

# Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health

Paul Montgomery,<sup>1</sup> Ani Movsisyan,<sup>2</sup> Sean P Grant,<sup>3</sup> Geraldine Macdonald,<sup>4</sup> Eva Annette Rehfuess<sup>5</sup>

**To cite:** Montgomery P, Movsisyan A, Grant SP, *et al*. Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health. *BMJ Glob Health* 2019;**4**:e000848. doi:10.1136/bmjgh-2018-000848

**Handling editor** Seye Abimbola

Received 22 March 2018

Revised 4 July 2018

Accepted 6 July 2018



© World Health Organization 2019. Licensee BMJ.

<sup>1</sup>School of Social Policy, University of Birmingham, Birmingham, UK

<sup>2</sup>Centre for Evidence-Based Intervention, Department of Social Policy and Intervention, University of Oxford, Oxford, UK

<sup>3</sup>Pardee RAND Graduate School, RAND Corporation, Santa Monica, California, USA

<sup>4</sup>School for Policy Studies, University of Bristol, Bristol, UK

<sup>5</sup>Institute for Medical Information Processing, Biometry and Epidemiology, Pettenkofer School of Public Health, Ludwig Maximilian University, Munich, Germany

**Correspondence to**

Dr Paul Montgomery; p.x.montgomery@bham.ac.uk

## ABSTRACT

Public health interventions and health technologies are commonly described as ‘complex’, as they involve multiple interacting components and outcomes, and their effects are largely influenced by contextual interactions and system-level processes. Systematic reviewers and guideline developers evaluating the effects of these complex interventions and technologies report difficulties in using existing methods and frameworks, such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE). As part of a special series of papers on implications of complexity in the WHO guideline development, this paper serves as a primer on how to consider sources of complexity when using the GRADE approach to rate certainty of evidence. Relevant sources of complexity in systematic reviews, health technology assessments and guidelines of public health are outlined and mapped onto the reported difficulties in rating the estimates of the effect of these interventions. Recommendations on how to address these difficulties are further outlined, and the need for an integrated use of GRADE from the beginning of the review or guideline development is emphasised. The content of this paper is informed by the existing GRADE guidance, an ongoing research project on considering sources of complexity when applying the GRADE approach to rate certainty of evidence in systematic reviews and the review authors’ own experiences with using GRADE.

## INTRODUCTION

### Systematic reviews on the effects of interventions in global health

Systematic reviews that estimate the effects of interventions can have a significant influence on subsequent decisions to either implement or disinvest in an intervention. In biomedicine, there are established methods for synthesising and rating certainty in the effects of medications and other single-component interventions.<sup>1</sup> Researchers in public health and health technology assessment (HTA) report difficulties in using these

## Key questions

### What is already known?

- ▶ The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach is an internationally prominent system for rating certainty of the evidence in systematic reviews estimating intervention effects.
- ▶ Researchers conducting systematic reviews on public health and health system interventions report difficulties in using the GRADE approach.

### What are the new findings?

- ▶ Applying a ‘complexity perspective’ can help identify aspects of using the GRADE approach that require particular consideration when rating certainty in the estimates from systematic reviews estimating the effects of global health interventions.
- ▶ These aspects include: sources of complexity when framing the review questions, such as important dimensions of context and implementation and other potential mediators and moderators of effect; a choice of thresholds or ranges for certainty of evidence ratings that matches the needs of intended users of the review; assessment of evidence from non-randomised study designs; the criteria within each GRADE domain for rating certainty of evidence and coherence of evidence across the hypothesised causal pathway of the intervention.

### What do the new findings imply?

- ▶ This primer can help systematic reviewers, health technology assessors and guideline developers better assess evidence relating to complex interventions and systems, which could enhance the use of such evidence in global health policy and practice decisions.

methods,<sup>2 3</sup> largely stemming from failure to account for ‘what interventions work’,<sup>4</sup> for whom, and under what circumstances’,<sup>4</sup> to further inform development of context-specific recommendations.<sup>5 6</sup>

