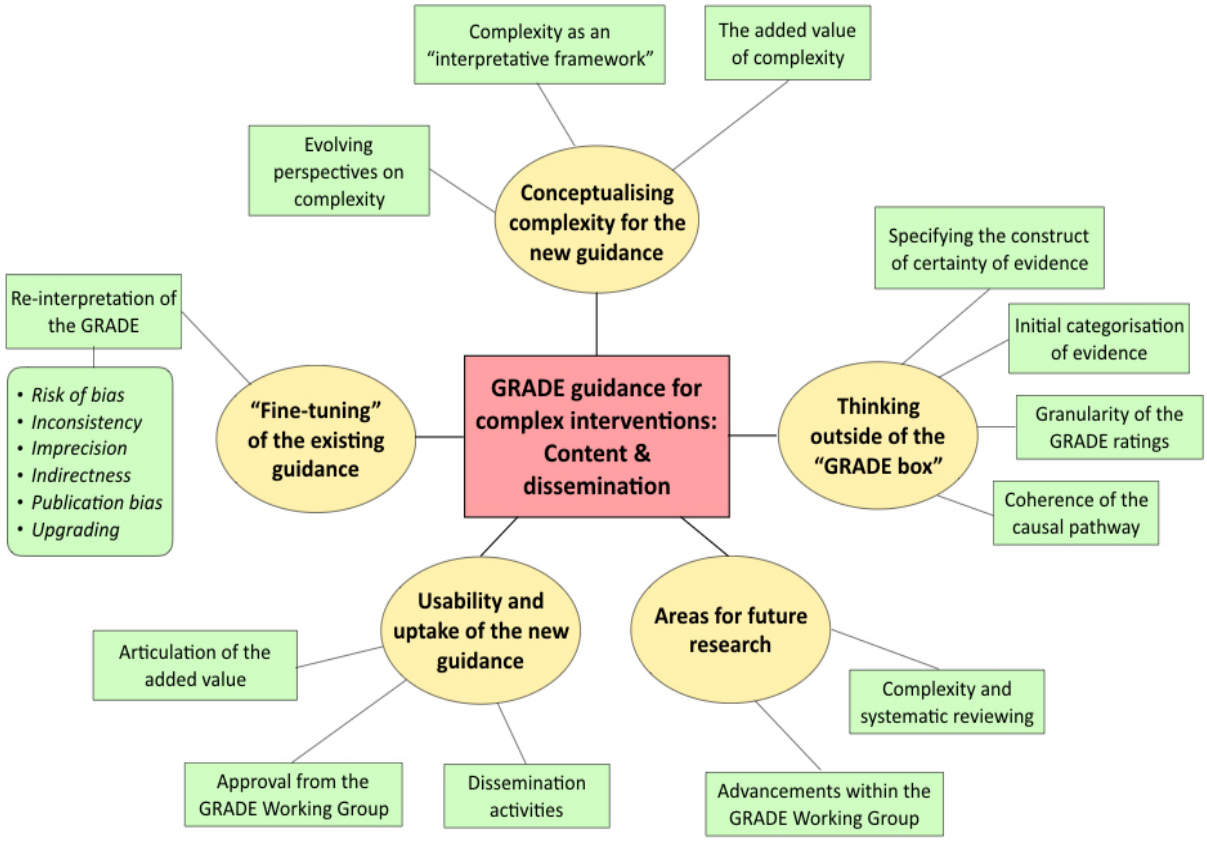


Figure 6.4. Thematic network map showing organising and basic themes on the content and dissemination of the GRADE guidance for complex interventions

Basic themes
Organising themes
Global theme

Conceptualising complexity for the new GRADE guidance

Throughout different sessions and days of the meeting, participants deliberated over how best to define "complex interventions" for the new GRADE guidance, including discussions on how the concept of complexity has been developed and used in intervention research in the recent decade, a decision to move away from framing of "complex interventions" to viewing "complexity as an interpretative framework", and finally, reflections on the added value of adopting a complexity perspective in GRADE and in systematic reviews more broadly.

Evolving perspectives on complexity

Discussions around how to conceptualise complexity for the new GRADE guidance started with a broader discourse within the group on how the concept of complexity has evolved in intervention research (in both primary and secondary research) over the last decade. Two interrelated perspectives were reviewed, namely "complex interventions" and "complex systems". The widely cited dimensions of complexity from the UK Medical Research Council (MRC) and other methods that were further developed based on this framework, such as the recently developed iCAT_SR tool and PRISMA-CI were referred to in the group as following the "complex interventions" perspective. In this perspective, sources of complexity are primarily located within the design of the intervention itself, including interventions being comprised of multiple (interacting) components, involving various outcomes, levels of target, behaviours, as well as flexibility in implementation across different settings. By comparison, the "complex systems" perspective was differentiated to designate a systems-thinking approach that was viewed as gaining prominence in health sciences in the recent decade. From this perspective, interventions

were described as *"events in systems"* that evoke changes at different levels of the system:

> *P10: The way I think about this is a system having many nuts and bolts, and you may be screwing some of those bolts [by introducing an intervention], and there may be multiple small changes, and overall these changes may lead to, um, system changes. (Guideline developer & methodologist; Cochrane)*

While the "complex systems" perspective was regarded in the group as an important future direction of intervention research, concerns were raised that this perspective still remains largely theoretical, and methods are currently in development in both primary and secondary research to adequately model and measure the system features, such as non-linear relationships and emergent properties. With regard to using this perspective to inform development of the new GRADE guidance, a few participants noticed that systems-thinking may be fundamentally incompatible with the principles of the GRADE approach:

> *P8: GRADE is very linear. Um, it's purposively so. It's reproducible, so, um, it's transparent. And I'm just, I think I'm trying to reconcile two things that possibly aren't reconcilable. I don't know … (Systematic reviewer & methodologist; DECIPHer)*

Several participants also noticed that the "complex systems" perspective necessarily involves broadening the scope of intervention evaluation to include further questions on intervention feasibility, acceptability and cost-effectiveness, which go beyond assessment of the estimates of effect of an intervention. Many of the considerations informed by system-thinking were, therefore, thought to be more relevant for the Evidence to Decision (EtD) frameworks for grading the strength of recommendations for practice, rather than the GRADE approach for rating the certainty in the estimates of effect of interventions. It was clarified in the group that the goal of

this meeting is to think about complexity with regard to *"studies of effectiveness, and confidence in what comes out of a systematic review of studies of effectiveness".* In this view, a few participants further argued that, even at a conceptual level, both of the perspectives of complexity (i.e., complex interventions and complex systems) can serve to elucidate important dimensions in systematic reviewing to produce more informative results:

> *P10: If we can just conceptualise these things [how we do systematic reviews and the way we synthesise and rate evidence] differently, they will be better as a result. It doesn't always need a very different method, sort of further on in the process. But I think just thinking, I mean, I work a lot with logic models, and I think just getting all these aspects of complexity on the table and thinking, oh, how does that influence my synthesis strategy? How does that influence my definition of "certainty of evidence?" (Guideline developer & methodologist; Cochrane)*

Complexity as an interpretative framework

As the discussion on how best to conceptualise complexity for the new GRADE guidance unfolded throughout the meeting, opinions in the group largely converged on viewing complexity as an *"interpretative framework",* rather than as a *"descriptor"* to distinguish between simple and complex types of interventions. This complexity framework can potentially be applied to any intervention provided it enhances the understanding of intervention effects:

> *P6: If we think of an interpretative framework that requires, for example, taking context into account, then we have to think of all interventions as complex interventions. So, even a very simple RCT of a pharmacotherapy, if you rolled this pharmacotherapy out in the US and in India, for example, you might get different results, which have to do with different adherence to the prescriptions. Um, it's about as simple of an intervention as you could get, but it's the context that's kicking in. So, is the argument really that there are no simple interventions? (Trialist in behaviour medicine)*

> *P20: I would argue in reality there are no simple interventions, but, you know, there are "simple enough" interventions that it's not terribly harmful to conceptualise them as simple. And then, there are "complex enough" interventions that it is terribly harmful to conceptualise them as simple. And I think, you know, where that line is drawn is, you know, I think is subjective. (Guideline developer & methodologist; CDC).*

Participants further discussed that many interventions in clinical practice, such as pharmacological treatments might not need a complexity perspective as, for example, the contextual factors might not substantially alter the effects, which occur in relatively shorter time spans; by contrast, interventions, which commonly operate via behavioural and social mechanisms, including many public health and health system interventions would most frequently benefit from a complexity perspective and a more nuanced conceptualisation of how the effects may unfold over longer periods of time in multi-factorial environments. Health system interventions, such as introduction of a national law, were discussed in particular as examples of "simple" interventions in design, which, however, require a complexity perspective to adequately assess *"how the effects may reverberate through the system"* to produce system changes in a long-run. In this view, one of the participants referred to Occam's razor as an important principle to guide the decision on adopting a complexity perspective in a systematic review:

> *P10: Sometimes we don't need to worry about all of the details of complexity… And I would like to suggest we go with Occam's Razor – as simple as we can be, but we shouldn't be too simplistic … So, it's often sensible to start with something relatively simple, ask a simple question. As you then realise, you can't, you know answer – you still don't understand what's going on – make it more complex over time. (Guideline developer & methodologist; Cochrane)*

Participants also discussed whether complexity should be viewed as a spectrum of approaches or whether there is a need for a more explicit differentiation between "simple" and "complex" perspectives. The majority supported the former:

P6: It just seems important to make some statements about what those two [simple and complex] scenarios are. (Trialist in behavioural medicine)

P10: But I don't think there is a "hard and fast" rule. I mean, that's the whole thing. We can't say this is simple and it will always be simple. It depends on the perspective you adopt and the question you want to answer. (Guideline developer & methodologist; Cochrane)

P18: Yes, I think this sort of perspective of thinking about a range of complexity and simplicity is more helpful, isn't it? (Systematic reviewer & methodologist)

With regard to operationalising complexity in systematic reviews and when using the GRADE approach, many participants felt that incorporation of a complexity perspective does not necessarily require "a paradigm shift" in the practice of systematic reviewing. As one of the participants argued: "*I think the whole differentiation between doing a complex review and a simple review is how you view variability in the data" (P20).* This point was further elaborated in the group on the example of how a complexity perspective can be used to scope a review, when using the PICO (Population, Intervention, Comparator, Outcomes) framework as an established approach for formulating review questions. Participants agreed that this framework does not need to change to incorporate sources of complexity in systematic reviews. However, further elements of complexity were perceived necessary to be brought into the process of how reviews are scoped, such as considering potential differences in implementation strategies across different settings, contextual mediators and moderators of effect and outcomes at different levels (Pfadenhauer et al., 2017). In this regard, another participant noted that complexity can be addressed within the existing paradigm of practice:

P9: By the end of the day, complexity is just about "more to think about". Like, you do a review on an intervention and you need to go and dig and understand what are the differences in the characteristics, and sources of variability … I think in my mind, it becomes more challenging and more complex, when there are more

> *factors or variables to think about; but it's conceptually the same approach of trying to understand what are the factors that are leading to that variability. (Guideline developer & methodologist; GGG)*

The approach towards operationalising the complexity perspective for the new GRADE guidance, which extends on the MRC framework was met positively by the group. In this approach, complexity is defined through description of *key sources of complexity* in systematic reviews. When considering these sources of complexity, however, suggestions were made in the group to extend beyond those that are informed by the properties of the interventions themselves (as it is currently described in the MRC framework), to also consider sources of complexity that stem from system properties, such as interactions of effects with different contextual factors, as well as causal pathways involving different intermediate and long-term outcomes.

The added value of the complexity perspective

Most frequently, participants discussed the value of the complexity perspective in enhancing the "explanatory power" of systematic reviews, that is, providing a framework to better describe and explain variation in the effects of an intervention:

> *P19: I think the big part of the answer to the question why bother with complexity is that if you ignore it, it makes the review simpler, but it turns all the interesting and important variation just back into "noise" – you know, back into heterogeneity. And what we tend to want to know with the evidence is whether something can be implemented effectively in a particular setting. Fundamentally it can be in any setting, which is why we need to take account of these, these specific aspects of complexity. (Systematic reviewer & methodologist; MRC/CSO)*

In relation to this, another participant provided an example of how a complexity perspective may inform the rating of the domain of inconsistency in the GRADE approach when estimating the effect of an intervention to prevent road traffic injuries. In a simple conceptualisation, the review may produce evidence of inconsistent effects (by way of

using the index of I^2 or a statistical test for heterogeneity), as all the contextual factors will be ignored. However, from a complexity perspective, which accounts for important contextual factors across different countries, variance will be broken down into *"error variance"* and *"substantive variance"* due to these contextual effects, and analysis will be stratified accordingly. A few participants mentioned that this conceptualisation provides a reviewer with a mind-set that *"there are unexplained sources of variance, which are not error variance"*. By corollary, it highlights the importance to think about a range of intervention effects, which are contingent upon various implementation and contextual factors, rather than aiming to estimate one true effect. Similarly, another participant provided an example of how using a complexity perspective has helped better explain the effects of sanitation interventions:

> *P10: For sanitation interventions, we have learned that you need something like 98% of the population to have a toilet and to have a functioning sewerage system, so that I, as an individual can actually benefit from it. So, what we think is going on is actually a herd effect. But we only learned about this in the last ten years or so. All the RCTs we have done in the past, show highly inconsistent results, because we didn't take into account that it actually mattered whether it was 10%, 50% or 90% of the population that had toilets. So that's a nice example of these sorts of multiple outcomes at multiple levels, and how conceptualising the thing differently as a complex system – we can literally think about the sewerage systems – makes a real difference. (Guideline developer & methodologist; Cochrane)*

Thinking outside of the GRADE box

A second overarching theme of the meeting discussions concerned advances of the GRADE approach, which were perceived as *considerable* changes to the existing GRADE guidance. Participants commonly referred to these changes as "thinking outside of the GRADE box", and agreement around these changes was not fully realised in the group. More specifically, these changes related to the need for further specification of

the construct of "certainty of evidence"; decisions regarding the initial categorisation of evidence in GRADE; addition of a new domain of evidence in GRADE, namely "coherence of the causal pathway"; and extension of the four categories of the GRADE ratings.

Specifying the construct of "certainty of evidence"

Overall, meeting participants thought positively of the recent paper published by GWG, which describes different approaches to defining the construct of "certainty of evidence". As one of the participants noted: *"Um, I think, conceptually, this is, I absolutely love this approach. I think it addresses all of my concerns that I've had forever about, um, the certainty, um, approach in GRADE" (P20).* Participants highlighted that the original conceptualisation of certainty of evidence with a focus on the accuracy of the point estimates of the effect did not work well for many complex interventions, because estimates of the effect can never be accurately measured in these interventions, and there is always a range of true effects that reviewers should be considering. Although participants valued the flexibility of the options that the new conceptual paper suggests to rating the certainty of evidence depending on the context of the review (i.e., noncontextualised, partially contextualised and fully contextualised), there was some disagreement among participants as to the level of flexibility that the new guidance should provide to reviewers. A few participants, specifically, representatives of GGG, argued that provision of *"more than one way of doing things"* is a more progressive approach, and that the GRADE Working Group itself has moved from promoting specific approaches to providing different options for reviewers depending on their circumstances. Another participant supporting this thinking, further clarified that *"flexibility doesn't necessarily mean changing your mind half-way through, but rather*

that there are different options available to you to start with; commit to one and when it becomes ridiculous, think again" (P26). A few participants, however, felt that this flexibility may cause confusion in the interpretation of the GRADE ratings and made the case for a unified approach among reviewers in the same discipline:

> *P2: I think potentially our issue is going to be to get every one of the reviewers to ultimately synthesising the evidence for a policy audience – to start using the same framework. Otherwise, you've potentially got someone saying I've got high confidence that there is an effect, and I've got moderate confidence that the effect is this big. (Journal editor & systematic reviewer; Campbell)*
>
> *P9: Yes, I think there has to be some sort of consistency, otherwise we confuse people with what we are rating. So, for a certain audience, we have to stick to one interpretation. And the one that we think is best for them to have an appropriate finding – so, people might argue that in public health what you need – is to say that there is an effect [noncontextualised] … But in other fields, you have to know if the effect is small, moderate, or large [partly contextualised]. (Guideline developer & methodologist; GGG).*

While the question on the most useful approach for defining the certainty of evidence in reviews of complex interventions kept recurring throughout different sessions of the meeting, many participants, specifically, reviewers and guideline developers in public health made the case to follow the noncontextualised approach using the null effect as the default threshold. The nonnull effect was justified as the most feasible threshold, which overcomes the challenges associated with the partially contextualised approach and the need to specify the magnitude of the effect for different outcomes (i.e., what is considered as large, medium, small or trivial effect for different outcomes):

> *P10: In the context of this meeting, for criminological, educational, and public health interventions, and if we adopt a complexity perspective, I would suggest, you know, the nonnull effect as the most suitable way forward, because it will remain in this kind of noncontextualised world and be relevant for Cochrane and Campbell reviews, and it would also be applicable to WHO guidelines, because*

those are also done at the global [noncontextualised] level. (Guideline developer & methodologist; Cochrane)

P20: I 100% agree with your statements … From a practical perspective, it avoids the systematic reviewer having to go through a process of coming up with some semi-arbitrary thresholds. So, the job of systematic reviewer is to present the evidence with whatever estimate of effect comes out of the review, plus the rating of the certainty of a nonnull effect. And that I think would appropriately leave the decision about what the threshold is for small, medium, large, or meaningful effects to the actual decision making body [e.g., in local or national guidelines]. (Guideline developer & methodologist; CDC)

Participants felt that the certainty of evidence is currently interpreted rather implicitly and inconsistently by reviewers (e.g., low certainty of evidence is frequently interpreted as low confidence that there is a protective effect rather than low confidence in a specific magnitude of effect). In this view, a new GRADE guidance which clearly articulates and takes a position with regard to the construct of "certainty of evidence" was perceived to bring further clarity and consistency into systematic reviewing. As a potential challenge to developing a guidance, which suggests a single approach towards defining the certainty of evidence, participants discussed that the choice of the appropriate approach might depend on the decision-making context, which might vary across disciplines and practice domains. For example, participants argued that in health and social policy, decision-makers are mostly interested in the direction of the effect, and whether the policy may demonstrate an overall protective effect; by contrast, representatives from the fields of criminology and education argued that in these fields decision-makers are most frequently interested in the size (i.e., magnitude) of the effect. In response to this, proponents of the noncontextualised approach further argued that rating the certainty of evidence in the nonnull effect does not diminish the importance of considering the magnitude of the effect in systematic reviews:

> *P10: Let's say we have moderate certainty in that toilets could generally protect against diarrhoea. And the Cochrane review would also provide an effect estimate with the confidence interval. So, the whole idea of the magnitude of effect isn't gone … It's just not what we place the rating on … I don't think it would really change what we are doing. We would just be very explicit about what can be done in this sort of noncontextualised scenario and how much, you know, confidence we have in what is coming out. (Guideline Developer & Methodologist; Cochrane)*

Participants also discussed the noncontextualised approach for rating the certainty that the effect estimate lies within the 95% confidence interval. While participants did not make particularly strong arguments for or against adopting this approach in the new GRADE guidance, one participant suggested to consider the 95% prediction interval instead of the 95% confidence interval as an important range that predicts the next population estimate of the effect. However, others in the group commented that it might introduce further complication into using the GRADE approach, as *"most people [i.e., reviewers] are bad enough even at trying to think about what a confidence interval is"* (P27).

When discussing the construct of "certainty of evidence", one of the participants made a suggestion to the group to think about its implications for operationalising the other domains of the GRADE approach: *"as we are sort of revisiting what certainty of evidence actually means, this may actually impact how we view some of the kind of domains we apply lower down"* (P10). While this topic was not extensively discussed in the group, a few participants commented how adopting the approach to rating the certainty of evidence in the nonnull effect may particularly inform assessment of risk of bias and inconsistency in GRADE:

> *P23: I wonder if inconsistency is particularly one of those domains where it really matters how the construct of "certainty of evidence" is specified, that is, rating certainty that (1) the effect is in this range, (2) the intervention actually "works", (3) the effect is this size, or (4) the effect is big enough. I wonder how does the*

inconsistency assessment differ if we are following the intervention "works" approach? (Systematic Review & methodologist)

Initial categorisation of evidence

One of the most controversial themes of the meeting related to the initial categorisation of evidence in the GRADE approach. Discussions on this issue recurred in most of the meeting sessions. Participants were first informed that GWG has recently revised its approach to initially categorising the evidence based on study design, and the most up-to-date GRADE guidance (to be published soon) suggests two options: either to follow the original approach of initially rating the body of evidence from RCTs as "high" certainty and the body of evidence from other study designs as "low" certainty, or to drop this initial categorisation of evidence and initially rate the body of evidence from all study designs as "high" certainty. In the latter approach, the guidance requires reviewers to use a rigorous risk of bias tool for nonrandomised studies, which assesses a full spectrum of biases, including residual confounding and selection bias. Representatives of GGG clarified that the decision in the original approach to initially categorise evidence based on study design was informed by the lack of rigorous tools to assess selection bias and confounding in nonrandomised studies. While the revised GRADE guidance does not specify a specific tool for assessing risk of bias in nonrandomised studies, this revision has been largely motivated by the development of the ROBINS-I tool:

P9: The issue of starting "high" did not happen overnight … The issue of assessing nonrandomised designs was a major concern in GWG for years. The original categorisation of evidence based on study design was being criticised for taking part of what is risk of bias assessment and putting it in with the study design. Um, and putting nonrandomised studies at a disadvantage. And I think one of the points that we [GWG] had to deal with was that ROBINS-I was coming out and there was a demand for guidance on how to deal with GRADE when you are using ROBINS-I. (Guideline developer & methodologist; GGG)

The vast majority of the meeting participants supported this revision as suggesting a conceptually more consistent and *"elegant way to prevent double penalisation of nonrandomised evidence [when nonrandomised evidence starts off as low and is further downgrade for risk of bias]" (P25).* However, major concerns were raised in the group regarding this new approach, which largely relies on reviewers adequately implementing the ROBINS-I tool. Application of the ROBINS-tool in the context of the GRADE ratings was in fact another most frequently discussed topics of this meeting. While participants thought that ROBINS-I is an important methodological development, concerns were raised that it is a very complex tool and requires epidemiological expertise in implementation, and *"even Cochrane authors might not be in a position to apply it in a meaningful way given capacities, skills and time constraints"* (*P10*). A further concern was raised that the ROBINS-I tool is currently designed for cohort studies and the work is still in progress to extend it for other nonrandomised study designs:

> *P10: My worry is, we're looking five years to the future where we could, maybe ten, I don't know, let's say about ten. But we need to know when we can feasibly do that, because at the moment we have ROBINS-I for cohort studies. We are in the process of developing ROBINS-I for five other groups of study designs, including, for example, interrupted time series or controlled before and after studies. Um, we then, we need to test all of that to make sure that the tools really work … I would be a bit worried now to base our recommendations on a development that we will see three, to five, to seven years in the future. (Guideline developer & methodologist; Cochrane)*

A few participants also noted that this new approach towards the initial categorisation of evidence following the ROBINS-I tool has not been adequately tested in GWG, which contradicts to the principles of rigour and *"working through examples"* within the Group. The major concern was raised regarding the application of the GRADE upgrading criteria in this new approach, and whether these would be maintained given

that evidence from all study designs is initially rated as "high" certainty. To this, the GGG representatives responded that rating of the "plausible residual confounding" should be integrated with the ROBINS-I tool, however, uncertainties were raised regarding how the revised guidance suggests to rate the remaining criteria for upgrading evidence, including the "dose-response relationship" and the "large magnitude of effect". Participants concluded that these issues haven't been adequately thought through in GWG, and this new approach requires further clarification and extensive testing. A few participants mentioned that GWG has mainly applied ROBINS-I to a body of evidence with cohort studies, and most of the evidence has been downgraded to "low" certainty. A general call for further examples was, therefore, made on the types of evidence that may have a "moderate" rating when applying ROBINS-I as part of the GRADE risk of bias rating.

As a potential solution to the foregoing challenges associated with the revised approach towards the initial categorisation of evidence, which largely relies on the use of ROBINS-I as the most rigorous tool for assessing risk of bias in nonrandomised studies, the group also discussed a third alternative. This suggests to modify the original GRADE approach so that certain nonrandomised studies are initially rated as "moderate" certainty. This approach was justified as being more practical and *"easy to apply"* across systematic reviewers with different levels of expertise:

> *P10: I would argue – it makes sense to have some early sorting into categories [high, moderate and low], and then we do fine-tuning through our usual, you know, risk of bias assessment processes using the best available tools, which may be ROBINS-I, once it's out for all study designs, or maybe something else. Um, so that's what we are doing for RCTs as well, right? I mean, behind the RCT label is lots of things, and we still kind of correct for that as we do the grading. (Guideline developer & methodologist; Cochrane)*

However, this approach was supported by the minority of the meeting participants (see the voting results presented in Meeting Minutes). The major counterargument against this approach described the large amount of work that would need to be done to differentiate the set of study designs, which could be initially rated as "moderate" certainty. The group acknowledged that the terminology for describing nonrandomised studies is rather *"messy"* in the field, and research communities label and define nonrandomised studies differently. Furthermore, one participant observed that this approach will have very low chances of being approved by GWG, as it suggests *"regression towards hierarchies of evidence"*. An argument for dropping the initial categorisation of evidence based on study design that many participants agreed with related to the variance of certainty of evidence within versus between different study designs:

> *P19: My view is that the variance in study quality is mostly within study designs, not between study designs. So, the difference between a good and a bad interrupted time series study will be much bigger than the difference between a good interrupted time series study and a good trial. You could say the same thing about regression discontinuity studies, if you have strong confounding forces. So, automatically demoting them all to low/moderate, I think is a huge mistake. We just have to develop ways to assess risk of bias for this studies and incorporate it into the GRADE procedure. (Systematic reviewer & methodologist; MRC/CSO)*
>
> *P10: This is the most convincing argument I've heard for why we should start off as high, so I think this was really helpful. (Guideline developer & methodologist; Cochrane)*

Coherence of the causal pathway

There was unanimous consensus among meeting participants that the new GRADE guidance should better describe the causal pathway of an intervention. This was perceived to be consistent with the recently published PRISMA-CI extension, which

requires integration of a conceptual framework/logic model in reviews of complex interventions. Uncertainties and disagreements, however, were raised in the group on how to operationalise the assessment of the causal pathway as part of the GRADE approach:

> *P14: I don't know where that's going to go … But somehow that [assessment of the entire causal pathway] seems to be the key thing that complex interventions have to have that GRADE right now doesn't have. (Systematic reviewer & methodologist; AHRQ)*

A few participants highlighted that assessment of the causal pathway might require substantial *"thinking outside of the GRADE box"* as instead of looking at outcomes *independently,* it would require reviewers to think of intervention outcomes (proximal and distal outcomes) in a *linked* (interdependent) chain. On a related note, a few participants suggested to operationalise the causal pathway assessment as a separate domain in the GRADE approach, namely coherence of the causal pathway, to enable upgrading of the certainty of evidence in the distal outcomes based on evidence in the intermediate links:

> *P16: We often need to look at all the pieces of evidence in the causal pathway (i.e., different links in Figure 6.3), because we do actually know that the pathway, which directly links the intervention with the distal outcome (i.e., pathway A in Figure 6.3) is either absent or is giving us a biased effect … And if we can systematically measure all the pieces of evidence all along the pathway (like pathways B, C, and D), I can't understand how you wouldn't, you know, upgrade certainty of evidence in the overall effect. (Guideline developer; LSHTM)*

Another participant argued that assessment of the coherence of the causal pathway *"boils down to assessment of consistency across methodological features" (P20).* Studies that are assessing proximal outcomes are expected to have different methodological features than those which are assessing the distal outcomes (e.g.,

different threats to internal validity). In this view, coherence of the causal pathway could be operationalised as a domain to upgrade the certainty of evidence in the distal outcomes based on the complementary methodological features of different pieces of evidence in the causal pathway, rather than simply demonstrating consistency with an a priori logic model.

A few participants raised a concern that this approach might need extending the scope of systematic reviews to include multiple bodies of evidence for different links in the causal pathway of an intervention. However, others provided examples of reviews, which used the same body of evidence for different links in the casual pathway. One of the participants argued that the existing GRADE guidance for network meta-analysis might serve as a useful resource for operationalising this domain considering that in both cases reviewers deal with indirect bodies of evidence (Puhan et al., 2014; Salanti et al., 2014). Participants, therefore, also discussed the option of integrating causal pathway considerations when assessing the GRADE domain of indirectness:

> *P27: One way of thinking about this might be, imagine you have a high certainty evidence, at least in terms of low risk of bias, for pathway C [see Figure 6.3]. And then you have this issue about the evidence being indirect. Then let's say you have kind of observational data, so probably greater risk of bias data that gives you information on pathway A. The question is, should we be downgrading for indirectness? So, in that situation, I am wondering if actually we can probably have greater certainty that we don't need to worry about indirectness if we've got additional information. (Systematic reviewer & methodologist; MRC/CSO)*

In general, the representatives of GGG mentioned that operationalisation of this domain would be supported by GWG, as it may link well with ongoing work of the Group looking at data modelling. It was mentioned, that so far, when dealing with linked bodies of evidence, GWG has been looking at the link in the causal pathway with the lowest certainty of evidence to make judgment about the certainty of evidence in the distal

outcomes. A few participants however, criticised this approach as potentially distracting reviewers from looking at linked bodies of evidence:

> *P9: And we've been through this with modelling for which we had a three-day meeting a couple of weeks ago, where we had linked bodies of evidence … If one of the links in the pathway is rated as low certainty, then the entire sequence would be rated as low. So, it would be the lowest among the ratings. (Guideline developer & methodologist; GGG)*
>
> *P10: I just want to play devil's advocate in relation to this point … Because, if I am looking at sort of two or three links in the pathway and then combining bodies of evidence in the way you've described, I am not doing myself any service. Chances are, one of these links is going to be low. So, I might be better off just looking at one link and not even entering this area of linked evidence. Because there is not much benefit to gain from it, if, you know, the lowest certainty link is going to drive what happens. (Guideline developer & methodologist; Cochrane)*

Towards the end of the meeting, and as the issue of assessing the coherence of the causal pathway was revisited, a few participants suggested to assess this domain outside of the GRADE framework, and as a separate consideration in systematic reviews of complex interventions:

> *P10: My suggestion would actually be to, kind of, address this almost outside of the, GRADE framework. To have it at the very beginning as a way of, you know, scoping your review. And a way of also determining what outcomes we want to look at, what analysis we want to do and all of that. But, it's more of a systematic review feature … And I would then suggest that you really look at the causal pathway at the end of the process again. And then look at the coherence across the pathway overall. You know, where are the big gaps? And where are bits in the pathway that we actually feel quite strongly about? So, um, we can certainly use it as a book mark at either end of the process. But, I don't think it sort of directly impacts on how we do rating in GRADE. (Guideline developer & methodologist; Cochrane)*

Granularity of the GRADE ratings

Another key topic that was extensively discussed at the meeting with controversial viewpoints related to the concern also raised in interviews with review

authors and GRADE methodologists (see Chapter 4), namely, the lack of granularity of the four categories of the GRADE ratings (i.e., high, moderate, low and very low). Participants thought that these categories do not allow for adequate description of distribution of the certainty of evidence in complex interventions, specifically, at the low certainty end, and discussed extending the current scale of the GRADE ratings.

> *P27: I used to work with Governance – if I were writing GRADE type summaries for a lot of policy options, and if all policy options come out as very low certainty, that is of no value at all. So, so, there needs to be some way of communicating that maybe you have a greater level of certainty for some of these [very low certainty] options … It suggests that maybe we need five or six categories. But, it's this issue of actually having anything presented as very low certainty is of no use for decision-making. (Systematic reviewer & methodologist; MRC/CSO)*

A few participants noted that this may be one of the reasons why the GRADE approach is not being used in many policy contexts, as it renders the majority of evidence as "low" or "very low" certainty despite that it might be the *"best evidence ever to be achieved"* for those interventions. Participants felt frustration that a body of evidence comprised of many studies with *good* designs, such as interrupted times series gets the same low or very low certainty of evidence rating as a body of evidence with case series only:

> *P2: I mean for so many reviews outside of an RCT world, you don't have a scale of 1 to 4, you have a scale of 1 and 2 [low and very low], and a scale of 1 and 2 just isn't it … This evidence, this evidence, and this evidence, there is a difference between them and we can't capture that in any way that is meaningful for communication without applying a narrative over it. And if we have to apply a narrative over it, I'm already doing that without using GRADE. (Journal editor & systematic reviewer; Campbell)*
>
> *P20: Right now as GRADE is constituted, the discriminant validity is, is really for RCT-based bodies of evidence and there is very poor discriminant validity for observational studies. So, it would be akin to giving the Graduate Record Exam to high school students. It's just not suitable. (Guideline developer & methodologist; CDC)*

Divergent views were expressed in the group regarding the suggestions to extend the GRADE categories of evidence. Some participants, particularly the representatives of GGG, felt that the main issue lies in how the certainty of evidence ratings are communicated and presented to decision-makers. These arguments maintained that it is important to relay the uncertainty associated with the estimates of effect that are rated as "low". In this view, suggestions were made to frame the "low" certainty of evidence ratings as evidence that decision-makers should be acting upon, rather than evidence that should be dismissed; however, the need for further research on how to adequately communicate the "low" certainty of evidence ratings to decision-makers was acknowledged, particularly in those situations, when it is "the best evidence possible" that is downgraded. Participants also mentioned that efforts should be put to educate decision-makers to better interpret the certainty of evidence ratings, rather than aiming to present "low" certainty of evidence ratings as higher. In relation to this, other participants noted that authors of systematic reviews should take the responsibility to interpret the "low" certainty of evidence ratings by way of discussing the implications for further research not merely in terms of further research for more accurate point estimates, but rather, for better understanding of the intervention impact more broadly, such as issues related to assessing feasibility of an intervention and people's experiences:

> *P26: In the scenario, where you cannot get to high certainty, I think the onus is on you to say, well, you know, we can't achieve that certainty, and actually there might be reasons for it in the field that you are working in. We need to start looking at evidence that helps you understand what's going – adherence, admission to hospital … Let's start investing in resources, you know, in things that actually help us understand this rather than, you know, looking only at point estimates… (Journal editor & systematic reviewer; Cochrane)*

One participant raised a concern of the downstream effects associated with extending the current scale of the GRADE ratings, that is, how the assessment of different domains of the GRADE approach would need to be implemented with regard to the new scale. Another participant also mentioned about the potential confusion in the GRADE application that a new scale might introduce:

> *P17: How would you explain that to people who are used to seeing just the four categories of evidence? That's an example and there are others as well, when education examples changed the grades, the employers throw their hands up and say, you know, I thought I knew what a certain grade was, and now students are coming through to me with a different grade. (Methodologist; WWW)*

A few participants also raised concerns that a new scale for GRADE ratings will have low chances of approval by GWG. In this light, suggestions were made to, first, work through examples to build the rationale for extending the GRADE categories of evidence by way of demonstrating the lack of sensitivity of the current ratings, and to support extending the categories slightly, rather than changing the entire scale. This was perceived as providing higher chances of being approved by GWG:

> *P27: So, the impression I get is that within the GRADE Working Group, there would be real hostility to having a different kind of level of certainty across different topics and fields. I think that would be a "no" go. However, something I wondered about, um, for a little while, is whether you could, instead of having just the 4 pluses, if you could have half pluses at the kind of 1, 1 and a half, 2, 2 and a half kind of way … It might allow the level of nuance between different evidence types to be conveyed at least a bit. (Systematic reviewer & methodologist; MRC/CSO)*
>
> *P21: Yeah, so, when comparing different guidelines, I came across a guideline group, which had ten different levels of recommendations. So, what is the distinction between level 6 and level 7? What do you tell differently? So, I think this is where, everyone has to be careful ... On the other hand, with the discussion we had, I fully understand that especially in the lower area, the observational area, yeah, de facto most of the time we only have low or very low, and the need to better distinguish good studies from not so good studies. So, this is something I*

would carry forward to the GRADE Working Group. (Guideline developer & methodologist; GGG)

Fine-tuning of the existing GRADE guidance

In addition to the more substantive changes to the GRADE guidance as presented above, there was overall agreement in the group that many acknowledged challenges of using GRADE in reviews of complex interventions can be addressed by way of providing more detailed guidance through worked examples of interventions from different social disciplines. This was commonly referred by participants as "fine-tuning" of the existing GRADE guidance. Participants discussed that many reviewers do not adhere to the GRADE guidance, in terms of exercising the needed level of judgment and flexibility in GRADE application. In this view, provision of further guidance on how to be more explicit in judgments of the GRADE domains of evidence was perceived appropriate:

P10: We're talking now about fine-tuning. I think there is quite a lot of agreement on this thing. So, for example, with regard to inconsistency domain, there tends to be a focus on statistical heterogeneity in what we are doing in reviews. And I think one point that is really important in complex intervention reviews is about carefully documenting heterogeneity and trying to make sense of it at whichever level. So, it's trying to get away from, you know, looking at just the I-square and looking at statistical heterogeneity. But that's a very minor downstream effect … (Guideline developer & methodologist; Cochrane)

It was further discussed that some of the reported challenges of GRADE use, such as those described in relation to indirectness and inconsistency assessment (see Chapters 1 and 4), can be addressed by further specifying the review purpose. This may include doing a systematic review for the purposes of summarising the available evidence (e.g., adopting a "lumping" approach) or doing a review with a clearly defined population and setting to inform a recommendation (e.g., adopting a "splitting" approach). Participants noted that these decisions made at a systematic review level can impact how the

domains of the GRADE approach are interpreted. For instance, in reviews, which adopt a lumping approach, that is, reviews, which combine estimates of the effect from a range of interventions, populations and settings, it should be expected that high levels of observed heterogeneity may produce low certainty of evidence ratings. As one of the participants observed:

> *P12: It might be to the extent that you can actually kind of lump it all together and generate a single effect estimate. It might be that actually if you have a huge heterogeneity, you have got low confidence in your effect estimate. That's not because there is anything wrong with the methodological work behind the review, but actually it may be more about how you split your effect estimates according to different contextual factors. (Systematic reviewer & methodologist; DECIPHer)*

In this view, participants highlighted the importance of making explicit decisions at a systematic review level, including specifying the review purpose and accordingly scoping review question(s), as they further exhibit *"downstream effects"* on the GRADE ratings. The value of logic models as providing a conceptual framework and outlining important considerations for the review and the GRADE ratings was further emphasised (Kneale, Thomas, & Harris, 2015; Rehfuess et al., 2017).

Re-interpretation of the existing GRADE domains

In dedicated sessions of the meeting, participants discussed how existing domains of the GRADE approach may need to be *"fine-tuned"* to address challenges associated with reviewing of complex interventions. This mainly involves re-interpretation of the existing GRADE domains of evidence on the examples of interventions from social disciplines, such as public health, criminology and education, as well as based on the adopted approach towards defining the construct of "certainty of evidence".

Risk of bias

Participants unanimously agreed that the difficulty of participant and provider blinding does not provide sufficient grounds for ignoring the bias that might arise from the lack of blinding:

> *P8: I wouldn't like to see a position, where just because an intervention is difficult or impossible to blind, um, it's suddenly not a problem in terms of potential bias. The potential is still there. (Systematic reviewer & methodologist; DECIPHer)*

> *P16: So, there is absolutely no question that what you are saying is correct and it's been demonstrated empirically in our field, so that's not the suggestion. (Guideline developer; LSHTM)*

Participants, however, noted, that lack of blinding should be differentiated from situations where awareness of the intervention, such as a mass media campaign is inherent to the intended intervention itself. A further comment was made to distinguish between lack of participant and provider blinding from risk of performance bias associated with the lack of blinding. Participants agreed that in many situations, and when integrating other considerations, lack of blinding may not necessarily lead to downgrading evidence for risk of bias in GRADE, and that it is important to provide an explanation for this decision not to downgrade:

> *P23: Downgrading for lack of blinding is not an autopilot. We need to ask whether this lack of blinding, combined with other considerations, say, lack of blinding with a subjective outcome, gives a risk of bias large enough to warrant downgrading our confidence. (Systematic reviewer & methodologist)*

> *P9: Yes, so, the Cochrane Risk of Bias tool has a question – was there blinding or no? So, the answer may be, no, there was no blinding. The next question that we should ask is [when doing GRADE] – would that lead to bias? And the answer might be no, even if there was lack of blinding. So, that would be there is no bias. (Guideline developer & methodologist; GGG)*

> *P21: Yeah, and I think if you don't grade down for bias, you need to give an explanation. (Guideline developer & methodologist; GGG)*

In a similar vein, participants observed that while there is always a possibility to blind outcome assessors, lack of blinding may not always lead to downgrading a body of evidence for the risk of detection bias. For example, this would apply to situations when objective outcome measures are employed, such as mortality.

With regard to assessing fidelity to intervention implementation as part of risk of bias rating in GRADE, one of the participants commented that the GRADE approach has adopted the Cochrane Risk of Bias (RoB) tool, and, therefore, any major modifications to the GRADE domain of risk of bias would require changes in the Cochrane RoB tool itself. In this light, a few participants noted that while fidelity to intervention implementation is not addressed in the current Cochrane RoB tool, it can be assessed under the broader domain of bias due to deviations from intended interventions in the revised Cochrane RoB 2.0 tool. A few other participants observed that fidelity to intervention implementation may be an issue related to the GRADE domain of indirectness, rather than risk of bias.

As part of the GRADE risk of bias rating, participants discussed the process in the GRADE approach of moving from assessing risk of bias in individual studies to assessing risk of bias in a body of evidence. A few participants noted that this process is not explicitly described in the existing GRADE guidance, and raised concerns with regard to rating risk of bias in a body of evidence comprised of different nonrandomised studies. Specifically, participants argued that different nonrandomised studies may have different assumptions and risks of bias at an individual study level. Therefore, when combing these studies, the different methodological features may interact with each other and produce evidence with a higher or lower risk of bias at the level of an aggregated body of evidence:

> *P20: And it seems to me that the logical way to approach the ratings on the ROBINS-I domains at the aggregated level is to think through, um, how those different weight of evidence considerations may influence your ratings, so that you could conceivably have a body of evidence where, uncertainty due to confounding on any individual study is very high, but as you aggregate the studies, your uncertainty due to confounding becomes very low, especially if you are taking the nonnull effect approach. (Guideline developer & methodologist; CDC)*

Participants further discussed that methodological features of individual studies may be different even in a body of evidence, which is comprised of one type of study design, such as a body of evidence with RCTs, some of which use blinding of participants and some of which don't (e.g., looking at results from 50 studies, which don't use blinding and 5 studies, which use blinding). Participants highlighted that it is important to consider how these features interact and complement each other when aggregating them to rate risk of bias in a body of evidence.