WHO is the leading institution for producing evidence-informed guidelines at a global level. WHO recommendations for practice and policy are underpinned by systematic reviews of evidence on the effects of interventions and health technologies and aim to follow a transparent and evidence-based process.<sup>7</sup> Review and HTAs that inform recommendations often need to consider a range of populations, interventions with different implementation strategies, multiple health and non-health outcomes and various contextual factors that may interact and modify intervention effects.<sup>8</sup> Given the pressing needs and limited resources in low-income and middle-income countries (LMICs), appropriate application of evidence synthesis on a broad range of health interventions is crucial for optimal decisions about implementation.<sup>9</sup>

### Using a complexity perspective in reviews of intervention effects

Recently, there has been increased attention to evaluating effects of complex health interventions implemented within complex systems.<sup>10</sup> This complexity creates challenges for guideline development and HTAs, from the conceptualisation of the questions asked to synthesising diverse types of evidence, assessing or rating the evidence and formulating recommendations. This paper is one of a series exploring the implications of complexity for systematic reviews, HTAs and guideline development.

Using a complexity perspective in systematic reviews of the effects of interventions can facilitate the more nuanced conceptualisation and assessment of interventions ultimately needed for health decision making.<sup>11</sup> An earlier paper in this series differentiates between two main perspectives on sources of complexity in the evaluation of interventions.<sup>10</sup> A 'complex interventions perspective' locates sources of complexity in the features of interventions themselves, such as interventions with different components addressing different and multiple causes of problems.<sup>12</sup> A 'complex systems perspective', on the other hand, locates sources of complexity in the properties of systems into which interventions are introduced, such as how the intervention interacts with, and impacts on, the system as a whole.<sup>13</sup> We use the overarching term 'complexity perspective' to encompass both perspectives and acknowledge the many sources of complexity.<sup>10</sup> Depending on the priority questions of a review, an HTA, or a guideline, either perspective or a combination may be adopted.

In planning and undertaking systematic reviews, a 'complex systems perspective' necessarily entails broadening the scope of a review to include evidence on how the wider system changes when the intervention, such as a complex technology, is introduced. This may involve collecting qualitative evidence on social norms and the dynamics of social networks to describe the broader system impact of the intervention. Not all sources of complexity are relevant to every systematic review and HTA. Researchers should take a pragmatic approach

### Box 1 Examples of guidance and tools for addressing sources of complexity

- ▶ Context and Implementation of Complex Interventions framework<sup>63</sup>
- ▶ Guidance on the integrated assessment of complex health technologies: the INTEGRATE-HTA model<sup>64</sup>
- ▶ Intervention Complexity Assessment Tool for Systematic Reviews<sup>65</sup>
- ▶ Preferred Reporting Items for Complex Interventions for Systematic Reviews and Meta-analyses<sup>66</sup>
- ▶ Template for Intervention Description and Replication tool<sup>6 67</sup>

that focuses on the key aspects of interventions, their causal pathways and the levels of target relevant to the specific aims of the review and users' needs. There is a growing body of literature and guidance, which can be helpful in deciding on the important sources of complexity to consider in systematic reviews, HTAs and guidelines, including approaches described in earlier papers in this series, by Booth *et al*,<sup>14</sup> Petticrew *et al*<sup>10</sup> and Rehfuss *et al*<sup>15</sup> (see box 1).

### The GRADE approach to rating certainty of evidence

The GRADE Working Group has taken a leading role in developing guidance and methods for using research evidence to inform healthcare recommendations. Grading of Recommendations Assessment, Development and Evaluation (GRADE) offers an explicit and transparent system for rating certainty in the body of evidence underpinning conclusions in a systematic review, an HTA or a guideline (box 2). In GRADE, certainty of the effect

### Box 2 Summary of the GRADE process for rating the certainty of evidence for intervention effects

- ▶ The Grading of Recommendations Assessment, Development and Evaluation (GRADE) process starts with an initial certainty rating based on the design of studies included in the body of evidence: if the body of evidence contributing to an outcome consists of randomised controlled trials, certainty is initially rated as 'high', whereas a body of evidence consisting of observational or non-randomised studies (NRSs) is initially rated as 'low'.
- ▶ The assessing team then uses five domains for potential downgrading of the initial certainty rating: study limitations, indirectness, inconsistency, imprecision and publication bias.
- ▶ Next, the team assesses three further domains for potential upgrading of the initial certainty rating: magnitude of the effect, dose-response relationship in the effect and counteracting plausible residual bias or confounding.<sup>43</sup> These upgrading domains are primarily relevant to NRSs (eg, cohort, before-after and interrupted time series).
- ▶ Evidence Profiles and Summary of Findings tables are used to summarise the effect estimates and the certainty ratings for those estimates for each main outcome in the assessment.
- ▶ The GRADE ratings are further used as one of the criteria in the Evidence to Decision frameworks to inform recommendations about implementing interventions in practice, where high-certainty evidence is more likely to result in a strong recommendation compared with low-certainty evidence.<sup>68</sup>