Inconsistency

As noted above, assessment of the GRADE domain of inconsistency was viewed by participants as contingent upon the approach adopted for defining the construct of "certainty of evidence" (i.e., rating the certainty in the nonnull effect or in a specific magnitude of effect) and the review purpose (i.e., a lumping review to summarise available evidence or a splitting review to inform a recommendation). By way of illustration, if a reviewer chooses the nonnull effect as the threshold for rating the certainty of evidence, then high levels of observed statistical heterogeneity or a large value of I^2 will not be a reason to downgrade evidence, as long as the estimates of the effect have a consistent direction.

A key highlight was made by participants to try the best to document the sources of heterogeneity in reviews of complex interventions. In this view, one of the participants

suggested to follow the *"5-E"* approach, namely, Explore, Embrace, Exhibit, and Explain (this was referred to as a 5-E approach, however, only four words were outlined in the discussions). The role of logic models was further emphasised in the group as serving an important tool to help identify potential effect modifiers in reviews of complex interventions:

> *P10: You always have to pre-specify the potential effect modifiers. In clinical interventions there are not many modifiers, so they may be too simplistic. In this case you would need logic models to have a good argument of what effect these factors may have on the outcome. And I think this is consistent with the current GRADE approach. (Guideline developer & methodologist; Cochrane)*

A few participants further observed, however, that in many complex interventions it might be difficult to pre-specify effect modifiers, and, therefore, they suggested to use logic models and refine them, as necessary, later in the review process (Kneale et al., 2015; Rehfuess et al., 2017). In general, participants made the case for using different methods and tools to try to explore sources of heterogeneity in systematic reviews of complex interventions. For these, participants mentioned the potential value of qualitative comparative analysis (QCA) and tabulation of effect estimates in relation to possible effect modifiers:

> *P28: I don't think it [exploring sources of heterogeneity] always has to be by quantitative analysis. So, it seems that maybe, um, the emphasis here is that when we think about inconsistency not to be rigid in thinking there has to be a sub-group analysis that explains it. But, that other methods might help you to understand heterogeneity. (Systematic reviewer & methodologist; Campbell)*

Imprecision

In general, participants noted that assessment of the GRADE domain of imprecision isn't much subject to *fine-tuning*, as the main criteria for assessing imprecision described in the existing GRADE guidance, such as the sample size

requirements, should apply regardless of the complexity perspective. Two main points, however, were highlighted that were perceived to require a more detailed guidance. Those related to differentiation of situations, when observed imprecision can be explained by high levels of heterogeneity across included studies, and extension of the Optimal Information Size (OIS) criterion for population-level interventions.

Participants discussed that in reviews of complex interventions, the observed high levels of imprecision in the pooled estimate of effect may often be caused by variation of the estimates of effect across included studies, i.e., heterogeneity rather than variation within included studies. The need for further guidance to distinguish between these situations was mentioned to avoid double downgrading of evidence:

> *P9: I think if there are differences across studies for the same complex intervention, the imprecision may often be due to heterogeneity, and there might be some double counting. You might have very precise studies with very large samples, but because they are so heterogeneous, you have them on both sides, and they are very scattered when you pool – you still have your wide confidence interval, because they are just so widely dispersed. So in this case, the real problem is heterogeneity, and imprecision is kind of the direct result of the heterogeneity. (Guideline developer & methodologist; GGG)*
>
> *P18: So, would it follow from that then the sort of guidance that we might want to do around this point would include a clear example or couple that…that demonstrate what to do with high heterogeneity? And then further to clarify the avoidance of double counting? (Systematic reviewer & methodologist)*
>
> *P9: Yeah, yeah!*

Participants agreed that OIS serves as a useful criterion to help judge the precision of the estimates of the effect in systematic reviews. A few participants, however, noted that for population-level interventions, a more *"aggressive approach"* for judging imprecision might be needed. An argument was raised that population-level interventions usually include large sample sizes, therefore, even a single study in a review may be judged to have a "precise" estimate of effect following the current GRADE

criteria. Participants, however, did not discuss how the OIS criterion needs to be extended to account for the sample size considerations for reviews of population-level interventions.

Indirectness

In addition to the discussions on incorporating "coherence of the causal pathway" as part of the GRADE indirectness assessment (see above), participants discussed how further *fine-tuning* of the GRADE guidance might address the challenges reported in relation to assessing this domain. Participants clarified that the GRADE domain of indirectness is conceptualised as *"relevance of the evidence in relation to the research question"*. Further discussions predominantly revolved around how decisions on lumping versus splitting of review topics may help with indirectness rating. A few participants argued for a splitting approach as providing a more informative approach for judging the directness of evidence:

> *P9: If you are interested, for example, in high- and low-income countries [i.e., a lumping question is asked], and the only evidence you find is for high-income countries. So, for high-income countries, you have direct evidence, there's no issue. For low-income counties, you could either say we have no evidence or say, okay, the only evidence is from high-income countries. So, you do consider this, but for low-income countries you grade down for indirectness. Same evidence. So, it's a question of splitting, as opposed to, you know, lumping both [high income and low income settings] together. (Guideline developer & methodologist; GGG)*

Participants further agreed that using logic models in reviews of complex interventions may also help with GRADE indirectness assessment in terms of informing intervention components and outcomes (intermediate and distal) of interest (Kneale et al., 2015).

Publication bias

While participants agreed that many challenges of publication bias assessment are common across systematic reviews, such as use of funnel plots with a few included studies, participants highlighted several considerations, which might be specifically relevant for reviews of social and public health interventions. First, participants made the case for thorough multilingual searches beyond peer-reviewed databases, including searches of grey literature, review of individual study protocol databases, citation tracking, and contacting authors and relevant experts. Participants acknowledged that while a comprehensive search will not eliminate publication bias, it is essential for locating relevant literature in certain disciplines, such as social policy and economics, as *"culturally",* researchers may not immediately publish their studies in peer-reviewed journals, but rather, first, set out the findings in working papers. Issues related to developer bias, when involvement of the intervention developer affects decisions to publish the study results, was also highlighted, along with its variant of allegiance bias, when *"commitment to a certain way of delivering an intervention"* affects publication of results from a new study. Finally, participants emphasised the importance of being mindful of industry sponsorship and research funding processes more broadly, and how the need for further research funding may lead to *"cherry picking"* of results in research communities. One of the participants further emphasised the need to search for the sources of research funding beyond what might be reported in the publication itself:

> *P9: So, we are doing some work about the conflict of interest and funding of studies. We have discovered – so, there is one paper showing that Coca Cola funds more than 90 foundations, non- for profit organisations. And then, when you look at the paper, it's funded by, you know, the National Foundation for Health Lifestyle or whatever. So, just by looking at the paper you would not suspect that this is industry funded. (Guideline developer & methodologist; GGG)*

Upgrading criteria

As noted above, participants expressed uncertainties regarding the application of the upgrading criteria in the revised GRADE approach, whereby evidence from any study design can initially be rated as "high" certainty. In the original approach, the upgrading criteria are discussed mainly in relation to nonrandomised evidence, which is initially rated as "low" certainty, because of lack of randomisation and, therefore, lack of appropriate accounting for confounding and selection bias. Furthermore, as currently specified in the GRADE guidance, these criteria only apply in situations, when evidence has not been downgraded for risk of bias, imprecision, inconsistency, indirectness, or publication bias. Participants felt that these rules may not be compatible with the revised GRADE guidance, as the latter suggest dropping the initial categorisation of evidence. Although it was agreed, that the use of the upgrading criteria in the revised GRADE approach need further clarification by GWG, a few participants suggested to integrate these criteria with risk of bias assessment as a logically more consistent approach:

> *P20: I would think dose-response and residual confounding would, kind of, have places in ROBINS-I, that could cause you to not downgrade. So, I guess, logically it makes sense to not upgrade at all, but to be very focused on training people to not be overly aggressive in identifying risks of bias … To make clear, that assessment of risk of bias is relative to whatever ultimately is decided as the right basis for what your certainty is in [e.g., rating certainty in the nonnull effect] … (Guideline developer & methodologist; CDC)*
>
> *P8: That's a really good point. If you are starting at high, why do you even need to upgrade? (Systematic reviewer & methodologist; DECIPHer)*

While participants agreed that the upgrading criterion of "plausible residual confounding" can be assessed in the risk of bias domain, such as, when applying ROBINS-I, opinions in the group diverged regarding the criterion of the "dose-response

relationship". Several participants argued for assessing confounding and dose-response in the same domain, while others made the case for separating them:

> *P27: I think the dose-response issue is fundamentally about confounding. I think if you see a dose response then you might think actually the reason, um, you might be less convinced that there's confounding by looking across the body of evidence. So, I don't think you would need it as a separate issue. I think it should be part of how you move from your individual studies to your assessment of the body of evidence. (Systematic reviewer & methodologist; MRC)*
>
> *P16: I will make a case for separating out confounding and dose-response … With self-reported outcome you have this bias that operates – "oh, thank you for giving me this shiny thing. My kid hasn't had diarrhoea in the last 2 weeks." Um, the more you go back to the household, and the more you provide them with something shiny, the more the effect will go up. So, the two are sort of slightly separate. (Guideline developer; LSHTM)*

Usability and uptake of the GRADE guidance for complex interventions

A large amount of time at the meeting was dedicated to discussions related to issues of usability and uptake of the GRADE guidance for complex interventions. Participants discussed important considerations that were perceived to enhance the use of the new guidance across different stakeholder groups, including (a) articulation of the added value of the new guidance, by way of, for instance, provision of worked examples from relevant areas of social practice, (2) activities to disseminate the guidance among both academic and non-academic users, and (3) efforts to seek collaboration and approval from the GRADE Working Group.

Articulation of the added value of the GRADE guidance for complex interventions

The added value of the new GRADE guidance was discussed throughout different sessions of the meeting to the extent of questioning the need for developing a new guidance for complex interventions whatsoever. Opinions converged that a new GRADE

guidance for complex interventions is needed, and participants provided justifications as to why this was the case. Participants argued that a new guidance can help enhance the transparency and consistency of the certainty of evidence judgments in reviews of complex interventions. Most importantly, participants highlighted that in order to have wide uptake, the guidance itself needs to articulate its added value by providing further clarification of the constructs, allowing for appropriate changes to *"accommodate the specifics of the evidence for complex interventions"* (e.g., lack of RCT evidence) and working through examples from relevant disciplines:

> *P21: I think if you want to get widespread acceptance, yeah, you need to show that the guidance is useful. So, this really means you have to work through examples and show how it provides more clarity to what you are looking at. Only then, people will be willing to accept that additional burden because they get a clearer picture of what is actually going on, like, synthesising evidence that is easier to interpret, easier to deduct a guideline from it and then, it is more transparent … But you have to work through examples – from what is your area, education? Yeah, so work with examples from the education area and show people how it works and what are the challenges. (Guideline developer & methodologist; GGG)*

In multiple occasions at the meeting, participants observed that the GRADE guidance has predominantly developed through biomedical examples, which was perceived as one of the major barriers to *"endorsing GRADE outside of health sector".* To appeal to reviewers outside of health, it was found critical that the new guidance employs examples from wider social and public health interventions. In terms of the write-up of the new guidance, a suggestion was made to produce a paper following the format of the existing official publications of the GRADE Working Group (such as those published in the *Journal of Clinical Epidemiology*) and add discipline-oriented annexes, which could be separately published in discipline-specific journals, such as the *Review of Educational Research* and the *Journal of Experimental Criminology*.

Issues of feasibility were also mentioned by many participants as important in terms of affecting the uptake of the new guidance. A few, in particular, drew on their decades-long experience of working with reviewers and guideline developers to argue that the new guidance should take into account the methodological knowledge and working conditions of the reviewers, who might not always have the level of methodological sophistication as the meeting participants themselves:

> *P20: People already have so many demands and so little time, that the issues of practicality become really important. So, to suggest a way to assess certainty of evidence to CDC review authors that people see as more valid, while at the same time, not having a process that is so burdensome, both in terms of time and expertise, is very tricky, but essential to get more buy-in. (Guideline developer & methodologist; CDC)*

> *P24: Yeah, I mean I work with WHO staff who are world experts on whatever disease or system entity, but, you know, they need something they can use, something that is practical, something that is "how to" … But it's extremely essential that we don't want to miss the high-powered group-think conceptual issues either. (Guideline developer; WHO)*

Dissemination activities

Participants agreed that in general the GRADE approach has not been widely disseminated outside of healthcare, which many thought may partly explain the *reluctance* of reviewers in social disciplines, such as in public health and international development to fully embrace GRADE:

> *P22: I think you do kind of have a PR problem of GRADE outside of health. Um, because people don't know it and, you know, all the examples are about stroke and bleeding … So, I think it could be sold outside. (Journal editor & systematic reviewer; Campbell)*

Participants discussed different strategies to help with the dissemination of the new guidance. These included efforts to make GRADE more visible by way of involving journal editors to recommend the use of GRADE to review authors, as well as

development of outputs for non-academic audiences. For the latter, participants suggested to design learning webinars similar to those used in Cochrane Learning Live, *"how to"* short videos providing a quick overview on different aspects of GRADE and short guides on using GRADE in complex interventions in the form of checklists.

The need to train review authors to *appropriately* apply the GRADE approach and decision-makers to *adequately* interpret the GRADE ratings was also emphasised by a few participants.

> *P13: I just want to make a pragmatic call for training – training to be well-designed …There are different Cochrane groups that not only talk about risk of bias, but also GRADE as it currently is, is not always well understood. The, the uptake of the summary of findings table amongst Cochrane reviews is not uniform by any means. Um, it's more about training and roll out. It would have to be superior in this case, given the degree of complexity that we are considering asking. (Journal editor & systematic reviewer; Cochrane)*
>
> *P17: Yeah, thanks. Um, I want to kind of build on this point about thinking about the users, because, as I mentioned yesterday, beyond the review authors – downstream are the people who are quote users of GRADE in the sense that they see the reports that are written. So, I think there are some issues around explaining to those people what is going on as well. (Methodologist; WWW)*

Approval from the GRADE Working Group

As described above, publications, which are officially approved by the GRADE Working Group (these are referred to as official GRADE guidance and are commonly published in the *Journal of Clinical Epidemiology*) are generally perceived to have higher visibility and impact. Throughout the meeting, participants frequently stressed the importance of working in collaboration with the GRADE Working Group and following the required procedures for obtaining approval from the Group to enhance the uptake of the new guidance. It is, however, worth noting, that the process of obtaining an approval from the Group was often perceived difficult in light of the high threshold set by the

Group (the papers need to have 80% approval from the Group members to qualify as GRADE guidance). Awareness of this issue often led participants to doubt some of the suggestions for changing the domains of the GRADE approach, especially when those implied a considerable departure from the existing GRADE guidance:

> *P18: I suggest we bear in mind that in whatever we decide to do in the end, we are going to have to think carefully about uptake and how these things can be implemented. And going back to the systematic review results [see Chapter 3], GRADE is in many ways the absolute mark, and that if we want to get uptake and buy-in, we would do well to tweak GRADE rather than to throw out and start again, so we might want to think along those lines. (Systematic reviewer & methodologist)*

As the GRADE Working Group is a lively community and has many active project groups aiming to advance the GRADE methodology for different contexts of use, representatives of the Group present at the meeting emphasised the need to *"cross-fraternise"* with these other project groups to incorporate the relevant developments and updates into the GRADE guidance for complex interventions. Among these, participants highlighted the GRADE project group, which tackles the GRADE domain of indirectness and applicability, the one working on linking different streams of evidence, such as animal data in environmental interventions, another group working on assessing a body of evidence from nonrandomised studies, and finally, a project group that aims to explore the appropriate ways of communicating the GRADE ratings to decision-makers. In terms of the timelines for drafting the new GRADE guidance, participants agreed that presenting a version for feedback to the GRADE Working Group in a year time would be most feasible. A few participants further commented that approval from the GRADE Working Group should be the aim, however, in case of disagreements with the Group, the new guidance can be published separately:

P3: Yeah, I think my view would be – let's stick to the results of this discussion. So, the guidance should reflect whatever we discussed here, whatever the evidence suggests from previous project phases, and let's aim for approval by the GRADE Working Group. And then see what actually happens, what is their feedback. If there are any major disagreements in the process, I think then we will need to decide how to proceed; if it doesn't work with the Group, we would still like this to be published as a, um, separate guidance. I don't know – that would be my understanding. (Systematic Reviewer)

Areas for future research

The final overarching theme that emerged from the meeting discussions related to areas of future research, including advancements within the GRADE Working Group, as well as methodological work around complexity and systematic reviewing more broadly. While participants noted that the new GRADE guidance for complex interventions might not be able to directly incorporate these insights, a few mentioned that these areas of further methodological work could be outlined as important *"placeholders"* in the new guidance for future updates.

Advancements within the GRADE Working Group

As described above, the GRADE Working Group operates in smaller project groups to advance different aspects of the GRADE methodology. When discussing the challenges of using GRADE in complex interventions throughout the meeting, representatives of GGG described how the ongoing work of the Group could address some of those. By way of illustration, a few participants mentioned about a newly published paper for rating the certainty of evidence in narrative synthesis of results and that there is an interest in the Group to produce on official GRADE guidance on synthesising evidence when estimates of the effect are not quantitatively pooled in meta-analyses (Murad, Mustafa, Schunemann, Sultan, & Santesso, 2017). It was noted, that this paper might also have

relevance for the GRADE guidance for complex interventions. Furthermore, one participant observed that the project group that currently works on producing a revised approach towards initial categorisation of evidence when using a rigorous risk of bias tool for nonrandomised studies, such as ROBINS-I, is also interested to address the challenges of rating mixed bodies of evidence, i.e., bodies of evidence comprised of studies with different designs and methodological features:

> *P9: The issue of combining randomised evidence with nonrandomised evidence, this is more an issue of meta-analyses, as opposed to evidence rating. So, there is some work in the systematic review/meta-analysis field about this, and what GRADE is doing is trying to, you know, work in parallel to that effort. But this is still a work in progress. There's no consensus yet on what is good and how to do it. (Guideline developer & methodologist; GGG)*

A few participants noted that there is currently strong interest in the GRADE Working Group to further investigate the processes of decision-making and the needs of decision-makers. This may further help address issues related to adequate communication of the GRADE ratings to decision-makers, and whether there is a need to extend the GRADE categories of evidence. It is, however, worth noting, that this is currently discussed in the GRADE Working Group mainly in the context of clinical decision-making, rather than broader health policy-making.

Complexity and systematic reviewing

Throughout the meeting, participants outlined areas for further methodological work related to complexity and methods of systematic reviewing more broadly. These involved further development of methods for incorporating a complex systems approach into systematic reviewing, which might include methods to deal with system attributes, such as non-linearity and emergent properties, modelling and synthesis of linked bodies of evidence:

> *P10: In terms of how do we deal with this whole systems business, I think the shorter answer is that we need a broader range of methods to try to understand what's going on and certainly a simple, excuse the word, randomised controlled trial, isn't going to give us all the answers. So, it's really looking at multiple methods to try to understand different elements of the system, including modelling. We have systems modelling efforts. Um, but I think we're a long way away from really understanding what's going on. (Guideline developer & methodologist; Cochrane)*

Participants also mentioned a few relevant initiatives in the field that may further inform the use of complexity perspective in systematic reviewing and evidence assessment, such as the ongoing project by WHO to strengthen the *Retrieval, Synthesis and Assessment of Complex Health Interventions* (WHO, 2017). It was highlighted that the new guidance should be *cognizant* of the relevant ongoing work in the field, in terms of both research into complexity, as well as methods of evidence synthesis and assessment more broadly, such as extensions of the ROBINS-I tool for assessing risk of bias in specific nonrandomised study designs and their incorporation into GRADE.

Discussion

Main findings

A group of stakeholders from a range of disciplines were brought together over the course of three days to make decisions regarding the content and dissemination of the new GRADE guidance for complex interventions. Throughout the meeting, participants engaged in thorough discussions and provided rich views on the importance of considerations for the new guidance. In general, there was agreement that *"fine-tuning"* of the existing GRADE guidance through provision of more detailed explanations of the GRADE constructs and domains and worked examples from social disciplines is crucial for widespread uptake of the GRADE approach. In this view, the proposed GRADE

guidance for complex interventions may follow the similar format as other official publications of the GRADE Working Group, such as the GRADE equity guidelines (Welch et al., 2017). In the meantime, a comprehensive reporting of the thematic analysis of the meeting discussions will provide an explicit account of the decisions made about the content of the new guidance, as well as serve as a resource for researchers interested in updating the guidance in light of future methodological developments.

While the broad aim of this meeting was to reach agreement on the considerations for the new guidance, it should be noted that in several instances the agreement was not fully realised. This mostly concerned considerations, which were perceived by participants as introducing substantial changes to the content and structure of the GRADE approach (the corresponding theme is referred to in the chapter as "thinking outside of the GRADE box"). One explanation for the lack of consensus around these changes may lie in the strategy adopted by this project towards developing a new guidance for rating the certainty of evidence in systematic reviews of complex interventions. As highlighted in the previous chapters and communicated to participants throughout different phases of research, a deliberate decision was made at the outset of this project to work within the existing GRADE framework and to aim to extend the existing GRADE guidance, rather than to develop a new method (Moher et al., 2010). This has largely been driven by considerations of the primacy and large uptake of the GRADE approach by leading organisations in evidence synthesis. In this view, changes to the GRADE approach that might substantially alter the existing approach, such as changing the scale of the GRADE ratings, might fall beyond the remit of this initiative. In the meantime, it is important that these suggestions and discussions are transparently

reported to inform further methodological work in the field (see Discussion of the thesis findings in Chapter 7).

The observed disagreements and controversies around issues, such as how to specify the construct of "certainty of evidence" in complex interventions, how to rate evidence from nonrandomised studies, as well as how to incorporate assessment of the "coherence of the causal pathway" into GRADE ratings may also be partly explained by uncertainties and lack of established methods to address these issues in systematic reviewing more broadly. As mentioned by the participants themselves, methods to synthesise complex evidence are still evolving (Lorenc et al., 2016). In the meantime, discussion data demonstrate that socio-cultural factors may also provide some explanation for the observed disagreements. For instance, the findings provide some examples of differences in the values and practices within "epistemic communities" of reviewers in healthcare (Cochrane-/GRADE-affiliated reviewers) and in social disciplines (Campbell-/non-GRADE-affiliated reviewers) regarding their use of nonrandomised studies, as well as prioritisation of magnitude versus direction of the intervention effect in decision-making (Knorr-Cetine, 1999). The challenging task for the project team tasked with the write-up and publication of the GRADE guidance for complex interventions would, therefore, be to liaise these distinct communities. This would entail formulating a guidance in a way that would speak to the values and needs of social researchers on one hand, and in the meantime, meet the requirements for an official GRADE guidance set by the GRADE Working Group, which is currently mainly comprised of researchers and practitioners in biomedicine. This might require further rounds of discussions and communication among the members of the project team and the GRADE Working Group.

In addition to the differences in epistemological perspectives across communities of reviewers, discussion data also reveal differences in purposes of systematic reviews, which may further challenge formulation of a single approach whereby to interpret the "certainty" of evidence. During the meeting, participants commonly discussed how different purposes of systematic reviews may yield different standards for the certainty of evidence. By way of illustration, it was argued that reviews, which aim to summarise available evidence in a noncontextualised way (i.e., these reviews can be viewed to serve as *"enlightenment"* tool without providing recommendations for practice), may need different interpretation of the construct of "certainty of evidence", such as rating the certainty in the nonnull effect, as opposed to those reviews, which aim to inform a recommendation in a specific context (i.e., reviews playing an *"instrumental"* role in decision-making) (Murad, Almasri, Alsawas, & Farah, 2016; Weiss, 1977). Participants noted that, currently, these perspectives are not clearly explicated in systematic reviews, and that a new GRADE guidance, which openly describes how these differences and various purposes embedded in reviews may play out in the GRADE ratings would be more likely to have high usability and enhance transparency in systematic reviewing of complex interventions.

Strengths and limitations

As a final stage of expert consultation recommended by Moher et al. (2010), this study was able to elucidate what a group of key stakeholders thought about the content of the new GRADE guidance for complex interventions and to establish areas of agreement and disagreement. A particular strength of this study is that reported findings are analysed based on transcriptions of the meeting audio-recordings, which provide a

more comprehensive account of the meeting discussions as compared to meeting notes or retrospection. Based on meeting notes and personal communication, participants seemed to be satisfied with the extensive discussions and how the meeting was generally organised and managed. However, no formal feedback was sought on participants' level of satisfaction with the meeting logistics and the discussed content. Participants were purposively sampled to include a range of social disciplines and professional roles, however, they predominantly came from high income countries, specifically, the UK and the US. Furthermore, despite the moderators' best efforts to engage all participants, some stakeholder groups were more vocal in the discussions than others. This included representatives from the GRADE Working Group (specifically, the GRADE Guidance Group) and reviewers and methodologists from the field of public health. Investment of these participants in the meeting discussions may be stimulated by concerns for their specific areas of research and practice, which may diminish the broad usability of the new GRADE guidance as it seeks to equally address the concerns of stakeholders reviewing complex interventions across different domains of social practice. Nevertheless, communities of researchers in the UK and the US, as well as from the field of public health include a relatively larger number of systematic reviewers in social disciplines, providing a rationale to sample more participants with these characteristics. It is also worth noting that participants with these demographics showed greater methodological interest and acceptance to attend the meeting. The sample was overall gender-balanced.

In line with the best-practices of qualitative research (Creswell & Miller, 2000; Tong et al., 2007), the presented thematic network analysis was independently reviewed by the research assistant who helped with transcribing the audio-recordings and the

candidate's supervisor, Prof Paul Montgomery, who helped with organising and moderating the meeting sessions. In addition, the meeting minutes were also circulated among all meeting participants and revised according to the feedback received. For publication purposes, however, it might additionally be worth checking the thematic network analysis with a larger group of meeting participants. This will help to validate any subjective decision made by the DPhil candidate when coding and interpreting the data. Furthermore, it is possible that participant fatigue could have influenced aspects of the discussions, especially at the end of Days 1 and 2. As such, communication with the meeting participants throughout the write-up and finalisation of the guidance will additionally serve to verify whether their views on any of the main themes have changed post-meeting.

Further steps

This chapter provides a transparent account of the discussions on the content of a new GRADE guidance for complex interventions. While the discussions were deliberately designed around the constructs and domains of the GRADE approach for the purposes of this thesis work and the ongoing project, they also provide interesting insights on systematic reviewing more broadly, such as how to define and address complexity in reviewing. It is, therefore, recommended that this chapter be adapted and published as a manuscript in an open-access peer-reviewed journal for greater visibility and use. This meeting discussions should further be used for developing the content of the new GRADE guidance for complex interventions. Even though the write-up of the new GRADE guidance would require further rounds of informal meetings and discussions within the project team, discussions of this meeting will serve as the basis for moving

forward with designing the guidance, which would then need be tested through feedback from meeting participants and members of the GRADE Working Group. Interested participants from the online modified-Delphi process should also provide feedback on drafts prior to publication. While the new guidance may follow the format of other official publications of the GRADE Working Group, certain constructs and domains may need to be re-defined and re-phrased to better describe the use of the GRADE approach in areas of practice beyond biomedicine. In a similar vein, alongside publishing the guidance in the *Journal of Clinical Epidemiology*, discipline-specific annexes to the guidance should be simultaneously published in relevant social science journals. The project team may also need to re-think how to properly name the guidance to convey the use of "complexity" as an interpretative perspective across areas of application. Further implications of this study and the entire thesis research for the write-up of the GRADE guidance for complex interventions are presented in the following *Discussion* chapter.

References

Alonso-Coello, P., Schunemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., . . . GRADE Working Group. (2016). GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ, 353*.

Aronson, J. (1994). A pragmatic view of thematic analysis. *Qual Report, 2*(1).

Attride-Stirling, J. (2001). Thematic networks: an analytic tool for qualitative research. *Qual Res, I*(3), 385-405.

Balshem, H., Helfand, M., Schunemann, H. J., Oxman, A. D., Kunz, R., Brozek, J., . . . Guyatt, G. H. (2011). GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol, 64*(4), 401-406.

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., Ravaud, P., & Group, C. (2008). Methods and processes of the CONSORT Group: example of an extension for trials assessing nonpharmacologic treatments. *Ann Intern Med, 148*(4), W60-66.

Boyatzis, R., E. (1998). *Transforming qualitative information: thematic analysis and code development*. London: Sage Publications Ltd.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qual Res Psychol, 3*(2), 77-101.

Burr, V. (1995). *An introduction to social constructionism*. London: Routledge.

Guise, J. M., Butler, M. E., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017). AHRQ series on complex intervention systematic reviews-paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol*, *90*, 43-50.

Guyatt, G., Oxman, A., D., Akl, E., A., Kunz, R., Vist, G., Brozek, J., . . . Schunemann, H., J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol, 64*(4), 383-394.

Hultcrantz, M., Rind, D., Akl, E. A., Treweek, S., Mustafa, R. A., Iorio, A., . . . Guyatt, G. (2017). The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol, 87*, 4-13.

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ, 311*(7001), 376-380.

Kneale, D., Thomas, J., & Harris, K. (2015). Developing and Optimising the Use of Logic Models in Systematic Reviews: Exploring Practice and Good Practice in the Use of Programme Theory in Reviews. *PLoS One, 10*(11).

Knorr-Cetine, K. (1999). *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., . . . Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

Lincoln, Y. S., & Guba, E. G. (2000). *Paradigmatic controversies, contradictions, and emerging confluences.* In Denzin N. K. & Lincoln Y. S. (Eds.), The handbook of qualitative research (2nd ed.). Beverly Hills, CA: Sage.

Lorenc, T., Felix, L., Petticrew, M., Melendez-Torres, G. J., Thomas, J., Thomas, S., . . . Richardson, M. (2016). Meta-analysis, complexity, and heterogeneity: a qualitative interview study of researchers' methodological values and practices. *Syst Rev, 5*(1), 192.

Miles, B. M., & Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2).

Murad, M. H., Almasri, J., Alsawas, M., & Farah, W. (2016). Grading the quality of evidence in complex interventions: a guide for evidence-based practitioners. *Evid Based Med, 22*(1), 20-22.

Murad, M. H., Mustafa, R. A., Schunemann, H. J., Sultan, S., & Santesso, N. (2017). Rating the certainty in evidence in the absence of a single estimate of effect. *Evid Based Med, 22*(3), 85-87.

Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health Technol Assess, 2*(3), i-iv, 1-88.

Pfadenhauer, L. M., Gerhardus, A., Mozygemba, K., Lysdahl, K. B., Booth, A., Hofmann, B., . . . Rehfuess, E. (2017). Making sense of complexity in context and implementation: The Context and Implementation of Complex Interventions (CICI) framework. *Implement Sci, 12*(1), 21.

Puhan, M. A., Schunemann, H. J., Murad, M. H., Li, T., Brignardello-Petersen, R., Singh, J. A., . . . GRADE Working Group. (2014). A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ, 349*.

Reed, J., & Payton, V. R. (1997). Focus groups: issues of analysis and interpretation. *J Adv Nurs, 26*(4), 765-771.

Rehfuess, E. A., Booth, A., Brereton, L., Burns, J., Gerhardus, A., Mozygemba, K., . . . Rohwer, A. (2017). Towards a taxonomy of logic models in systematic reviews and health technology assessments: A priori, staged, and iterative approaches. *Res Synth Methods*, *9*(1),13-24.

Ryan, G. W., & Bernard, H. R. (2000). *Data management and analysis methods.* In Denzin, N. K. and Lincoln, Y. S. (Eds.), Handbook of qualitative research (2nd ed.): Sage.

Riesmman, C., K. (1993). Narrative analysis. In Huberman, M., & Miles, M. (Eds). Qualitative researcher's companion. London: Sage Publications Ltd.

Salanti, G., Del Giovane, C., Chaimani, A., Caldwell, D. M., & Higgins, J. P. (2014). Evaluating the quality of evidence from a network meta-analysis. *PLoS One, 9*(7).

Saldana, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Los Angeles, Calif.; London: SAGE publications.

Smith, J. A., & Osborn, M. (2003). *Interpretative phenomenological analysis.* In Smith J. A. (Ed.), Qualitative psychology: a practical guide to methods. London: Sage.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care, 19*(6), 349-357.

Toulmin, S. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Weiss, C., H. (1977). Research for policy's sake: The enlightenment function of social research. *Policy Anal, 3*(4), 531-545.

Welch, V. A., Akl, E. A., Guyatt, G., Pottie, K., Eslava-Schmalbach, J., Ansari, M. T., . . . Tugwell, P. (2017). GRADE equity guidelines 1: health equity in guideline development-introduction and rationale. *J Clin Epidemiol*, *90*, 76-83.

World Health Organization. (2017). *Retrieval, Synthesis and Assessment of Evidence on Complex Health Interventions.* Maternal, new-born, child and adolescent health. Retrieved Oct 18, 2017 from http://www.who.int/maternal_child_adolescent/guidelines/development/complex-health-interventions/en/

Chapter 7. Discussion

A paper including drafts from this chapter has been submitted to *BMJ Global Health*

Chapter overview

Research into complex interventions has gained traction in the recent years, and specific initiatives have been launched to develop methods and guidance for designing, evaluating and systematically reviewing complex interventions (Craig et al., 2008; Guise, Chang, Butler, Viswanathan, & Tugwell, 2017; Lewin et al., 2017; Petticrew, Anderson, et al., 2013). The research conducted in this thesis provides a timely investigation into the challenges of rating the certainty of evidence in systematic reviews of complex interventions, which takes into consideration the ongoing work and debates in the field and aims to inform development of a new GRADE guidance for complex interventions.

This final chapter discusses findings across the four phases of the thesis research and considers contributions and limitations of each phase. Drawing on the findings from the thesis work, the chapter outlines specific implications for writing up and disseminating the new guidance papers, as well as suggests next steps for consideration by the team working on the project to develop the *GRADE Guidance for Complex Interventions.* The chapter concludes with recommendations for future methodological research and the practice of systematic reviewing more broadly.

Discussion of the thesis findings

This thesis drew on the best-practice techniques for developing research reporting guidelines to inform development of a new guidance for rating the certainty of evidence in systematic reviews of complex interventions (Moher, Schulz, Simera, & Altman, 2010). Figure 7.1 below maps the thesis chapters on to the phases of the thesis research, which broadly aimed to examine the challenges of using the GRADE approach in systematic reviews of complex interventions and explore ways to address those challenges. This further helped to draw implications for the write-up and dissemination of the GRADE Guidance for Complex Interventions.

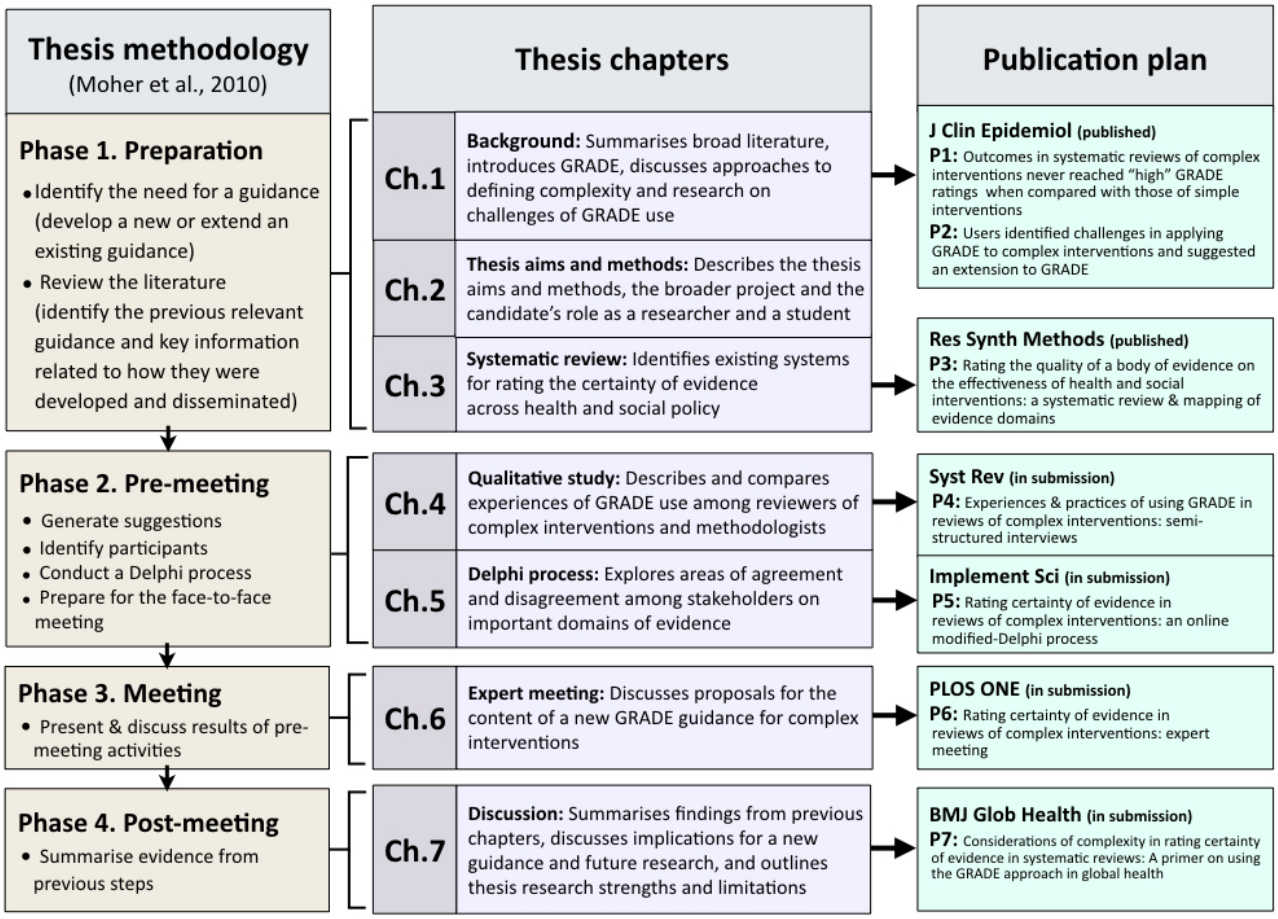


Figure 7.1. Thesis logic model outlining how the thesis chapters link with specific phases of the thesis research (Moher et al., 2010) and papers published or in submission

Findings from each research phase are discussed below; following this, implications for the GRADE Guidance for Complex Interventions are outlined.

Phase 1. Preparatory activities: Systematic review

Main findings

Phase 1 of the thesis research asked two research questions, namely, (1) what are the challenges of using GRADE in systematic reviews of complex interventions? And (2) what is the empirical evidence supporting the need for a new GRADE guidance for complex interventions? The preliminary empirical investigation conducted by the DPhil candidate (see Chapter 1) revealed that outcomes in systematic reviews of complex social and public health interventions more frequently had "low" and "very low" GRADE ratings as compared to outcomes in reviews of pharmaceutical interventions (Movsisyan, Melendez-Torres, & Montgomery, 2016a). None of the outcomes in complex intervention reviews were given "high" GRADE ratings in the sample of reviews considered. In the meantime, the most frequent reasons for downgrading these outcomes included *inconsistency* of evidence, *performance bias* and *study design* (i.e., use of NRS designs). Through further contacts with review authors regarding their experiences with GRADE, it was found that while review authors appreciated the structure and transparency that GRADE added to the review, challenges were reported with using GRADE in reviews of complex interventions (Movsisyan, Melendez-Torres, & Montgomery, 2016b). Specific challenges related to the assessment of NRSs in GRADE, the domains of inconsistency and performance bias. Application of these domains as they are currently described in the GRADE guidelines were perceived by review authors to contribute to frequent downgrading of the "best evidence possible" for complex

interventions. Thus, recommendations were made to advance the GRADE guidance for reviews of complex interventions. In line with one past investigation on using GRADE in public health, these included suggestions for minor modifications, such as provision of more detailed and tailored GRADE guidance for complex interventions, as well as major modifications to the existing GRADE guidance, such as addition of new domains of evidence (Rehfuess & Akl, 2013).

Building on this preliminary empirical work, the systematic review reported in Chapter 3 aimed to further examine the need for a new GRADE guidance for complex interventions. Seventeen systems for rating the certainty of evidence on the effectiveness of health and social interventions were identified, including the GRADE approach. The systems varied greatly in the domains they included and how they operationalised them, and most had important limitations in their development and dissemination. The construct of certainty of evidence was defined in only a few systems largely extending the GRADE approach. GRADE was found to be unique in its comprehensive guidance, rigorous development and dissemination strategy. Despite extensive searches amongst the grey literature, very few systems were identified which included domains of evidence speaking to the dimensions of complexity in systematic reviews. Most frequently these included ad-hoc modifications to the GRADE approach, such as additions of new domains of evidence (e.g., coherence of evidence in the GEPHI approach). Considering the status of GRADE among the existing evidence rating systems and its adoption by the leading organisations in evidence synthesis, this systematic review established that a new GRADE guidance which accounts for the dimensions of complexity of broader social and public health interventions will be most appropriate to ensure widespread dissemination and uptake by stakeholders. By providing a rationale

for extending the existing GRADE guidance for complex interventions, Chapters 1 and 3 also helped to set research questions for subsequent thesis chapters.

Contributions and limitations

In line with the recommendations by research reporting guideline developers (Moher et al., 2010), starting this thesis project with a preliminary empirical investigation into the use of GRADE in systematic reviews of complex interventions served as a strong foundation for scoping the thesis work and informing further research objectives. In addition, the systematic review of evidence rating systems in health and social policy considered whether an adequate system already exists that would address dimensions of complexity when rating the certainty of evidence in systematic reviews, as well as identifying potential deficiencies and limitations. This approach to guidance development, which begins with needs assessment, including a comprehensive literature review of existing guidelines, was most optimal to avoid research waste and burdening systematic reviewers with unnecessary guidance (Chalmers & Glasziou, 2009; Moher et al., 2016).

Considering the focus of this thesis work on the GRADE approach for rating the certainty in the estimates of effect in systematic reviews of complex interventions, it was appropriate to limit the scope of the systematic review reported in Chapter 3 and the entire thesis work more broadly to evidence on intervention effectiveness. It should, however, be noted that this focus on intervention effectiveness itself has been the core of the arguments of researchers supporting a complexity perspective in systematic reviewing (Petticrew, 2015). As discussed in Chapter 1, such a perspective promotes broadening the scope of systematic reviews to also consider how interventions work,

that is, to examine the mechanisms whereby interventions operate and in which contexts and circumstances (Petticrew, Anderson, et al., 2013; Wong, Greenhalgh, Westhorp, Buckingham, & Pawson, 2013). In addition to traditional Cochrane-style systematic reviews, researchers are currently considering an amalgam of approaches to evidence synthesis, including qualitative and mixed method systematic reviews that integrate quantitative and qualitative findings (Harris et al., 2015; Petticrew, Rehfuess, et al., 2013). While rating the certainty of evidence is equally important for quantitative, qualitative and mixed-method evidence syntheses, broadening the scope of this thesis work, and, specifically, the systematic review in Chapter 3, to also consider systems for rating the certainty of evidence from different types of evidence synthesis would have substantially increased the volume of the work making it impossible to manage within a single thesis project. Separate initiatives have been launched to this end, including development of the GRADE-CERQual approach for rating the certainty of evidence from qualitative evidence syntheses (Lewin et al., 2015; Noyes, Booth, Flemming, et al., 2017). A recent series of papers published in *Implementation Science* provide detailed guidance for applying GRADE-CERQual domains to qualitative evidence synthesis findings (Lewin et al., 2018). As methods for synthesising evidence of complex interventions expand, future work on rating the certainty of evidence in systematic reviews may need to take a broader view to evidence assessment, such as when integrating different types and sources of evidence in a systematic review (Candy, King, Jones, & Oliver, 2013; Harden et al., 2017; Thomas, O'Mara-Eves, & Brunton, 2014).