estimate for each outcome is ultimately assigned one of four categories: high, moderate, low or very low. The GRADE approach has been widely adopted by systematic reviewers, authors of HTA and guideline developers in healthcare, including over 100 organisations worldwide.<sup>16</sup> Among these, WHO uses GRADE to inform global health recommendations,<sup>16</sup> and the Cochrane Collaboration mandates use of GRADE in Cochrane intervention reviews.<sup>17</sup>

Despite its wide uptake in biomedicine, systematic reviewers and guideline developers report difficulties applying GRADE in reviews of broader health technologies, health system and public health interventions.<sup>2,3</sup> These challenges are frequently attributed to the complexity of these interventions, often requiring sophisticated consideration and analysis.<sup>3,13</sup> For example, high levels and various sources of heterogeneity in reviews of public health interventions often lead to challenges in deciding how and whether to downgrade for inconsistency.<sup>3,18</sup> Another common challenge results from the difficulty, if not impossibility, of using RCTs to evaluate policy-level and health system interventions.<sup>2,3</sup> With all types of non-RCT evidence starting off at a 'low' certainty level, public and global health researchers have voiced concerns that GRADE may inadvertently produce ratings that steer decision-makers away from implementing important system-level interventions.<sup>19-21</sup> Concerns have also been raised on how to conceptualise the construct of 'certainty' in reviews of global health interventions, and consensus is currently lacking.<sup>3</sup> Consequently, global health researchers could benefit from targeted guidance

on how to rate certainty when encountering these challenges.<sup>2,3,9,21,22</sup>

## Objectives

As part of the overall series intended to stimulate thinking about how methods for reviewing and assessing evidence in guideline development can be enhanced to take account of complexity, this paper clarifies how a complexity perspective may be applied when using the GRADE approach to rate certainty of evidence. The GRADE Working Group is actively working to advance the GRADE methodology for different applications, including for diagnostic tests, prognostic studies and qualitative evidence.<sup>16,23</sup> This paper focuses on using the GRADE approach for rating certainty in the evidence from systematic reviews *estimating the effects* of complex interventions and technologies in global health.

## METHODS

This paper is largely informed by an ongoing, mixed-methods research project, *GRADE Guidance for Complex Interventions*, involving five key studies that follow an established methodology on developing guidance for health research.<sup>24</sup> In Study 1, we investigated GRADE certainty ratings in 24 'complex' and 16 'simple' systematic reviews<sup>18</sup> and obtained feedback from review authors on 19 of these reviews about their process of applying GRADE.<sup>2</sup> In Study 2, we compared domains and criteria across GRADE and 16 other systems for rating certainty of evidence in health and social interventions.<sup>25</sup> In Study 3, we interviewed 10 Cochrane review authors

**Table 1** Mapping the main sources of complexity onto difficulties in rating estimates of the effect of interventions (data taken from Movsisyan *et al*, 2016; Petticrew *et al*, 2013; Petticrew *et al*, 2019; Rehfues and Akl, 2013)<sup>2,3,10,69</sup>

Source of complexity	Difficulties in rating estimates of the effect of interventions
Multiple components	<ul style="list-style-type: none"> <li>▶ Interventions are comprised of different components, which may interact (synergistically or dysynergistically)</li> <li>▶ Need to assess the effects of interventions as bundles or specific intervention components</li> </ul>
Flexibility or tailoring or non-standardisation of implementation	<ul style="list-style-type: none"> <li>▶ Ambiguities around how to assess fidelity to intervention implementation</li> </ul>
Long causal pathways	<ul style="list-style-type: none"> <li>▶ Lack of direct evidence linking interventions with distal outcomes</li> <li>▶ Need to integrate different pieces of evidence from potentially different bodies of evidence to estimate the distal effects</li> </ul>
Effects are contingent on recipients' and providers' agency	<ul style="list-style-type: none"> <li>▶ It may be impossible to blind recipients and providers of interventions</li> </ul>
Multiple outcomes	<ul style="list-style-type: none"> <li>▶ Need to prioritise between a range of important (health and non-health) outcomes</li> </ul>
Effects at different levels, for example, individual and population levels	<ul style="list-style-type: none"> <li>▶ Need to consider outcomes at different levels (eg, individual, family and societal levels)</li> <li>▶ Population-level interventions are frequently impossible to evaluate using RCTs, which results in downgrading the 'best evidence possible' for these interventions because of initial categorisation of evidence in GRADE based on study design</li> </ul>
Moderating effects of context	<ul style="list-style-type: none"> <li>▶ Need to account for various implementation and contextual factors, when conceptualising and rating estimates of the effect</li> </ul>