The conclusions of thesis Phase 1 research, as reported in Chapter 1, are limited by the focus on a few Cochrane review groups (Movsisyan et al., 2016a, 2016b). Therefore, they may not capture a wide range of perspectives on applying GRADE in

reviews of all types of complex interventions. While this limitation also applies to conclusions drawn from Chapter 4, it should be noted that the Cochrane Collaboration is one of the leading organisations in evidence synthesis and includes a substantive number of reviews of complex interventions that use GRADE. Furthermore, the findings triangulate and corroborate the results from the past investigation conducted by Rehfuess & Akl (2013), which includes perspectives on GRADE use from a range of institutions, such as WHO, NICE, Public Health Agency of Canada, European Centre for Disease Prevention and Control, and Norwegian Knowledge Centre for the Health Services.

Phase 2. Pre-meeting activities: Stakeholder consultations

Main findings

The conclusions drawn from Phase 1 research served to further shape the thesis plans and objectives. Specifically, a need was established for augmenting the existing GRADE approach by providing additional guidance to address the reported challenges of GRADE use in reviews of complex interventions. Phase 2 thesis research, therefore, asked two research questions, specifically, (1) how should the challenges of using GRADE in systematic reviews of complex interventions be addressed? And (2) what domains of evidence should be considered for the GRADE guidance for complex interventions? To address these questions a range of stakeholders were consulted by the DPhil candidate using qualitative semi-structured interviews (Chapter 4) and an online-modified Delphi process (Chapter 5).

Through feedback from the representatives of the GRADE Working Group at the 24th Cochrane Colloquium in Seoul, it was established that a comparison of experiences

of GRADE use by review authors and GRADE methodologists would help to distil the "real" challenges of using GRADE in reviews of complex interventions from those related to lack of knowledge, training and inadequate implementation of GRADE. Findings from the qualitative interviews revealed that many of the reported challenges of GRADE use in reviews of complex interventions (discussed in Chapters 1 and 4) related to how review authors "positioned" application of GRADE in the review process and the time they allocated to the GRADE ratings. While review authors reported implementing GRADE at the end of the review process, adequate implementation of GRADE was conceived by GRADE methodologists to require a thorough consideration from the beginning of the review process. In relation to this, domains of the GRADE approach that were perceived by review authors as particularly challenging to implement in reviews of complex interventions (mostly inconsistency and indirectness) were often seen by GRADE methodologists as influenced by earlier review stages (such as how review questions are framed and important elements are specified in the first place to further inform the structure of the Summary of Findings tables). Lack of team capacity, methodological expertise, time constraints, as well as the spread of the GRADE guidance papers across multiple papers were found to additionally hinder critical application of the GRADE domains by review authors of complex interventions. In the meantime, discontent and uncertainties remained around the initial categorisation of evidence in GRADE and rating of evidence from NRSs.

Using a modified Delphi method, an online expert panel, involving 114 participants, explored areas of agreement and disagreement about the domains of evidence to include in the GRADE guidance for complex interventions (Khodyakov et al., 2011). These domains were derived from the systematic review reported in Chapter 3, as

well as from qualitative interviews with key stakeholders. While participants agreed on all of the 50 domains considered, 32 of these domains were rated as critically important, 17—as important, but not critical and only 1 domain was rated to be of limited importance to consider in the GRADE guidance for complex interventions. The latter related to the suggestion to drop the initial categorisation of evidence based on study design.

Contributions and limitations

As reported in Chapter 2 (thesis methodology), the DPhil candidate has been involved and took a leading role in the running of the project on developing *GRADE Guidance for Complex Interventions* from a very early stage (including drafting of grant applications). In fact, the candidate accomplished Phase 1 of the research prior to the project launch. The successful grant application presented many opportunities for the candidate in terms of providing adequate resources to attend the GRADE Working Group meetings and liaising with important stakeholders and collaborators as the project unfolded. This also meant that issues were identified in time and resolutions sought through team approach. One such issue, for example, included the need for semi-structured interviews with review authors and GRADE methodologists, which was not part of the original project plan. The need for this additional research was, first, identified through feedback from the GRADE Working Group and further discussed within the project team upon completion of Phase 1 of the research.

Collaboration with the GRADE Working Group from the beginning of this research was particularly insightful for the project execution. In the last ten years, the GRADE Working Group has engaged substantively with systematic reviewers and guideline

developers in healthcare aiming to address their needs through advancement of the GRADE guidelines. The feedback received from this group at the 24th Cochrane Colloquium in Seoul was therefore critical in highlighting research priorities, which otherwise might have been overlooked. Specifically, based on their vast experience, members of the group felt that many of the perceived challenges of GRADE use, as reported in Chapter 1, might be due to lack of training and inappropriate implementation of the GRADE domains by review authors rather than the GRADE guidance itself. This served as an impetus for the cross-examination of the experiences of GRADE use by review authors and GRADE methodologists. While the GRADE Working Group itself provides training options through workshops, webinars and online materials, findings from Chapter 4 further support the concern that lack of training and methodological expertise remains an important barrier to critical application of the GRADE approach in reviews of complex interventions.

Engagement with the GRADE Working Group also helped to keep abreast of new projects of relevance for the GRADE guidance for complex interventions. Two recent articles, published by the GRADE Working Group after the launch of this project, discuss issues largely related to the challenges of GRADE use reported in Chapter 1. Specifically, the paper by Hultcrantz et al. (2017) provides a clarification of the GRADE construct of certainty of evidence, and the paper by Schünemann et al. (2017) discusses the initial categorisation of evidence when using ROBINS-I and other tools to assess risk of bias in nonrandomised studies. Both of these aspects of GRADE have been extensively discussed and debated in the context of complex interventions (see Chapters 1, 4 and 6), and in line with the suggestions made at the expert meeting, the write-up of the GRADE guidance for complex interventions should integrate and capitalise on the insights from these

papers (see more detailed discussion of these papers below). It should, however, be noted, that assessing nonrandomised evidence in systematic reviews is a "living" field, as new methods are developed and tested, and subsequent updates of the GRADE guidance will be required to incorporate the novel insights.

While findings from the work in Phase 2 reflects opinions of a large group of stakeholders internationally and across a number of practice domains, several aspects of the online-modified Delphi process are worth noting. The online expert panel served to prioritise domains of evidence and topics for discussion in the subsequent face-to-face expert meeting. However, as a method of stakeholder consultation, it was limited to provide in-depth empirical support for any domain of evidence in terms of the relative importance of the domains to each other, but rather, explored areas of agreement and disagreement around important aspects of the GRADE guidance. Furthermore, the GRADE guidelines do not follow an itemised checklist format, and transformation of the content of the GRADE guidance into a Delphi checklist was somewhat artificial and forced. In addition, as discussed in Chapter 5, the Delphi-based panels are limited to allow participants to make clarifications and resolve misunderstandings. This questions the reliability of deciding the content of the GRADE guidance for complex interventions by the online expert panel results only. It was, therefore, crucial that the panel results were used as general stakeholder advice to guide the discussions of the expert meeting, rather than as hard and fast decisions upon which to base the content of the GRADE guidance. This approach towards using *stakeholder consultation* results from large online panels to inform further *stakeholder collaborations*, such as expert consensus meetings, are increasingly promoted in the field (Khodyakov et al., 2017; Khodyakov et al., 2011; Michie et al., 2013; Moher et al., 2010; Montgomery et al., 2013).

Phase 3. Face-to-face expert meeting

Main findings

Phase 3 of the thesis research aimed to clarify the remit and build consensus around the content and dissemination strategy of the GRADE guidance for complex interventions. Specifically, it sought to answer (1) how should the construct of certainty of evidence be conceptualised in the GRADE guidance for complex interventions? (2) And what domains of evidence should be included in the guidance and how should these be operationalised? To this end, a three-day expert meeting was organised in Oxford in May 2017 by the DPhil candidate and the project co-investigators, where a group of 28 stakeholders from different practice domains and with various professional roles engaged in discussions in light of the evidence from the previous project phases. Participants agreed that "fine-tuning" of the existing GRADE guidance, that is, enhancing the guidance through more detailed explanations of constructs and domains and worked examples on complex interventions would benefit the community of reviewers across social disciplines. However, consensus around more radical changes to the GRADE framework, such as changes to the initial categorisation of evidence based on study design, an alternative conceptualisation of the construct of certainty of evidence, and addition of a new evidence domain on assessing "coherence of the causal pathway" was not fully realised. Furthermore, concerns were raised that substantial deviations from the structure of the GRADE approach, such as extension of the scale of the GRADE ratings (i.e., addition of further categories of ratings to the existing four, including, high, moderate, low, and very low) may not be supported by the GRADE Working Group, which would undermine the uptake of the GRADE guidance for complex interventions.

Collaboration with and official approval from the GRADE Working Group were on the other hand seen as instrumental for widespread dissemination of the GRADE guidance for complex interventions.

Contributions and limitations

The lack of consensus at the expert meeting around important constructs and domains of evidence in rating the certainty of evidence poses challenges for the project team in terms of write-up and dissemination of the new guidance, as further rounds of feedback and revisions will be required to seek resolutions. In this view, while contrary to the adopted strategy for developing research reporting guidelines (Moher et al., 2010), whereby a consensus meeting follows and is informed by a broader Delphi –based process, the conduct of a confirmatory (rather than exploratory) Delphi process after the expert meeting might have proven informative and efficient. In the meantime, the lack of consensus is also indicative of the broader debates and caveats in the field of evidence synthesis, and it is questionable whether resolutions over the raised concerns are actually realistic in the scope of a single project. One such area is the assessment of risk of bias in nonrandomised studies (NRS) in systematic reviews. While assessment of NRSs in systematic reviews can be a challenging task as reported by the methodologists themselves (see Chapters 4 and 6), a key advancement in the recent years has been the development of the ROBINS-I tool for assessing risk of bias in NRSs of interventions (Sterne et al., 2016). In response, the GRADE Working Group has recently updated their guidance for initial categorisation of evidence. In their paper, Schünemann et al. (2018) discuss how to use ROBINS-I to rate the certainty of evidence in GRADE. Since, ROBINS-I includes a more nuanced and comprehensive assessment of selection bias and

confounding, it is proposed to drop the much contested initial categorisation of evidence when using ROBINS-I to inform GRADE certainty of evidence ratings.

A few important challenges with this approach are, however, worth discussing. First, as findings from Chapters 4 and 6 show, use of ROBINS-I requires extensive methodological expertise, which might be beyond the capacities of current systematic review teams. Second, concerns have been raised with the use of ROBINS-I for assessing risk of bias from studies exploiting natural experiments, which are more commonly used for evaluating complex health and social policies (Humphreys, Panter, & Ogilvie, 2017). By comparing an assessment of an individual NRS against a target RCT, ROBINS-I is perceived to set the threshold of methodological acceptability unattainably high for these studies. In the meantime, natural experimental studies are argued to have methodological features different from those of RCTs. For example, blinding and allocation concealment can never be achieved in studies exploiting natural experiments, simply because researchers do not have control over the intervention. This concern also speaks to the points highlighted across stakeholder consultations, specifically, that ROBINS-I is currently designed for cohort-type studies and lacks signalling questions needed to adequately assess risk of bias in quasi-experimental studies, such as interrupted time series. Finally, there is no practical example, where use of ROBINS-I has given a GRADE rating of “high” or “moderate” in a body of evidence from NRSs, where no traditional GRADE upgrading domains applies. This concern was mentioned by a few participants in qualitative interviews (Chapter 4), as well as in the paper by Schünemann et al. (2018). In light of these challenges, it is questionable whether application of ROBINS-I within the GRADE approach is indeed the adequate response to the concern of frequent downgrading of the *“best evidence possible”* for many complex interventions

(see discussion in Chapter 1). Testing the feasibility and acceptability of this approach, as well as ongoing projects to extend ROBINS-I for different types of NRS designs may provide further insights and solutions.

Findings from the systematic review reported in Chapter 4 suggest that conceptualisation of the construct of certainty of evidence is a topic in evidence synthesis that has not been thoroughly examined apart from the initiatives of the GRADE Working Group. The few existing attempts to describe this construct largely extend the definition of the GRADE approach. To remind, in the original GRADE guidelines published in the *Journal of Clinical Epidemiology,* certainty of evidence is defined as *"the extent to which one can be confident that the estimates of effect are correct"* (Guyatt, Oxman, Akl, et al., 2011, p. 394) . In response to the criticism over the lack of nuance and conceptual bias for this definition, the GRADE Working Group has recently published a paper aiming to provide further clarification (Hultcrantz et al., 2017). Specifically, an alternative conceptualisation of certainty of evidence is advocated in the paper, whereby reviewers do not assess their confidence in point estimates of effects, but rather confidence in where effects lie relative to particular thresholds. Implementation of this threshold approach is referred in the GRADE paper as *contextualisation* of certainty of evidence ratings (see Chapter 6 for more details on this approach).

While this threshold approach adds more nuance and transparency to the GRADE ratings of certainty of evidence, findings of Chapter 6 show that challenges remain as to how to set the thresholds, particularly when systematic reviews are conducted without a specific context of application, as in case of most of Cochrane reviews. Specification of thresholds for what may be considered as trivial, small, moderate, and large effects is likely to vary across outcomes, and until now, there is no consensus on any outcome for

complex interventions. Furthermore, there might be cultural differences in reviewers' and decision-makers' accounts of the thresholds across different areas of practice (Knorr-Cetine, 1999). As expert discussions suggest, consideration of the magnitude of effect (i.e., whether effects are large, moderate or small) can be of great importance in the fields of education and criminology, while in broader public health policy areas, decision-makers are perceived to be more interested in the overall direction of intervention effects (i.e., whether there are non-null effects). This further questions whether a single recommendation on the construct of certainty of evidence will be applicable across different domains of practice (see implications below).

It should further be noted that the implications of the way the construct of "certainty of evidence" is defined on other GRADE domains was mentioned, but not thoroughly discussed at the expert meeting, because of the lack of time as it would require further brainstorming and in-depth discussions of each domain in relation to different approaches to conceptualising "certainty of evidence" (e.g., as confidence in the nonnull effect or confidence in a specific magnitude of effect). This further raises the question whether an entirely different strategy towards developing a new evidence rating system for complex interventions, which examines the fundamental concepts in evidence rating, such as the meaning of the construct of "certainty of evidence" in complex interventions would have provided more insights and resolutions. This strategy would have, however, been different from that adopted in this thesis work, which largely aimed to "fine-tune", that is to say, to extend the existing constructs and domains of the GRADE approach in the context of complex interventions. The GRADE approach, itself, has been primarily developed using a pragmatic approach (i.e., working through examples), and does not have an explicit theoretical and conceptual basis. It is partly the

criticism of this lack of conceptual basis that has driven the GRADE Working Group to publish a new paper, post factum, aiming to provide further clarifications for the construct of "certainty of evidence" (Hultcrantz et al., 2017).

This thesis project has aimed to take a transdisciplinary approach towards development of the GRADE guidance for complex interventions by targeting a set of social disciplines. This approach has been inspired by the notion advanced by Gibbons (1994) regarding the ongoing changes in the mode of knowledge production in social sciences. Authors identify and examine a number of attributes suggesting a transformation in the mode of knowledge production in social sciences. In contrast to the traditional knowledge production generated within a single disciplinary context (i.e., Mode 1 knowledge), the knowledge production in the modern societies are increasingly carried out in a context of application and through a transdisciplinary effort (i.e., Mode 2 knowledge). This means that efficient solutions need to be sought beyond a single contributing discipline and through integration of different skills in a framework of action. While a delineated set of social disciplines were targeted from the outset of this thesis work, including public health, social policy, psychology, education, criminology and international development, there are no doubt other disciplines involved in complex intervention research that have not been as adequately represented, such as economics, political science and philosophy of science. Despite recruitment of researchers identifying with more than ten disciplines in the online expert panel, the expert meeting was, perhaps, limited in number.

The online-modified Delphi process and the expert meeting were also limited by primary sampling of participants from the US, UK and Europe, geographically, and those from "public health" than other areas of practice, disciplinarily. While the number and

the diversity of stakeholders consulted throughout research Phases 1 and 2 was large in comparison to other guidance development projects (Akl, Welch, et al., 2017; Michie et al., 2013), stakeholders from outside the US, UK and Europe, as well as those without a training in public health were underrepresented in the thesis research. This can be partly explained by a higher response rate of stakeholders with these demographic profiles in the online expert panel and the expert meeting, as well as the employed purposive sampling; these geographic regions and disciplines are more invested in the GRADE approach and produce larger amounts of systematic reviews and practice guidelines. Nevertheless, further involvement of stakeholders from LMICs and other areas of social practice, such as social policy would have enlarged the breadth of considerations and examples to be included in the GRADE guidance for complex interventions. Thus, directed dissemination activities (e.g., though editorials) will be required to reach these groups when writing up the guidance and for any future updates (see below).

Implications for the write-up and dissemination of the GRADE guidance for complex interventions

While the write-up and dissemination of the GRADE guidance for complex interventions, that is, the Phase 4 in the recommended techniques for developing research reporting guidelines (see Figure 7.1 above), is beyond the scope of this thesis work, implications of this thesis research are described below. These can serve to provide a solid foundation for moving forward with the write-up and dissemination of the new guidance.

Remit of the GRADE guidance for complex interventions

As discussed in detail in Chapter 1, the perspective towards defining complexity in this thesis draws on *sources of complexity* informed by the features of interventions, their causal pathways, and the larger systems in which interventions are implemented. This approach towards viewing complexity as an "interpretative framework" to guide the conduct of systematic reviews through highlighting the important considerations was supported by the stakeholder in the online panel and the meeting. In line with other initiatives on considering complexity in systematic reviews (see Chapter 1), description of sources of complexity to consider when using GRADE would be an appropriate approach for the GRADE guidance for complex interventions. It should be noted, however, that not all sources of complexity will be relevant to every systematic review. In each specific case, researchers should be recommended to take a pragmatic approach that focuses on the key aspects of interventions, their causal pathways, and the systems relevant to the specific aims of the review and users' needs. While a list of the sources of complexity to

consider in systematic reviews has been outlined in the previous chapters of the thesis (Petticrew, Anderson, et al., 2013), there is a growing body of literature which can provide helpful guidance, including, for example, the iCAT_SR tool (Lewin et al., 2017), the Context and Implementation of Complex Interventions (CICI) framework (Pfadenhauer et al., 2017) and the new PRISMA-CI guidance for reporting of systematic reviews and meta-analyses (Guise, Butler, et al., 2017). This literature can provide insights for structuring the GRADE guidance for complex interventions, for example, through mapping of the sources of complexity (see Table 1.7 in Chapter 1) onto reported challenges of the GRADE ratings and specific recommendations and examples to address them.

The adopted complexity perspective corresponds to a variety of disciplines and types of interventions that would fall under the remit of the new GRADE guidance. Examples of these are presented in thesis Introduction (see Table 1 in thesis Introduction). As highlighted in the expert meeting, it is important that the GRADE guidance for complex interventions uses examples of interventions from social disciplines, as all existing GRADE guidelines primarily focus on interventions from biomedicine. Working through "real-life" examples is a practice strongly encouraged by the GRADE Working Group, and, therefore, it will also increase the likelihood of the produced guidance being officially approved. However, in light of the foregoing limitations of the thesis research in terms of the breadth of disciplines considered, there might be a need for additional targeted dissemination activities to enhance the visibility and applicability of the guidance in some social disciplines, such as economics. As suggested at the expert meeting, one such solution could be to publish discipline-oriented annexes on the use of GRADE in specific journals, in addition to publishing the

guidance papers in the *Journal of Clinical Epidemiology*, which is home to all official GRADE publications. While this may require more time and effort for the project team tasked with the write-up and dissemination of the guidance, the discipline-oriented papers may also help to address the differences noted above in the values and practices within "epistemic communities" of reviewers across various disciplines (such as differences in the priorities for thresholds when specifying the construct of certainty of evidence).

Emphases and linkages

Although the procedures for obtaining approval from the GRADE Working Group may require several rounds of revisions and, therefore, more time, participants at the expert meeting stressed that publication of the guidance as an official GRADE paper (i.e., a paper approved by the GRADE Working Group) is likely to increase its visibility and impact. Possible tensions, however, may exist regarding the radical changes to the GRADE approach, such as introduction of new categories of GRADE ratings and domains (see "thinking outside of the GRADE box" in Chapter 6). Thesis findings reveal that the GRADE Working Group might be resistant to approve guidance on the GRADE approach, which suggests considerable changes to the existing GRADE structure and framework. A compromise, therefore, may need to be sought that will enable an augmented guidance, meanwhile, seeking collaboration and approval from the GRADE Working Group. This is in line with the overall non-disruptive strategy adopted in this thesis work, which supports extension of research guidelines through adaptations that enable operating within the existing framework of practice as opposed to developing an entirely new approach (Moher et al., 2010). According to the principles of the evidence-based practice

(Sackett, 2000), actions moving forward should include drafting the GRADE guidance for complex interventions in light of the evidence from this thesis research, and presenting the initial drafts to the GRADE Working Group. Received feedback will then need to be assessed within the project team to inform the course of further actions (such as, to pursue with official GRADE publications or to publish the guidance independently).

Although the key emphasis of this thesis research was on the GRADE certainty of evidence ratings in complex interventions, it is important that these ratings are integrated and contextualised within the broader process of systematic reviewing and guideline development. Figure 7.2 (also shown as Figure 1.3) situates the GRADE ratings within this process. As shown here, GRADE ratings are informed by earlier stages in systematic reviewing, including formulation of questions, retrieval and critical appraisal of studies and evidence synthesis. As suggested by stakeholders, it is important that the GRADE ratings are viewed in connection with other review stages, including how review questions are formulated and important PICO elements specified, so as to further inform the structure of the Summary of Findings (SoFs) tables and the GRADE Evidence Profiles. A holistic approach should therefore be taken in the GRADE guidance for complex interventions, including a section discussing how decisions at earlier review stages influence the design of the SoFs tables and the GRADE certainty of evidence ratings. Incorporation of the sources of complexity into the review, and ultimately, into the GRADE ratings at the earliest stages of the review process should be emphasised and illustrated. The literature reviewed in Chapter 1 on *framing systematic reviews of complex interventions* can provide helpful insights for designing this section of the guidance.

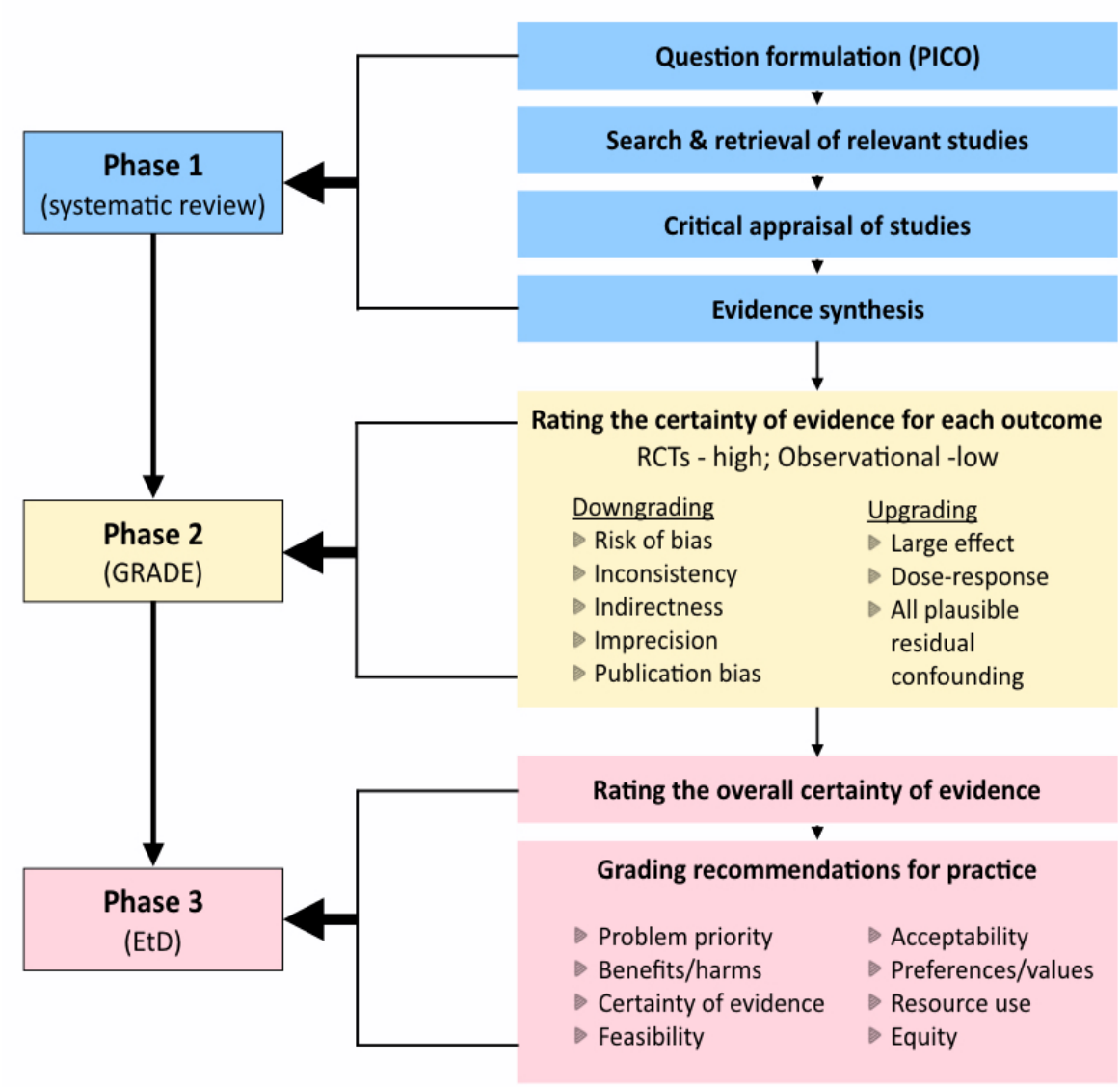


Figure 7.2. The GRADE process for systematic reviewing and developing practice recommendations, adapted from Guyatt, Oxman, Akl, et al. (2011) and Alonso-Coello et al. (2016)

In a similar vein, linkages need to be made in the guidance on how GRADE certainty of evidence ratings may integrate with other considerations to inform recommendations on complex interventions. In recent years, the GRADE Working Group has recognised that—in addition to assessing the certainty in the estimates of effect of interventions—several other factors should be systematically considered before making

practice recommendations. This has led to the development of the Evidence-to-Decision (EtD) frameworks, which delineate important criteria for health decision-making (see Figure 7.2). These criteria include values and preferences (in relation to outcomes), balance of benefits and harms, resource implications, priority of the problem, equity and human rights, acceptability and feasibility (Alonso-Coello et al., 2016). While assessment of the appropriateness of these criteria in the context of complex interventions is beyond the scope of this thesis work, it should be stressed that the certainty in the estimates of effect should always be brought together with other criteria of relevance to inform decisions and practice recommendations on complex interventions (see Table 7.1).

Formatting and publication

As highlighted by expert meeting participants, the recently published GRADE equity guidelines can provide insights into structuring and formatting the GRADE guidance for complex interventions (Welch, Akl, Guyatt, et al., 2017). GRADE equity guidelines include four publications in the *Journal of Clinical Epidemiology.* The first paper introduces the initiative and the methods for considering health equity in the GRADE methodology for developing clinical, public health and health system guidelines (Welch, Akl, Guyatt, et al., 2017). The second paper aims to provide guidance for guideline developers on how to consider health equity at key stages of the guideline development process (Akl, Welch, et al., 2017). The third paper provides guidance on how to assess health equity when rating the certainty of evidence (Welch, Akl, Pottie, et al., 2017). And finally, the fourth paper aims to provide detailed guidance on how to incorporate health equity within the GRADE evidence to decision (EtD) process (Pottie et al., 2017).

Table 7.1. Criteria for Evidence to Decision frameworks for five types of decisions, adapted from Alonso-Coello et al., 2016

	Clinical recommendation Individual perspective	Clinical recommendation Population perspective	Coverage decision	Health system & public health decisions	Diagnostic, screening & other tests
Priority of the problem	Is the problem a priority?				
Test accuracy	Not applicable				How accurate is the test?
Benefits and harms	How substantial are the desirable anticipated effects?				
	How substantial are the undesirable anticipated effects?				
Certainty of evidence	What is the overall certainty of the evidence of effects?				What is the certainty of evidence of test accuracy, any critical direct benefits, harms, burden, effects of the test, link between test results and management decisions?
Outcome importance	Is there important uncertainty about or variability in how much people value the main outcomes?				
Balance	Does the balance between desirable and undesirable effects favour the intervention or the comparison?				Does the balance between desirable and undesirable effects favour the test or the comparison?
Resource use	—	How large are the resource requirements (costs)?			
	—	What is the certainty of evidence of resource requirements (costs)?			
	Does the cost-effectiveness of the intervention favour the intervention (option) or the comparison?				Does the cost effectiveness of the test favour the test or the comparison?
Equity	—			What would be the impact on equity?	
Acceptability	Is the intervention (option) acceptable to key stakeholders?				Is the test acceptable to key stakeholders?
Feasibility	Is the intervention (option) feasible to implement?				Is the test feasible to implement?

Following this approach, the GRADE guidance for complex interventions can be published in three papers (see Box 7.1). The first paper, *GRADE guidelines for complex interventions 1: Rationale and methods*, will introduce the rational and methods for developing the GRADE guidance for complex interventions through a summary of the phases of research described in this thesis work. This will ensure transparent and explicit reporting of the procedures and decisions behind the new guidance. The second paper, *GRADE guidelines for complex interventions 2: Consideration of sources of complexity in systematic reviews and practice guidelines*, will describe a conceptual framework for how to consider sources of complexity in systematic reviews and practice guidelines through consolidation of existing literature on the topic; implications of complexity for systematic reviews and practice guidelines discussed in Chapter 1 can provide a foundation from which to draft this paper. The final paper, *GRADE guidelines for complex interventions 3: Rating the certainty of evidence,* will provide a detailed guidance through examples on how to address complexity when rating the certainty of evidence. Suggestions for the content of this paper are outlined below. It should be noted, that while this publication strategy may provide an extended format and detail, the potential downside is that it further adds many papers for systematic reviewers to consult. It is, therefore, important that the aims of these papers are clearly specified and that the papers follow a structure enabling systematic reviewers to use them (or only sections of them) according to their needs and interests.

Although assessment of the GRADE EtD criteria for complex interventions was beyond the scope of this thesis work, expert meeting discussions show that there is interest within the GRADE Working Group to develop an additional guidance on the use of the EtD frameworks in public health. It might therefore be worth considering

publication of this guidance alongside the three papers outlined above (for example, as the fourth paper in the series of the GRADE guidance for complex interventions). This will, however, depend on whether members of the GRADE Working Group may be found who are willing to collaborate and take a lead on this paper. Finally, as noted above, alongside publication of the three papers as official GRADE guidelines in the *Journal of Clinical Epidemiology,* discipline-oriented editorials and annexes may need to be adapted for publication in specific journals in social sciences as part of targeted outreach.

Box 7.1. Overview of *GRADE guidelines for complex interventions* papers

GRADE guidelines for complex interventions 1: Rationale and methods

Description: This paper will introduce the rationale for developing the GRADE guidance for complex interventions. The methods and findings from each research phase will be summarised, including the systematic review, the qualitative interviews the online-modified Delphi process and the expert meeting.

GRADE guidelines for complex interventions 2: Consideration of sources of complexity in systematic reviews and practice guidelines

Description: This paper will consolidate and cross-reference to available guidance on implications of complexity for different stages of systematic reviewing and guideline development. Consideration of key sources of complexity and GRADE when framing systematic reviews will be discussed and illustrated.

GRADE guidelines for complex interventions 3: Rating the certainty of evidence

Description: Considering the evidence from the thesis research, this paper will provide a detailed guidance to help address complexity when rating the certainty of evidence. Relevant approaches to specifying the construct of certainty of evidence will be outlined and assessment of specific domains will be illustrated through examples of interventions from social disciplines (e.g., public health policy).

GRADE guidelines for complex interventions 4: Evidence to decision process

Description: Considering the interest within the GRADE Working Group to develop a tailored guidance on GRADE EtD criteria for public health, this paper may provide specific guidance on how to use GRADE EtD criteria in complex public health interventions. It should, however, be noted that the write-up of this paper is beyond the scope of this thesis work or the project on developing *GRADE Guidance for Complex Interventions* and will be subject to contribution from other interested members of the GRADE Working Group.

Suggestions for the content of the GRADE guidelines for complex interventions 3: Rating the certainty of evidence

In light of the findings of this thesis research, key recommendations for rating the certainty of evidence in reviews of complex interventions are outlined below (see Table 7.2). While these recommendations can serve as the basis for drafting the third paper in the series of the GRADE guidelines for complex interventions, they have been largely developed by the candidate based on rounds of feedback and revisions from the co-investigators working on the project to develop the *GRADE Guidance for Complex Interventions* (a manuscript adaption of these has been submitted to in *BMJ Global Health*). For the guideline papers, feedback from the larger group of the expert meeting participants will be needed to validate and populate these recommendations with examples of complex interventions from social disciplines.

Defining "certainty of evidence" in reviews of complex interventions

Systematic reviewers need to explicitly define the construct of "certainty of evidence" in a way which speaks to the needs of the intended users of the review. According to the recent revisions by the GRADE Working Group (Hultcrantz et al., 2017), certainty of evidence ratings can be conceptualised in three different ways depending on the purpose of the review (i.e., whether a review is conducted as part of a guideline). In systematic reviews which are not specifically intended to inform a guideline (e.g., Cochrane and Campbell reviews), noncontextualised ratings will be more appropriate. This means that reviewers can conceptualise certainty of evidence as confidence that a nonnull effect is present (i.e., the effect of one intervention is different from another intervention), or alternatively, as confidence that the true effect lies within a given range

(i.e., a 95% confidence or prediction interval). Partially contextualised ratings would also be relevant in these reviews, whereby authors set thresholds based on a priori defined magnitude of effect (e.g., what may be considered as small, moderate, or large effect). In these ratings, the net benefit of an intervention cannot be determined, because of lack of information on other factors associated with the intervention, such as intervention harms, feasibility, and acceptability (Alonso-Coello et al., 2016). By contrast, fully contextualised ratings apply to systematic reviews which are conducted within a guideline development context, and, therefore, may be informed by considerations relevant for decision-making, for example, the feasibility and the costs associated with the intervention. In this case, reviewers should conceptualise "certainty of evidence" as confidence that the effect lies above a threshold which makes the intervention worthwhile to implement (Hultcrantz et al., 2017).

Considering that effects of complex interventions may vary widely across studies because of differences in implementation and contextual factors (Petticrew, Rehfuess, et al., 2013; Rehfuess & Akl, 2013), specification of magnitudes of effect for different outcomes which are important for all potential contexts of application can be a challenging task. In this view, the nonnull effect can serve as the most feasible threshold for rating the certainty in the effects of complex interventions. While reviewers will still need to report the point estimates of effect along with the corresponding confidence intervals, conceptualising and rating the certainty of evidence in terms of the confidence in the nonnull effect can provide useful insights into the general direction of the intervention effect (negative or positive). Since contextualisation of evidence requires a broad range of considerations beyond the evidence on intervention effectiveness (Alonso-Coello et al., 2016), it is appropriate that contextualisation and choice of relevant

thresholds are left to the end-users of evidence in specific contexts. They will be in better position to make context-specific judgments.

Finally, it is worth stressing that conceptualisation of certainty of evidence have further implications on how specific domains of GRADE are assessed and operationalised. For example, using variation in point estimates as an indicator of inconsistency will be only marginally relevant when rating the certainty in the nonnull effect given that point estimates across studies are consistent in direction; however, variation in point estimates would be more relevant when rating the certainty with regard to a specific magnitude of effect. It is, therefore, critical that reviewers provide a transparent account on the adopted approach towards rating the certainty of evidence to inform subsequent judgements on specific GRADE domains.

Initial categorisation of evidence based on study design

One of the most contested aspects of GRADE use in reviews of complex interventions has been the initial categorisation of evidence based on study design, specifically, a body of RCTs being initially rated as “high” certainty, and a body of NRSs as “low” certainty (Movsisyan et al., 2016b; Rehfuess & Akl, 2013). Many researchers in public health and social policy have argued that this categorisation of evidence results in downgrading of the “best evidence possible” for many complex interventions, which are practically impossible to evaluate using RCTs. In turn, this tends to bring about low uptake of these interventions by policymakers. Furthermore, this approach has also been criticised for lack of differentiation among various NRS designs. It has been argued that NRS designs differ in their ability to provide causal inferences about intervention effects,

and some designs are stronger than others for that aim (such as, a cross-sectional study versus a controlled interrupted-time series).

Since the recent publication by the GRADE Working Group on using tools such as ROBINS-I to assess risk of bias in NRSs, reviewers have two options: they may proceed with the original GRADE approach to categorise evidence based on study design, or they may drop this categorisation given they use a rigorous tool, such as ROBINS-I is to assess risk of bias in NRSs (Schünemann et al., 2018). The key advantage of the ROBINS-I tool in comparison to other tools is its thorough and comprehensive assessment of risk of bias through 7 distinct domains, including selection bias and confounding (Sterne et al., 2016). It should, however, be noted that use of this tool has its own challenges (Humphreys et al., 2017): it requires significant human resources and epidemiological expertise and currently is only designed for assessing risk of bias in cohort-type studies. Ongoing research and initiatives into assessing risk of bias for NRS designs other than cohort studies, as well as examples of ROBINS-I application in reviews of complex interventions are likely to result in future updates of the GRADE guidance for complex interventions.

Applying GRADE domains in reviews of complex interventions

Risk of bias: Assessment of performance bias is another commonly reported challenge in systematic reviews of complex interventions, primarily because it is often impossible to blind participants and/or providers (Foxcroft, 2016; Grant, Pedersen, Osilla, Kulesza, & D'Amico, 2016). In fact, human agency and active involvement is integral to the mechanisms whereby interventions in social disciplines are thought to operate (May, 2013). The question is whether the lack of blinding in studies of complex interventions

necessarily introduces a risk of bias that reduces confidence in the estimates of effect. It is important that reviewers differentiate between “lack of blinding” and the potential for “performance bias” caused by the lack of blinding (Grant, Pedersen, Osilla, & al., 2016). It is often the case that lack of blinding is an essential aspect of the intervention, especially when awareness of the intervention is an important aspect of its effectiveness, such as in a traffic safety enforcement campaign. Judgments around lack of blinding should thus be made in tandem with other important considerations, such as blinding of outcome assessors or the nature of the comparator (Schünemann, 2013). By way of illustration, in a systematic review on rehabilitation from chronic obstructive pulmonary disease, authors decided not to downgrade evidence for lack of blinding of intervention providers and recipients, because the procedures to blind outcome assessors were assessed as sufficient to guard against performance bias (McCarthy et al., 2015). Finally, it should be noted that judgments regarding performance bias will also depend on the PICO elements of the review (Schünemann, 2013). For example, subjective outcomes will be at higher risk of bias for lack of blinding as compared to objectively measured outcomes, such as all-cause mortality. Similarly, lack of blinding will be a more significant concern when systematic reviews use “usual care” over an active intervention as their comparator.

The new Cochrane risk of bias tool (RoB 2.0) and ROBINS-I provide a more nuanced assessment of performance bias ("Risk of bias tools for use in systematic reviews," Sterne et al., 2016). Specifically, in the revised RoB 2.0 tool, assessment of performance bias is conducted under the domain of *“bias due to deviations from intended interventions”.* This allows assessment of two different aims of a trial, including (1) assessment of the effect of assignment to the intervention, or (2) assessment of the effect of starting and adhering to the intervention. In the former case (also known as

"treatment offer"), lack of blinding of intervention recipients and providers may not warrant downgrading the certainty of evidence as all deviations from the intended intervention reflect real-world practices. In case of the latter, however, deviations such as poor implementation may lead to risk of bias providing a reason to downgrade the certainty of evidence.

The domain on the *"bias due to deviations from intended interventions"* may also inform assessment of fidelity to implementation. While fidelity to implementation is an important consideration that protects against a Type III error (Basch, Sliepcevich, Gold, Duncan, & Kolbe, 1985), that is, assessing an intervention that has not been adequately implemented, many complex interventions often need to be tailored to specific contexts (e.g., different educational and behaviour change interventions). As in case of assessing performance bias, reviewers should use their judgment regarding the level of differences in the implementation from what might be expected in a real-world practice to decide on whether to downgrade evidence for a potential bias.

Inconsistency: Heterogeneity in the estimates of effect is observed frequently in systematic reviews of complex interventions, primarily due to large variations in the implementation of these intervention in different settings and employed outcome measures (Movsisyan et al., 2016a; Rehfuess & Akl, 2013). Consideration of the potential sources of complexity at the outset of the systematic review, such as when framing the review questions can inform the rating for the GRADE domain of inconsistency at this later stage of the review. By way of illustration, a priori defined sources of complexity (such as contextual differences in implementation) can inform how included studies may be grouped and synthesised in the review. Furthermore, if the considered sources of

complexity explain heterogeneity, separate certainty of evidence ratings should be provided for each of the groupings (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al., 2011; Squires, Valentine, & Grimshaw, 2013; Thomas et al., 2014).

It is worth highlighting that judgments of inconsistency in the magnitude or direction of the estimates of effect should accord with how reviewers specify the construct of "certainty of evidence". For instance, if reviewers decide that the null effect is an important threshold for rating the certainty in the estimates of effect, then assessment of inconsistency in the direction of effect would be appropriate. Variation in point estimates and statistically significant heterogeneity should not warrant downgrading evidence for inconsistency in this case, provided that the estimates of effect across included studies are consistent in direction in relation to the point of no effect (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al., 2011; O'Connor et al., 2003). Multiple criteria for assessing inconsistency outlined in the original GRADE guidelines should be used otherwise, such as when reviewers specify the construct of "certainty of evidence" with respect to whether the average effect lies within the 95% confidence or prediction interval. These criteria include overlap of confidence intervals, degree of variation of the point estimates of the effect with respect to chosen thresholds, I^2 and the p-value for the Q test (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al., 2011).