GRADE, Grading of Recommendations Assessment, Development and Evaluation; RCT, randomised controlled trial.

and 5 GRADE methodologists on their views about the challenges of, and suggestions for, using GRADE in specific systematic reviews incorporating various sources and degrees of intervention and system complexity (Movsisyan *et al*, forthcoming). In Study 4, we conducted an online modified-Delphi process to explore areas of agreement and disagreement among 116 stakeholders about the importance of specific domains and criteria for rating certainty in systematic reviews of complex interventions (Grant *et al*, forthcoming). In Study 5, we held a 3-day consensus meeting to discuss proposals for the content of the new GRADE guidance for complex interventions with 28 stakeholders, purposively invited from the Delphi process, representing a range of subject areas (Movsisyan *et al*, forthcoming). All studies were approved by the Departmental Research Ethics Committee at the Department of Social Policy and Intervention, University of Oxford (SPI\_CIA\_16\_009). This project draws suggestions from several sources—including the existing GRADE guidelines and conceptual papers, previous work considering complexity in systematic reviews, HTAs and guideline development and consultation with relevant stakeholders—to advise how to apply the GRADE approach using a complexity perspective in the context of global health.

## RESULTS

Several aspects of using GRADE require particular consideration when using a complexity perspective in systematic reviews and HTAs on the effects of interventions in global health (see [table 1](#)). Particularly important is that authors consider GRADE from the outset of the review or HTA and not at the end when evidence has already been synthesised. In this way, the totality of the evidence will become an integral part of the assessment from its inception. As systematic reviews represent an important source of evidence and are integral to most HTAs and guidelines in global health, below we describe how specific constructs and domains of GRADE can be used in systematic reviews using a complexity perspective.

### Considering complexity and GRADE when framing the question(s) and conducting the systematic review(s)

The certainty ratings at the final stage of the review are inextricably linked to the purpose and key questions established at the beginning of each complex intervention review. Similar to the approach described in Petticrew *et al*, 2019<sup>10</sup>, reviewers and guideline developers should identify sources and degrees of complexity inherent in interventions themselves, as well as the systems in which they are implemented and intended to influence (see [table 1](#)). As emphasised in the WHO-INTEGRATE framework (see Rehfuess *et al*, 2019<sup>15</sup>), incorporation of sources of complexity into the review and, ultimately, into the GRADE ratings should be considered at the earliest stages of the review process: explicitly addressing sources

### Box 3 Consideration of sources of complexity in a protocol for a Cochrane review of environmental interventions to reduce consumption of sweetened beverages (von Philipsborn *et al*)<sup>27</sup>

von Philipsborn *et al* (2016) developed a system-based logic model taking into account:

- ▶ Beverage choices and diet-related health and non-health outcomes
  - ▶ Physiological and psychological mechanisms linking sugar-sweetened beverages with health outcomes at an individual level
  - ▶ Interventions aimed at policy (macro) and settings (meso) and interpersonal and intrapersonal factors (micro level)
  - ▶ Determinants of diet-related outcomes and related interventions
- This logic model was used to guide data extraction, analysis and interpretation.