Imprecision: The chosen thresholds for rating the certainty of evidence should inform judgments on the GRADE domain of imprecision. If reviewers choose the null effect as the important threshold, then downgrading evidence for imprecision may be

warranted only when the confidence or prediction interval includes the null effect. In this case, reviewers should be able to differentiate whether the evidence is imprecise, for example, due to small number of events or participants, or whether the evidence is precise and the intervention does not really have a significant effect (Hultcrantz et al., 2017). The latter conclusion will be valid only when the confidence or prediction interval is narrow enough around the null effect to exclude a "meaningful" effect established a priori (Hultcrantz et al., 2017). Finally, if authors choose to rate the certainty that the average effect lies within a chosen range, such as 95% confidence interval, then imprecision should not be a concern to warrant downgrading the evidence (Hultcrantz et al., 2017).

Indirectness: GRADE suggests to assess indirectness of evidence in relation to the PICO elements in the review question. If collected evidence differs widely in the PICO elements from what is specified in the review, this may warrant downgrading evidence for indirectness. It is important that reviewers make careful judgment informed by understanding of intervention mechanisms regarding the level of difference that may cause concerns of indirectness. It is often rare that the intended populations and interventions are identical to those in included studies, and evidence should be downgraded only when the differences are judged to be large enough to impact the outcomes (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Falck-Ytter, et al., 2011).

A specific challenge in assessment of indirectness of evidence in systematic reviews of complex interventions may relate to completeness of available evidence with respect to the review question. Because of broad questions (i.e., following a "lumping"

approach), the available evidence may not always comprehensively address all the PICO elements (e.g., the review question may ask for evidence in both LMICs and high-income countries, but the evidence may only be available for high-income countries). If important differences are suspected, reviewers will need to downgrade the evidence for indirectness. However, a more tailored approach would be to split the question to be able to provide direct evidence for a subset of conditions (e.g., categorise and rate the certainty of evidence for LMICs and high-income countries separately). Lack of evidence may be reported for the remaining subset of conditions (e.g., LMICs), or reviewers may alternatively make an extrapolation based on available evidence and make judgments regarding the level of indirectness of that evidence. It is highly recommended that the potential factors that may modify intervention effects are specified at the beginning of the review process (including when specifying the PICO elements) to reduce reviewers' analytical flexibility and possible data dredging (Munafo et al., 2017).

Publication bias: Evaluations of complex interventions are often published as reports or working papers. It is, therefore, often critical that reviewers of complex interventions consider multi-component searches, including extensive searches in grey literature and expert contacts (the necessity for this should be judged by reviewers in each specific case depending on the results of preliminary scoping searches). Additional scrutiny of the sponsorship of included studies by any vested industries (such as intervention developers, representatives from industries benefiting from the status quo), as well as "allegiance bias" would be appropriate in judgments regarding publication bias in reviews of complex interventions (Dragioti, Dimoliatis, & Evangelou, 2015; Eisner & Humphreys, 2012).

Upgrading: Reviewers should follow the guidance of the GRADE Working Group to assess upgrading criteria in reviews of complex interventions (Guyatt, Oxman, Sultan, et al., 2011). The revised guidance by Schünemann et al. (2018) should be used to inform evidence upgrading when using ROBINS-I and consequently initially rating evidence from all types of study designs as "high" certainty. Reviewers are encouraged to interpret the existing upgrading criteria in GRADE more broadly in reviews of complex interventions. For example, the upgrading criteria of dose-response and counteracting plausible confounding can be extended to consider intervention implementation issues. Specifically, reviewers may consider upgrading evidence ratings in GRADE when (a) larger effects are observed in studies with better implementation (dose-response effect), or (b) positive results are observed in studies with low fidelity to implementation (i.e., counteracting plausible residual bias or confounding). It should, however, be noted that in many system-level interventions the relationships between the intervention implementation and the outcomes might not be as linear. An example of this is the herd protection from sanitation interventions (Cronin et al., 2017), where the community sanitation coverage needs to reach thresholds in the order of 60% or higher, to optimise health and nutrition gains.

Coherence of the causal pathway: Complex interventions are often described to have long and variable causal pathways. Stakeholder consultations reveal a strong interest in the field for a domain of evidence that looks at the coherence of the causal pathway in reviews of complex interventions. An example of such approach can be found in the evidence rating system used by the US Preventive Services Task Force (USPSTF). Specifically, causal chain diagrams are used to describe different links in the causal

pathway of an intervention and to inform the types of evidence that need to be located and synthesised (Sawaya et al., 2007; Whitlock, Williams, Gold, Smith, & Shipman, 2005). Furthermore, availability of rigorous evidence in different links in the causal pathway of an intervention (e.g., links 4 and 7 in Figure 7.2) may increase the certainty in the estimates of effect of its distal outcomes (e.g., link 5 in Figure 7.2).

Reviewers of complex interventions are encouraged to use logic models in the beginning of the review to visually depict important elements and links in intervention causal pathways (Kneale, Thomas, & Harris, 2015; Rehfuess et al., 2018). It is, however, equally important that these initial logic models are further revisited at pre-defined stages of the review process, and, specifically, at the end of the review. This will enable to assess the overall coherence of the causal pathway in light of the evidence collected and synthesised. This approach may prove particularly insightful for reviews of complex interventions where direct evidence linking the intervention with the distal outcomes is not available (such as link 5 in Figure 7.3). For now, reviewers are recommended to reflect on the coherence of intervention causal pathway outside of the GRADE framework through a narrative description; ongoing and future research may provide further solutions on how causal pathway considerations may be integrated into the GRADE ratings.

Table 7.2. Recommendations for rating the certainty of evidence in systematic reviews of complex interventions (developed in discussion with the project co-investigators; Montgomery et al., forthcoming)

Recommendation	Rationale
Deciding on the scope of the review	
1. Use logic models to develop PICO and review questions	Logic models help in scoping, defining and conducting the review and in making the review relevant to policy and practice.
2. Identify which tools to use to best describe the sources of complexity that users will require	There are several newly developed tools on using a complexity perspective in systematic reviews, including iCAT_SR, the CICI framework, and PRISMA-CI.
3. Using these tools identify contextual and implementation factors and other moderators of effect that may help explain heterogeneity and which will need separate GRADE certainty ratings.	In addition to the standard PICO question, identify in both the intervention and the system in which it is being used all the complexities and interactions that review users will want to know about. • Under intervention complexities, consider all aspects of its implementation, including theory of why and how the intervention is expected to work, the process, the components, implementers, causal pathways, and important process outcomes • Under system complexities, consider context, setting (e.g., individual or population level), and any other independent interventions taking place
Defining certainty of evidence	
Define "certainty" in a manner that matches the needs of the intended users of the review	• Decide among the three approaches to defining certainty of evidence: "noncontextualised", "partially contextualised" and "fully contextualised" • In each case, specify the threshold or ranges used to rate the certainty of evidence • For "noncontextualised" reviews, first consider the utility of using GRADE for the "nonnull" effect, then consider assessing certainty with respect to the effect lying within a given range
Rate the certainty of evidence using GRADE	
1. Initially rate outcomes from any body of	• Consider using Cochrane Risk of Bias (RoB 2.0) tool for randomised controlled trials

evidence as "high" if a rigorous tool is used to assess risk of bias (ROBINS-I). Alternatively, follow the "standard" GRADE approach.	• Consider using ROBINS-I for cohort-type studies
2. Give extra scrutiny to the impact of lack of blinding providers/participants on overall risk of bias for outcomes	If lack of blinding of either participants of providers is unlikely to affect assessment of outcome (such as when using objective outcome measures, e.g., mortality), then do not downgrade evidence for lack of blinding for that outcome.
3. Consider the effect of bias associated with deviation from the intended intervention	• Deviations such as poor implementation and co-interventions in relation to the effect of starting and adhering to an intervention may lead to bias and may be downgraded by 1 level • Do not downgrade if assessing the effect of assignment to the intervention, when deviations do not occur in relation to usual practice and groups remain balanced
4. Consider multiple criteria for judging inconsistency of evidence	• Assessment of heterogeneity should always start off with an appraisal of study heterogeneity, including heterogeneity in PICO elements as well as methodological aspects. • Assessment of heterogeneity should take account of multiple criteria for inconsistency (e.g., I^2 and its p value, overlap of CIs and degree of variation within chosen thresholds) • Consider whether definition of certainty of evidence influences nature of inconsistency assessment (e.g., when effect sizes across all studies are consistently in the same direction outside of the null effect or a given threshold of interest, then downgrading for inconsistency is not warranted despite other measures) • Consider different analytic methods to explain heterogeneity (e.g., subgroup analysis, meta-regression, qualitative comparative analysis)
5. Rate imprecision of evidence with regard to the adopted definition of "certainty"	• Consider whether definition of certainty of evidence influences nature of imprecision assessment: - For "noncontextualised" systematic reviews specifying certainty that the effect lies within estimated confidence or prediction intervals, a GRADE assessment for imprecision can usually be omitted as assessment of precision is dependent on the chosen range - For "partly-contextualised" systematic reviews, if the effect is close to the null, consider

	whether the point estimate would represent a trivial, small, moderate or large effect - For "fully contextualised" systematic reviews, simultaneously consider all important outcomes to determine precision of the effect estimate
6. Examine indirectness of evidence by way of assessing important differences in the evidence base beyond what is expected	• Consider grouping studies, synthesising evidence and rating certainty in the estimates of effect for separate outcomes according to the relevant sources of complexity identified at the start of the review • Consider splitting the questions to answer subset conditions, downgrading only for those with less certain evidence. Do not downgrade for indirectness if observed differences are unlikely to affect the outcome
7. Consider publication bias	• Conduct extensive grey literature searches and expert contacts • Consider sponsorship of studies by any vested industries as well as potential "allegiance bias"
8. Upgrading evidence	• Consider upgrading certainty of evidence in accordance with the existing GRADE guidelines when using the standard approach of initial categorisation of evidence based on study design • Consider the new guidance by Schünemann et al. (2018) for upgrading evidence when evidence from all types of study designs is initially rated as "high" certainty • Consider a broadened interpretation of upgrading criteria for reviews of complex interventions (e.g., implementation issues)
Use logic models to investigate coherence of evidence across the causal pathway	Consider assessing the coherence of evidence across different links in the causal pathway at the end of the review. This judgement should be made outside of the GRADE framework.

Notes: CI = "Confidence Intervals". GRADE = "Grading of Recommendations Assessment, Development and Evaluation". PICO = "Population/Problem, Intervention, Comparison, Outcome".

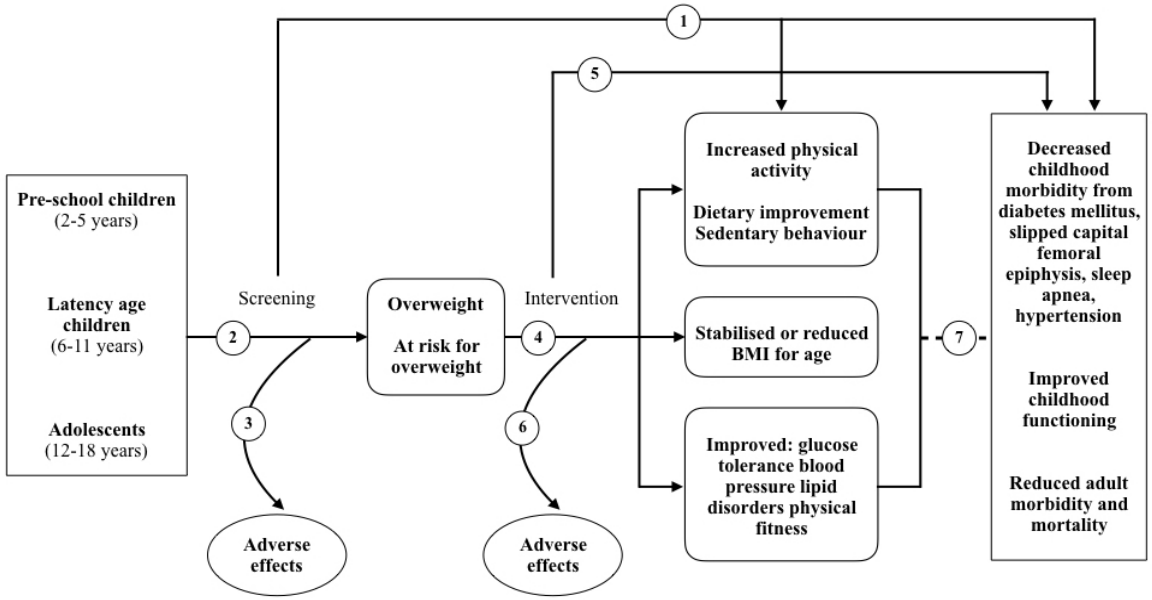


Figure 7.3. A causal pathway approach: screening and interventions for overweight in childhood, adapted from Whitlock et al., 2015

Arrow 1: Is there direct evidence that screening for overweight in childhood improves age-appropriate behavioural measures, or health outcomes?

Arrow 2: **a.** What are appropriate standards for overweight in childhood and what is the prevalence of overweight based on these?
b. What clinical screening tests for overweight in childhood are reliable and valid in predicting obesity in childhood?
c. What clinical screening tests for overweight in childhood are reliable and valid in predicting poor health outcomes in adulthood?

Arrow 3: What are the adverse effects of screening, including labelling? Is screening acceptable to patients?

Arrow 4: **a.** Do weight control interventions lead to improved intermediate outcomes?
b. What are common behavioural and health system elements of efficacious interventions?
c. Are there differences in efficacy between patient subgroups?

Arrow 5: Do weight control interventions lead to improved health outcomes and/or improved functioning?

Arrow 6: What are the adverse effects of interventions? Are interventions acceptable to patients?

Arrow 7: Are improvements in intermediate outcomes associated with improved health outcomes? (Only evaluated if there is no direct evidence for link 1 or link 5 and if there is sufficient evidence for link 4)

Next steps

Write-up of the GRADE guidance for complex interventions

For the project on developing *GRADE Guidance for Complex Interventions,* the important next step is the drafting of the guidance papers. The implications outlined above provide an important starting point. They consolidate findings from the thesis research phases, including the systematic review of the existing evidence rating systems in health and social policy, qualitative interviews with key stakeholders, the online expert panel, and the face-to-face expert meeting. While these implications reflect the amount of attention given to different aspects of the GRADE approach and the resolutions sought within the previous evidence rating systems and in stakeholder consultations, there are several considerations that need to be further discussed and detailed in the guidance (such as, how to address challenges associated with using the ROBINS-I tool and operationalisation of a new evidence domain on "coherence of causal pathway"). It is, therefore, critical that the drafting of the guidance papers follows an iterative process with rounds of feedback from the members of the GRADE Working Group and the expert meeting participants. This will also ensure that the papers integrate the recent insights from the relevant initiatives and subgroups within the GRADE Working Group. This process should also seek to find the best strategy for publishing the guidance that minimises the number of papers (as the large number of publications on GRADE can be a potential barrier to uptake), and, in the meantime, maintains comprehensiveness and the important details.

Further refinements to the suggestions on the content of the GRADE guidance for complex interventions should also be planned in light of piloting the initial drafts of the

guidance to particular disciplines by expert meeting participants. While this process may take additional time and resources, it will help to populate the guidance papers with examples of complex interventions to present to the GRADE Working Group, and, therefore, should be prioritised during the write-up. Through this process, further insights could also be gained on the controversial issues that were not fully resolved at the expert meeting. This would include decisions on the recommendations in the guidance regarding the conceptualisation of the construct of "certainty of evidence" and how it may impact the assessment of the subsequent GRADE domains. In the meantime, navigation through the GRADE Working Group throughout the write-up of the guidance papers can be facilitated by the project team already having sought collaboration and group membership.

As a final note, it is important that the project team tasked with the write-up of the guidance papers consider the ongoing and future developments in the methods of reviewing complex interventions. Many of the issues discussed in this thesis project, such as assessment of risk of bias of nonrandomised studies and approaches to synthesise evidence of complex interventions are still work in development. It is, therefore, critical that the guidance papers are framed accordingly, for example, as "living guidelines" (Akl, Meerpohl, et al., 2017), which will require further updates to incorporate nascent approaches and methods.

Dissemination of the GRADE guidance for complex interventions

Activities to adequately disseminate and implement the GRADE guidance for complex interventions have been emphasised from the beginning of this thesis project. As findings from different phases of the thesis research demonstrate, lack of training and

familiarity, as well as inopportune guidance (such as multiple publications on the GRADE approach) can be an important barrier to implementation of the GRADE approach. In this view, strategies discussed by the expert meeting participants may facilitate knowledge translation. Specifically, dissemination activities outside of healthcare, through involvement of journal editors and publication of "*how to"* short videos and materials, was found important in addition to publishing the official guideline papers. One potential strategy is to develop structured templates or worksheets highlighting important domains and criteria that review authors may need to consider at different stages of reviewing and rating evidence on complex interventions. Prototypes from which to design these worksheets can be found in the resources for authors of the Cochrane Effective Practice and Organisation of Care (EPOC) reviews. The EPOC worksheets for preparing a SoFs table aim to inform on the most important outcomes for each comparison, assess the certainty of evidence for each of those outcomes using GRADE, and prepare a SoFs table for the review ("Cochrane Effective Practice and Organisation of Care (EPOC)," 2017). An extension of these worksheets, which highlight how sources of complexity should be considered by review authors and at what stages of the review, can provide a structured approach for implementing the GRADE ratings. These worksheets can be appended to the guideline papers, as well as made available online as additional resources for review authors.

Future trajectories

This thesis work has important implications for future research, including future methodological projects. While implications of each phase of the thesis research were discussed across the thesis chapters with respect to developing the GRADE guidance for complex interventions, broader implications of the work for future methodological research and research practices are summarised below.

Methodological research

Methods for appraising evidence from NRSs

Stakeholder consultation conducted in this thesis work, including qualitative interviews (Chapter 4) and the expert meeting (Chapter 6), highlight that future research should seek to develop more nuanced approaches for appraising nonrandomised studies (NRS) of interventions. While development of ROBINS-I has been an important advancement in the recent years (Sterne et al., 2016), challenges are identified regarding the significant human resources and epidemiological expertise required to adequately implement this tool. Thus, further research into feasibility and acceptability of this tool will be needed, including when using it to inform the GRADE ratings. Furthermore, as discussed above, the current version of the ROBINS-I tool is designed for cohort-type studies and is not suited for use in appraising more sophisticated quasi-experimental study designs, such as interrupted time series and regression discontinuity. Future extensions of this tool with the addition of signalling questions to assess the specific characteristics of these studies might be needed. Finally, while the use of risk of bias tools are considered a good practice in examining the limitations of evidence within the evidence-based practice paradigm, concerns have also been raised that these tools may

be rigid in their structure to enable a comprehensive evaluation of NRSs, especially those exploiting natural experiments. This is because factors important for the design of these studies are outside of the control of researchers (Humphreys et al., 2017). In this view, calls have been made for a more pragmatic and contextualised approach, which supports evidence appraisal in light of the evidence *availability* and *appropriateness* (Humphreys et al., 2017; Parkhurst & Abeysinghe, 2014). While this may suggest a different perspective to evidence appraisal in comparison to using and adapting established tools and checklists, it might be worth considering this broader approach towards assessment of NRSs considering the acknowledged challenges with the standard methods.

Related to this, another approach towards developing an evidence rating system for use in complex social and public health interventions might be worth considering, which, in addition to describing specific domains of evidence, provides a conceptual and theoretical basis to organise these domains. It should be noted, that even the GRADE approach, which, currently, provides the most comprehensive framework in its guidance and developing, lacks such a basis. This approach towards developing a new guidance for evidence assessment might require novel thinking in terms of questioning the underlying concepts and constructs, such as the meaning of "certainty of evidence", as opposed to a more conservative strategy used in this thesis, which aim to extend the concepts from an existing framework. Stakeholders consulted in this thesis work made suggestions, which could be considered in such a project, but which, however, might be seen too radical or deviating to implement within the existing GRADE framework (e.g., suggestions to have a different scale/categories of evidence ratings or to conceptualise "certainty of evidence" as the extent of confidence in the nonnull effect). Such an initiative, however, would

require much more resources and time for development and dissemination in comparison to the projects, which work from within an existing framework.

Methods for synthesising evidence on the effects of complex interventions

Feedback from the stakeholders also stress the need for further methodological guidance on the synthesis and rating of mixed bodies of evidence (i.e., a body of evidence comprised of different study designs, such as RCTs and NRSs). Complex interventions, especially those targeting system factors, such as population health and policy interventions, are more frequently evaluated using NRSs than RCTs. Systematic reviews of these interventions therefore are likely to include both RCTs and NRSs in the same body evidence informing specific outcomes. Moreover, the arrival of ROBINS-I and the possibility to initially rate evidence from all types of studies as “high” certainty have further raised questions on whether and how to combine results from NRSs and RCTs. These questions, however, remain largely unresolved. Although the current best practice suggests to separate evidence from RCTs and NRSs in systematic reviews, including when constructing the SoFs tables and doing the GRADE ratings (Higgins & Green, 2011; Higgins et al., 2013), there are concerns that this approach may split the evidence too much and, therefore, compromise the best use of the available evidence (Schünemann et al., 2018). Further work both within the GRADE Working Group and in the wider field of evidence synthesis is needed to provide new insights and guidance.