Tools used for data collection:

- ▶ Template for Intervention Description and Replication framework to extract and report relevant data related to the intervention<sup>6</sup>
- ▶ Context and Implementation of Complex Interventions framework for contextual data<sup>63</sup>

Internal and external sources of heterogeneity were predefined at multiple levels, such as at policy or setting level:

- ▶ With or without behavioural cointerventions
- ▶ Targeted at: sugar-sweetened beverages, sugar-sweetened milk, beverages with non-nutritive sweeteners or beverages without added sweeteners
- ▶ Implemented in high-income, middle-income or low-income countries
- ▶ Targeted at the general population or at disadvantaged populations

of complexity when formulating review questions<sup>14 15</sup> and structuring the proposed GRADE Evidence Profiles and Summary of Findings (SoFs) tables.<sup>26</sup> Thinking through all relevant sources of heterogeneity at the beginning of the systematic review process will influence the types of data extracted and syntheses conducted. For instance, for a Cochrane review of environmental interventions to reduce the consumption of sugar-sweetened beverages and their adverse effects on health,<sup>27</sup> reviewers developed a system-based logic model to guide data extraction, analysis and interpretation. The frameworks to extract and report relevant data were prespecified as were internal and external sources of heterogeneity for subgroup analysis ([box 3](#)). Such an approach is essential for capturing heterogeneity in the methodology and Population, Intervention, Comparison, Outcome (PICO) elements, which would otherwise remain unexplained and almost inevitably lead to downgrading of evidence.

A frequent challenge for authors of intervention reviews in global health arises from posing broad review questions on bundles of conceptually similar interventions (often referred to as a ‘lumping’<sup>28</sup> or a ‘holistic’<sup>29</sup> approach) that may actually have very different characteristics.<sup>30 31</sup> In these cases, authors should carefully consider upfront which sources of complexity are critical to include in their research questions, such as the active or ‘prototypical’ components of an intervention that are most likely to modify intervention effects.<sup>32</sup>

Logic models may be particularly helpful in depicting intervention components and identifying potential effect modifiers.<sup>10 11 32 33</sup> A recent example involves a systematic review by Welch *et al* (2016) aiming to estimate the effects of deworming interventions on the developmental health and well-being outcomes of children in LMICs.<sup>34</sup> While evidence for deworming programmes had been debated,<sup>35–37</sup> review authors made considerable efforts to describe the complexities of the programme. Specifically, they developed a logic model at the outset of the review to elucidate the entire causal chain from worm infection to nutritional status and educational outcomes, how deworming—in combination with other strategies (such as hygiene promotion and sanitation)—intervenes in the pathway and which factors might be important in mediating or moderating the effects (such as poverty, prevalence and intensity of infection and spill-over effects).<sup>34</sup> This complexity perspective allowed them to add important questions to their overall question of ‘what works’, such as the effects of deworming according to the prevalence of infections, as well as the synergistic effects of cointerventions and treatment externalities for untreated children. They then constructed three separate GRADE SoF tables to provide certainty ratings depending on the levels of endemicity for which different strategies of mass deworming were relevant.<sup>34</sup>

### Defining the thresholds or the ranges for certainty of evidence ratings

The GRADE Working Group conceptualises ‘certainty of evidence’ as confidence that the true effect of an intervention lies on one side of a specified threshold or within a chosen range (see [table 2](#)).<sup>38</sup> In general, depending on the purpose of the assessment (ie, whether the systematic review informs a guideline or not), certainty

of evidence ratings are presented as ‘non-contextualised’, ‘partly contextualised’ and ‘fully contextualised’. Non-contextualised ratings are relevant for assessments conducted outside of a guideline (eg, Cochrane and Campbell reviews): in these circumstances, authors may prioritise the threshold of the null effect and conceptualise certainty of evidence as confidence that a non-null effect is present, that is to say, that the effect of one intervention differs from another. Alternatively, the range approach may be chosen, and certainty of evidence may be conceptualised as confidence that the effect lies within a given range (eg, a 95% CI or prediction interval). Finally, authors may instead choose a partly contextualised rating, setting thresholds of specified magnitudes of effect (eg, what may be considered as a trivial, small, moderate or large effect). Fully contextualised ratings are relevant when systematic reviews are conducted as part of a specific guideline development or decision-making process, which enables integration of other considerations relevant for a health decision. In this case, authors could rate the certainty that the effect lies above a threshold that makes implementation of the intervention worthwhile.<sup>38</sup>

The non-null effect is likely the simplest and most feasible threshold for rating certainty on the effects of public health interventions. Since these intervention effects may vary depending on implementation factors, context and settings, it may be very challenging for a review group to define specific magnitudes of effect for various outcomes that are practically important for *all* potential contexts of application. Rating certainty in the non-null effect would inform the broad global readership about the general direction (positive or negative) of an intervention effect. The task will then be left for

**Table 2** Approaches for setting thresholds or ranges for certainty of evidence ratings (adapted from Hultcrantz *et al*, 2017)<sup>38</sup>

Setting	Contextualisation	Threshold or range	How to set	What certainty rating represents
Primarily for systematic reviews and health technology assessment	Non-contextualised	Range: 95% CI	Using existing limits of the 95% CI, which implies that precision is not routinely part of the rating	Certainty that the effect lies within the CI
		OR≠1; RR≠1; HR≠1; RD≠0	Using the threshold of null effect	Certainty that the effect of one treatment differs from another
Primarily for systematic reviews and health technology assessment	Partly contextualised	Specified magnitude of effect	For example, 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 (eg, trivial, small, moderate or large)
Primarily for 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 for those ranges	Confidence that the direction of the net effect will not differ from one end of the certainty range to the other

RD, risk difference; RR, risk ratio.

the end-users of the evidence at the local level to further contextualise the evidence and set the corresponding thresholds informing implementation of the intervention depending on their specific circumstances. Importantly, contextualising the evidence usually involves a broad range of considerations and decisions not solely driven by evidence of intervention effectiveness.<sup>15</sup>

The choice of the thresholds or ranges will have implications for how the domains of the GRADE approach are applied in a given review or guideline. For example, the criteria of imprecision and inconsistency are only marginally relevant when assessing certainty in the non-null effect (as long as there is consistency in the direction of effect across studies), but are highly relevant when assessing certainty in a specified magnitude of effect. It is therefore critical for systematic reviewers and guideline developers to make their choice explicit and to carry it through into rating the evidence (see [table 3](#) for further details).

### Initial certainty rating based on study design

Following definitions of review scope and thresholds or ranges for rating the certainty of evidence, the initial rating is based on study design. Given the practical impossibilities of conducting RCTs for many complex interventions, global health researchers have struggled with the convention that a body of RCTs is initially rated as ‘high’ certainty in the GRADE approach and a body of NRSs as ‘low’ certainty.<sup>2 3 39</sup> Moreover, there are clearly differences with respect to one’s ability to draw causal inferences about intervention effects from a cross-sectional study, a very weak design for that aim, compared with a much stronger design, such as controlled interrupted-time series; however, both study designs would start off as ‘low’ certainty in GRADE.

Partly in response to these concerns, the GRADE Working Group has recently suggested a second approach, in which any body of evidence may receive a ‘high’ *initial* certainty rating provided that a rigorous tool has been used to assess risk of bias.<sup>40</sup> The new risk of bias in non-randomised studies (ROBINS-I) tool for cohort-type studies is the only tool recognised by the GRADE Working Group as meeting this standard for NRSs.<sup>40 41</sup> Compared with other tools, ROBINS-I offers a rigorous process for assessing risk of bias through seven distinct domains and overall 34 signalling questions, but requires both significant human resources and substantial epidemiological expertise.<sup>41 42</sup> This nuanced assessment of risk of bias in ROBINS-I, specifically, regarding the domains of confounding and selection bias, allows for RCTs and NRSs to be placed on a common metric for risk of bias. While this approach enables one to initially rate evidence from NRS designs as ‘high’ certainty, it is expected that the certainty of evidence will eventually be downgraded for risk of bias, so that the final rating is the same no matter the starting point (ie, whether a body of evidence from NRSs was initially rated as ‘low’ certainty and subsequently rated

up or down or whether it was initially rated as ‘high’ certainty and further downgraded). It is important to note that the current version of ROBINS-I is primarily designed for cohort studies. Although future initiatives may develop extensions to ROBINS-I for other types of NRS designs, following the original GRADE guidance for initial rating of evidence based on study design, the body of evidence comprised of NRSs other than cohort studies should be initially rated as ‘low’ certainty.<sup>40 43</sup> Authors should then explain their decisions for further downgrading or upgrading of the certainty of evidence (eg, subsequent downgrading the certainty of a body of evidence from cross-sectional studies for additional concerns over the risk of bias).

### Applying GRADE domains for rating certainty

We further highlight how reviewers and guideline developers in global health may address sources of complexity when making judgements on specific domains of the GRADE approach.