This thesis shows that methods for synthesising complex interventions are evolving. The traditional Cochrane-style systematic reviews which look at the evidence directly linking interventions with the outcomes of interest are being expanded to enable

further understanding of complexities of the system in which interventions operate, such as aspects of context, implementation and causal pathways (Caldwell & Welton, 2016; Candy et al., 2013; Harden et al., 2017; Noyes, Booth, Cargo, et al., 2017; Petticrew, Rehfuess, et al., 2013; Thomas et al., 2014). Thesis findings demonstrate that understanding of the mechanisms of action should be prioritised in systematic reviews of complex interventions, partly because these involve more complex chains and dynamics. Intervention mechanisms are not currently thoroughly addressed in systematic reviews and evidence rating systems (see Chapter 3), and many participants in the online expert panel and the expert meeting noted that the field in general would benefit from future initiatives that seek to describe mechanisms of complex interventions in addition to exploring evidence on intervention effects. This is in line with the arguments to view evidence of mechanisms as complementary (rather than inferior) to evidence of correlation in establishing causal claims (Clark, Gillies, Illari, Russo, & Williamson, 2014; Hill, 1965). There are a few examples on how to explore mechanisms of action in evidence synthesis, such as using model-driven meta-analyses, in which different sources of evidence are integrated in the causal path model similar to a directed acyclic graph (Becker, 2009; Brown, Becker, Garcia, Brown, & Ramirez, 2015). Moving forward, more guidance and examples are needed on the synthesis methods for better understanding of the causal pathways of complex interventions, and investigators in primary research should also explicitly theorise and aim to explore intervention mechanisms in their trials and programme evaluations.

Research practices

Increasing the value of systematic reviews

While systematic reviews aim to inform practice decisions through integration of the best evidence available, their value is largely contingent upon how researchers design, conduct, analyse, and report their primary investigations (Ioannidis et al., 2014). As described in 2014 Lancet series *"Increasing value: reducing waste"*, concerted efforts need to be taken across different stages in the entire pipeline of research production to reduce research waste, including (1) formulation of questions relevant for users of research, (2) use of appropriate research design, conduct and analysis, (3) efficient research regulation and management, (4) production of accessible and full research reports, and (5) unbiased and useable research publications (Macleod et al., 2014). As shown in this thesis work, lack of adherence to scientific standards at any of these stages may further complicate production of unbiased systematic reviews and informative evidence ratings. While this applies to research practices in health and social disciplines in general, issues of adequate reporting and specification of intervention components, contextual and implementation factors is particularly critical in the context of complex interventions, as these may importantly affect the observed effects (Petticrew, Anderson, et al., 2013). Adherence to the recently developed tools, such as the TIDieR checklists, the CONSORT-SPI and the CICI framework can prove useful (Hoffmann et al., 2014; Montgomery et al., 2013; Pfadenhauer et al., 2017).

In addition to enhanced efforts to reduce waste in primary research, protocols of systematic reviews of complex interventions also need to clearly pre-specify review elements, including intervention components of interest, contextual factors and methods

for managing multiple outcome measures and possible multiple analyses (Mayo-Wilson et al., 2017). Outcome definitions in complex interventions are expected to vary extensively across included studies, and lack of specification of outcome elements in systematic reviews (such as outcome domain, measure, metric, measure of aggregation, and time point) can pose genuine challenges to reproducibility and transparent rating of evidence (Munafo et al., 2017).

Review teams and workflow

Findings from qualitative interviews reported in Chapter 4 show that composition of systematic review teams is a significant factor that may influence the efficiency of conduct and methodological rigour of systematic reviews. As discussed by Uttley & Montgomery (2017), composition of review teams has received limited attention in the field of evidence synthesis despite an increasing body of literature showing flaws in the methodology and reporting of systematic reviews (Kirkham, Altman, & Williamson, 2010; Page et al., 2016; Wasiak, Tyack, Ware, Goodwin, & Faggion, 2017). For example, while good practice in systematic reviewing suggests consultation with information specialists, experienced systematic reviewers, statisticians and content experts, no licence is currently required to do a systematic review, and the review team may in fact include none of these specialists (Uttley & Montgomery, 2017). In the meantime, as noted above, the practice of systematic reviewing is becoming more specialised requiring enhanced methodological expertise and human resources (such as for applying the ROBINS-I tool to assess risk of bias in nonrandomised studies of interventions). Therefore, a closer and informed consideration of how review teams are composed is warranted.

In a similar vein, little is known on how the workload is distributed within the review teams. As discussed in Chapter 4, there are concerns that less attention and time is currently given to tasks, which require extensive methodological input and evidence interpretation, including the GRADE assessment. This is likely due to more time being spent within the review teams on title screening and data extraction. Use of new technologies, including machine automation and novel models of human contribution, such as working in dynamic review communities as opposed to siloed review author teams, may offer efficient solutions in the near future (Thomas et al., 2017). In the meantime, a closer attention to who should be conducting systematic reviews and how best to distribute the roles within the team to enable more time and resources for the later (more interpretative) stages of the review process is required.

Evidence to decision process

Finally, efficient translation of research evidence into decision-making contexts remains a significant issue, especially in public health and social policy, where decision-making is a complex process involving multiple perspectives, values and competing interests and priorities (Sanderson, 2009). As highlighted in Chapter 1 (Background to thesis), evidence to decision (EtD) frameworks were beyond the scope of this thesis work and have been advanced by a dedicated group of healthcare researchers within the GRADE Working Group (Alonso-Coello et al., 2016). While the GRADE EtD frameworks aim to ensure that all important criteria are systematically and transparently considered in the process of guideline development, there are concerns that they lack an interdisciplinary perspective and the need to collect and integrate across a range of evidence for broader public health and social policy contexts (e.g., consideration of social

and economic determinants of health) (Cromwell, Peacock, & Mitton, 2015). Further work, therefore, is warranted to test and advance the GRADE EtD frameworks to inform decision-making across different domains of social practice (Rehfuess et al., forthcoming). Moreover, findings from the expert meeting reported in Chapter 6 suggest that further research into the specific needs and training of decision-makers is likely to provide some solutions for optimal communication of research evidence. It is, however, important that an interdisciplinary approach is adopted, which extends the work of the GRADE Working Group to include a range of perspectives across different world regions and areas of practice.

Closing

The recent years have shown increased interest in complex intervention research. This has largely been driven by the acknowledgement of social and health problems as embedded in complex dynamic systems, where solutions need to be sought through concerted efforts that transcend strict disciplinary boundaries. This thesis work argues that evidence-based practices in clinical medicine can provide a useful, however, incomplete model for the areas of social practice. Specifically, while systematic reviews remain an important method to inform practice decisions, they require a wider set of approaches and perspectives for social policy and public health. This is because evidence in these areas of practice is not as homogeneous as in tightly defined clinical settings, and interventions are thought to operate through dynamic mechanisms and interactions with contextual factors (i.e., sources of complexity need to be explicitly considered and addressed). This thesis demonstrates that systematic reviewers and guideline developers of complex interventions are likely to benefit from a more tailored GRADE guidance on how to rate the certainty in the estimates of effect for these interventions. This will serve as a step forward in the transparent and rigorous practice of systematic reviewing across social disciplines, and, ultimately, in better informed policy and practice decisions.

References

Akl, E. A., Meerpohl, J. J., Elliott, J., Kahale, L. A., Schunemann, H. J., & Living Systematic Review, N. (2017). Living systematic reviews: 4. Living guideline recommendations. *J Clin Epidemiol, 91*, 47-53.

Akl, E. A., Welch, V., Pottie, K., Eslava-Schmalbach, J., Darzi, A., Sola, I., . . . Tugwell, P. (2017). GRADE equity guidelines 2: considering health equity in GRADE guideline development: equity extension of the guideline development checklist. *J Clin Epidemiol, 90*, 68-75.

Alonso-Coello, P., Schunemann, H. J., Moberg, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., . . . GRADE Working Group. (2016). GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ, 353*, i2016.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F., & Kolbe, L. J. (1985). Avoiding type III errors in health education program evaluations: a case study. *Health Educ Q, 12*(4), 315-331.

Becker, B. J. (2009). *Model-based meta-analysis. In: Cooper H, Hedges LV, Valentine JC, eds. The Handbook of Research Synthesis and Meta-analysis* (2nd ed*).* New York: Russell Sage Foundation.

Brown, S. A., Becker, B. J., Garcia, A. A., Brown, A., & Ramirez, G. (2015). Model-driven meta-analyses for informing health care: a diabetes meta-analysis as an exemplar. *West J Nurs Res, 37*(4), 517-535.

Caldwell, D. M., & Welton, N. J. (2016). Approaches for synthesising complex mental health interventions in meta-analysis. *Evid Based Ment Health, 19*(1), 16-21.

Candy, B., King, M., Jones, L., & Oliver, S. (2013). Using qualitative evidence on patients' views to help understand variation in effectiveness of complex interventions: a qualitative comparative analysis. *Trials, 14*, 179.

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet, 374*(9683), 86-89.

Clark, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi, 33*(2), 339-360.

Cochrane Effective Practice and Organisation of Care (EPOC). (2017). *EPOC worksheets for preparing a Summary of Findings (SoF) table using GRADE.* EPOC Resources for review authors. Retrieved March 20, 2018 from http://epoc.cochrane.org/resources/epoc-resources-review-authors

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, P. (2008). Developing and evaluating complex interventions: new guidance. Medical Research Council (MRC). Retrieved 20 Jun, 2017, from https://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/

Cromwell, I., Peacock, S. J., & Mitton, C. (2015). 'Real-world' health care priority setting using explicit decision criteria: a systematic review of the literature. *BMC Health Serv Res, 15*, 164.

Cronin, A. A., Gnilo, M. E., Odagiri M., & Wijesekara, S. (2017). Equity implications for sanitation from recent health and nutrition evidence. *Int J Equity Health, 16*, 211.

Dragioti, E., Dimoliatis, I., & Evangelou, E. (2015). Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal. *BMJ Open, 5*(6), e007206.

Eisner M., & Humphreys D. (2012). Measuring conflict of interest in prevention and intervention research - a feasibility study. In Bliesener, T., Beelmann, A., & Stemmler, M. (eds). Antisocial behaviour and crime: contributions of theory and evaluation research to prevention and intervention. Cambridge, Mass: Hogrefe.

Foxcroft, D. R. (2016). We cannot ignore bias, especially if effects are small, but we need better methods for evaluating prevention systems. *Addiction, 111*(9), 1532-1533.

Gibbons, M. (1994). *The new production of knowledge: the dynamics of science and research in contemporary societies*. London: Sage Publications Ltd.

Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. (2017). Retrieved March 15, 2018 from http://gradeworkinggroup.org/

Grant, S., Pedersen, E. R., Osilla, K. C., & al., e. (2016). It is time to develop appropriate tools for assessing minimal clinically important differences, performance bias and quality of evidence in reviews of behavioural interventions. *Addiction, 111*(9), 1533-1535.

Grant, S., Pedersen, E. R., Osilla, K. C., Kulesza, M., & D'Amico, E. J. (2016). Reviewing and interpreting the effects of brief alcohol interventions: comment on a Cochrane review about motivational interviewing for young adults. *Addiction, 111*(9), 1521-1527.

Guise, J. M., Butler, M. E., Chang, C., Viswanathan, M., Pigott, T., Tugwell, P., & Complex Interventions, W. (2017). AHRQ series on complex intervention systematic reviews-paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol, 90*, 43-50.

Guise, J. M., Chang, C., Butler, M., Viswanathan, M., & Tugwell, P. (2017). AHRQ series on complex intervention systematic reviews-paper 1: an introduction to a series of

articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol, 90*, 6-10.

Guyatt, G., H., Oxman, A., D., Akl, E., A., Kunz, R., Vist, G., Brozek, J., . . . Schunemann, H., J. (2011). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol, 64*(4), 383-394.

Guyatt, G., H., Oxman, A., D., Kunz, R., Atkins, D., Brozek, J., Vist, G., . . . Schunemann, H., J. (2011). GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J Clin Epidemiol, 64*(4), 395-400.

Guyatt, G., H., Oxman, A., D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., . . . GRADE Working Group. (2011). GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol, 64*(12), 1303-1310.

Guyatt, G., H., Oxman, A., D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., GRADE Working Group. (2011). GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *J Clin Epidemiol, 64*(12), 1294-1302.

Guyatt, G., H., Oxman, A., D., Sultan, S., Glasziou, P., Akl, E. A., Alonso-Coello, P., . . . GRADE Working Group. (2011). GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol, 64*(12), 1311-1316.

Harden, A., Thomas, J., Cargo, M., Harris, J., Pantoja, T., Flemming, K., . . . Noyes, J. (2017). Cochrane Qualitative and Implementation Methods Group guidance paper 5: methods for integrating qualitative and implementation evidence within intervention effectiveness reviews. *J Clin Epidemiol*. In press.

Harris, K. M., Kneale, D., Lasserson, T., McDonagh, V. M., Grigg, J., & Thomas, J. (2015). School-based self-management interventions for asthma in children and adolescents: a mixed methods systematic review protocol. *Cochrane Database Syst Rev*.

Higgins, J. P., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. Retrieved April 2, 2018 from http://www.handbook.cochrane.org/

Higgins, J. P., Ramsay, C., Reeves, B. C., Deeks, J. J., Shea, B., Valentine, J. C., . . . Wells, G. (2013). Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Res Synth Methods, 4*(1), 12-25.

Hill, A., B. (1965). The Environment and Disease: Association or Causation? *Proc R Soc Med, 58*, 295-300.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., . . . Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*, g1687.

Hultcrantz, M., Rind, D., Akl, E. A., Treweek, S., Mustafa, R. A., Iorio, A., . . . Guyatt, G. (2017). The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol, 87*, 4-13.

Humphreys, D. K., Panter, J., & Ogilvie, D. (2017). Questioning the application of risk of bias tools in appraising evidence from natural experimental studies: critical reflections on Benton et al., IJBNPA 2016. *Int J Behav Nutr Phys Act, 14*(1), 49.

Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . . Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet, 383*(9912), 166-175.

Khodyakov, D., Grant, S., Barber, C. E., Marshall, D. A., Esdaile, J. M., & Lacaille, D. (2017). Acceptability of an online modified Delphi panel approach for developing health services performance measures: results from 3 panels on arthritis research. *J Eval Clin Pract, 23*(2), 354-360.

Khodyakov, D., Hempel, S., Rubenstein, L., Shekelle, P., Foy, R., Salem-Schatz, S., . . . Dalal, S. (2011). Conducting online expert panels: a feasibility and experimental replicability study. *BMC Med Res Methodol, 11*, 174.

Kirkham, J. J., Altman, D. G., & Williamson, P. R. (2010). Bias due to changes in specified outcomes during the systematic review process. *PLoS One, 5*(3), e9810.

Kneale, D., Thomas, J., & Harris, K. (2015). Developing and Optimising the Use of Logic Models in Systematic Reviews: Exploring Practice and Good Practice in the Use of Programme Theory in Reviews. *PLoS One, 10*(11), e0142187.

Knorr-Cetine, K. (1999). *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., . . . Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. *Implement Sci, 13*(Suppl 1), 2.

Lewin, S., Glenton, C., Munthe-Kaas, H., Carlsen, B., Colvin, C. J., Gulmezoglu, M., . . . Rashidian, A. (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 Med, 12*(10), e1001895.

Lewin, S., Hendry, M., Chandler, J., Oxman, A. D., Michie, S., Shepperd, S., . . . Noyes, J. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Med Res Methodol, 17*(1), 76.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., . . . Glasziou, P. (2014). Biomedical research: increasing value, reducing waste. *Lancet, 383*(9912), 101-104.

May, C. (2013). Towards a general theory of implementation. *Implement Sci, 8*, 18.

Mayo-Wilson, E., Fusco, N., Li, T., Hong, H., Canner, J. K., Dickersin, K., & investigators, M. (2017). Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *J Clin Epidemiol, 86*, 39-50.

McCarthy, B., Casey, D., Devane, D., Murphy, K., Murphy, E., & Lacasse, Y. (2015). Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database Syst Rev*(2).

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., . . . Wood, C. E. (2013). The behaviour change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med, 46*(1), 81-95.

Moher, D., Glasziou, P., Chalmers, I., Nasser, M., Bossuyt, P. M., Korevaar, D. A., . . . Boutron, I. (2016). Increasing value and reducing waste in biomedical research: who's listening? *Lancet, 387*(10027), 1573-1586.

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Med, 7*(2), e1000217.

Montgomery, P., Grant, S., Hopewell, S., Macdonald, G., Moher, D., Michie, S., & Mayo-Wilson, E. (2013). Protocol for CONSORT-SPI: an extension for social and psychological interventions. *Implement Sci, 8*, 99.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016a). Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol, 78*, 22-33.

Movsisyan, A., Melendez-Torres, G. J., & Montgomery, P. (2016b). Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol, 70*, 191-199.

Munafo, M., R., Nosek, B., A., Bishop, D., V., M., Button, K., S., Chambers, C., S., Percie du Sert, N., . . . Ioannidis, P., A. (2017). A manifesto for reproducible science. *Nat. Hum. Behav., 1*(0021).

Noyes, J., Booth, A., Cargo, M., Flemming, K., Garside, R., Hannes, K., . . . Thomas, J. (2017). Cochrane Qualitative and Implementation Methods Group guidance series-paper 1: introduction. *J Clin Epidemiol*. In press.

Noyes, J., Booth, A., Flemming, K., Garside, R., Harden, A., Lewin, S., . . . Thomas, J. (2017). Cochrane Qualitative and Implementation Methods Group guidance paper 3: methods for assessing methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings. *J Clin Epidemiol*. In press.

O'Connor, A. M., Stacey, D., Entwistle, V., Llewellyn-Thomas, H., Rovner, D., Holmes-Rovner, M., . . . Jones, J. (2003). Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev*(2).

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., . . . Moher, D. (2016). Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Med, 13*(5), e1002028.

Parkhurst, J., O., & Abeysinghe, S. (2014). What constitutes 'good' evidence for public health and social policy making? From hierarchies to appropriateness. *SERRC, 3*(4), 34-46.

Petticrew, M. (2015). Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev, 4*(36).

Petticrew, M., Anderson, L., Elder, R., Grimshaw, J., Hopkins, D., Hahn, R., . . . Welch, V. (2013). Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol, 66*(11), 1209-1214.

Petticrew, M., Rehfuess, E., Noyes, J., Higgins, J. P., Mayhew, A., Pantoja, T., . . . Sowden, A. (2013). Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol, 66*(11), 1230-1243.

Pfadenhauer, L. M., Gerhardus, A., Mozygemba, K., Lysdahl, K. B., Booth, A., Hofmann, B., . . . Rehfuess, E. (2017). Making sense of complexity in context and implementation: The Context and Implementation of Complex Interventions (CICI) framework. *Implement Sci, 12*(1), 21.

Pottie, K., Welch, V., Morton, R., Akl, E. A., Eslava-Schmalbach, J. H., Katikireddi, V., . . . Alonso-Coello, P. (2017). GRADE equity guidelines 4: considering health equity in GRADE guideline development: evidence to decision process. *J Clin Epidemiol, 90*, 84-91.

Rehfuess, E. A., & Akl, E. A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health, 13*, 9.

Rehfuess, E. A., Booth, A., Brereton, L., Burns, J., Gerhardus, A., Mozygemba, K., . . . Rohwer, A. (2018). Towards a taxonomy of logic models in systematic reviews and health technology assessments: A priori, staged, and iterative approaches. *Res Synth Methods, 9*(1), 13-24.

Risk of bias tools for use in systematic reviews. Retrieved February 20, 2018 from http://www.riskofbias.info/

Sackett, D., L. (2000). *Evidence-based medicine: how to practice and teach EBM*. Edinburgh: Churchill Livingstone.

Sanderson, I. (2009). Intelligent Policy Making for a Complex World: Pragmatism, Evidence and Learning. *Political Stud, 57*(4), 699-719.

Sawaya, G. F., Guirguis-Blake, J., LeFevre, M., Harris, R., Petitti, D., & Force, U. S. P. S. T. (2007). Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med, 147*(12), 871-875.

Schünemann, H. J. (2013). Methodological idiosyncrasies, frameworks and challenges of non-pharmaceutical and non-technical treatment interventions. *Z Evid Fortbild Qual Gesundhwes, 107*(3), 214-220.

Schünemann, H. J., Cuello, C., Akl, E. A., Mustafa, R. A., Meerpohl, J., Thayer, K., . . . Group, G. W. (2018). GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol*. In press.

Squires, J. E., Valentine, J. C., & Grimshaw, J. M. (2013). Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol, 66*(11), 1215-1222.

Sterne, J. A., Hernan, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., . . . Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919.

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., . . . Living Systematic Review, N. (2017). Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol, 91*, 31-37.

Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Syst Rev, 3*, 67.

Uttley, L., & Montgomery, P. (2017). The influence of the team in conducting a systematic review. *Syst Rev, 6*(149).

Wasiak, J., Tyack, Z., Ware, R., Goodwin, N., & Faggion, C. M., Jr. (2017). Poor methodological quality and reporting standards of systematic reviews in burn care management. *Int Wound J, 14*(5), 754-763.

Welch, V. A., Akl, E. A., Guyatt, G., Pottie, K., Eslava-Schmalbach, J., Ansari, M. T., . . . Tugwell, P. (2017). GRADE equity guidelines 1: health equity in guideline development-introduction and rationale. *J Clin Epidemiol, 90*, 59-67.

Welch, V. A., Akl, E. A., Pottie, K., Ansari, M. T., Briel, M., Christensen, R., . . . Tugwell, P. (2017). GRADE equity guidelines 3: considering health equity in GRADE guideline development: rating the certainty of synthesized evidence. *J Clin Epidemiol, 90*, 76-83.

Whitlock, E. P., Williams, S. B., Gold, R., Smith, P., & Shipman, S. (2005) *Screening and Interventions for Childhood Overweight*. Rockville: MD: Agency for Healthcare Research and Quality.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: realist syntheses. *BMC Med, 11*, 21.