#### Risk of bias

Handling performance bias in certainty ratings has proven challenging in reviews of interventions, where it is often impossible to blind participants and/or providers.<sup>44 45</sup> A common source of complexity is the contingency of intervention effects on recipients’ and providers’ agency.<sup>46</sup> The challenge therefore is to assess if the lack of blinding introduces a risk of bias that implies reduced confidence in the effect estimates. To do so, review authors should be careful to differentiate between ‘lack of blinding’ and the judgement for the potential of ‘performance bias’ typically associated with the lack of blinding.<sup>47</sup> Lack of blinding does not always cause sufficient bias to warrant downgrading for risk of bias. Indeed, lack of blinding may be an essential aspect of the intervention of interest, particularly when knowledge of the presence of the intervention is an important aspect of its effectiveness as in a traffic safety enforcement campaign. In these circumstances, other considerations become even more important, such as blinding of outcome assessors or the nature of the comparator.<sup>48</sup> For example, for the outcome ‘quality of life’, authors conducting a review on rehabilitation for chronic obstructive pulmonary disease did not downgrade for lack of blinding of provider and participants because they judged the procedures used in included studies to blind outcome assessors sufficient to address any concerns about risk of bias.<sup>49</sup> Furthermore, the potential for bias due to lack of blinding will also depend on the PICO question of the review.<sup>48</sup> Subjective outcomes are more prone to bias than objectively measured outcomes, such as all-cause mortality. Lack of blinding would be a more important source of bias when the comparator in the review is ‘usual care’ than an active intervention, such as when comparing two different educational interventions.

It is worth noting that assessment of performance bias has been revised in the new Cochrane tool to assess

**Table 3** Key considerations for rating certainty in systematic reviews on the effects of complex interventions

Recommendation	Rationale
<b>Deciding on the scope of the review</b>	
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. Approaches have been developed to assist with this <sup>10 33 58 70</sup>
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, such as the approach by Petticrew <i>et al</i> 2019, <sup>10</sup> iCAT_SR, the CICI framework, TIDieR and PRISMA-CI <sup>6 63 65 66</sup>
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 <sup>10</sup> ▶ Under intervention complexities, consider aspects of its implementation, such as theory of why and how the intervention is expected to work, the components, implementers, mediators, moderators, and causal pathways ▶ Under system complexities, consider context, setting and any other independent interventions taking place
<b>Defining thresholds or ranges for certainty of evidence ratings</b>	
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: 'non-contextualised', 'partly contextualised' and 'fully contextualised' <sup>38</sup> ▶ In each case, specify the threshold or ranges used to rate certainty of evidence ▶ For 'non-contextualised' reviews, consider the utility of using GRADE for the 'non-null' effect
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 <sup>10</sup> ▶ 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, moderators, causal pathways (linear and non-linear) and important process outcomes ▶ Under system complexities, consider context, setting (eg, individual or population level) and any other independent interventions taking place
<b>Rating certainty of evidence using GRADE</b>	
1. Initially rate any body of evidence as 'high' if a rigorous tool is used to assess risk of bias in NRSs (ie, ROBINS-I), otherwise, use the 'standard' GRADE guidance	▶ Consider using Cochrane Risk of Bias (RoB V.2.0) tool for randomised controlled trials <sup>42</sup> ▶ Consider using ROBINS-I for cohort-type studies <sup>41</sup>
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 or providers is unlikely to affect assessment of outcome (such as when using objective outcome measures, for example, mortality), then consider not downgrading 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 adherence, poor implementation and cointerventions in relation to the effect of starting and adhering to an intervention, may lead to bias and may be downgraded by one level ▶ Consider not downgrading 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 rather than single criteria for inconsistency (eg, $I^2$ 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 (eg, 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 analytical methods to explain heterogeneity (eg, subgroup analysis, meta-regression and 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 <sup>38</sup> ▶ For 'non-contextualised' systematic reviews definition, a certainty that the effect lies within estimated CIs 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, consider whether the point estimate would represent a trivial, small, moderate or large absolute effect ▶ For 'fully contextualised' systematic reviews, simultaneously consider all important outcomes to determine precision of the effect estimate

Continued

Table 3 Continued

Recommendation	Rationale
6. Examine indirectness of evidence by way of assessing important differences in the evidence base beyond what is expected	<ul style="list-style-type: none"> <li>▶ 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</li> <li>▶ 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</li> </ul>
7. Consider publication bias	<ul style="list-style-type: none"> <li>▶ Conduct extensive grey literature searches and expert contacts to identify reports and working papers</li> <li>▶ Consider sponsorship of studies by any vested industries as well as potential 'allegiance bias'</li> </ul>
8. Upgrading evidence	<ul style="list-style-type: none"> <li>▶ Consider upgrading certainty of evidence for a dose–response relationship related to the level of implementation</li> <li>▶ Consider upgrading evidence for a body of evidence from studies with low implementation fidelity positive results which counteract plausible residual bias or confounding</li> </ul>
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 evidence synthesis. This judgement should be made outside of the GRADE framework

CICI, Context and Implementation of Complex Interventions; GRADE, Grading of Recommendations Assessment, Development and Evaluation; iCAT-SR, Intervention Complexity Assessment Tool for Systematic Reviews; NRS, non-randomised study; PICO, Population, Intervention, Comparison, Outcome; PRISMA-CI, Preferred Reporting Items for Complex Interventions for Systematic Reviews and Meta-analyses; ROBINS-I, risk of bias in non-randomised studies; TIDieR, Template for Intervention Description and Replication.

risk of bias in randomised trials (RoB V.2.0)<sup>42</sup> and in ROBINS-I.<sup>41</sup> In this revised version of the RoB tool, performance bias is assessed under the domain of 'bias due to deviations from intended interventions', which allows for assessing two different aims of the trial: either the effect of assignment to intervention or the effect of starting and adhering to intervention. When interest is in the effect of assignment to intervention (also known as 'treatment offer'), lack of blinding of intervention recipients and providers may not warrant downgrading certainty of evidence, as the deviations from the intended intervention should not create a prognostic imbalance between the trial groups. However, in relation to starting and adhering to the intervention, deviations, such as poor adherence, poor implementation and cointerventions, may lead to risk of bias.<sup>42</sup> Many interventions in global health, such as educational and behaviour change interventions, require tailoring to specific contexts. Accordingly, authors should exercise judgement on the level of differences in intervention implementation that are beyond what would otherwise be expected in a real-world context.

### Inconsistency

Interventions examined in global health reviews often vary in how they are implemented in different contexts and in outcome measures used across settings.<sup>3 18</sup> Consequently, reviews often find considerable heterogeneity in effect estimates. The proper consideration of sources of complexity when framing the review questions can facilitate assessments of whether to downgrade for inconsistency at this later stage of the review. Namely, review authors can group and synthesise the included studies according to the nature of relevant sources of complexity and, if these sources of complexity help explain heterogeneity, provide separate certainty ratings for each of these groupings.<sup>32 50</sup>

Judgements of inconsistency in the magnitude or direction of effects should correspond with the chosen threshold or range that the review team adopts for rating certainty of evidence. For example, if the review team chooses the null effect as a threshold for rating certainty in the estimate of effect, then judging inconsistency in the direction of effect (ie, beneficial or harmful) would be a relevant approach to follow. In this case, variation in point estimates and statistically significant heterogeneity may not warrant downgrading for inconsistency if the effects across studies are consistently in the same direction with respect to the null effect.<sup>50 51</sup> However, if reviewers are rating their certainty in whether the average effect lies within an estimated range, such as within the CIs, then authors should consider multiple criteria for inconsistency (eg, overlap of CIs, degree of variation with respect to chosen thresholds and  $I^2$  and its p value), rather than using only a single statistical measure of heterogeneity.<sup>50</sup>

### Imprecision

Judgements for imprecision are contingent on reviewers' chosen thresholds for rating certainty. If the chosen threshold is the null effect, then imprecision will not be a concern to warrant downgrading evidence unless the confidence or prediction interval includes the null effect, in which case the evidence is either (1) imprecise (due to small number of events or participants) or (2) precise and the intervention does not have an effect relative to the comparator.<sup>38</sup> For the latter, the confidence or prediction interval needs to be sufficiently narrow around the null effect to exclude a 'meaningful' effect established a priori.<sup>38</sup> If clearly stated, authors rating certainty of effects within 95% CIs may omit a precision assessment.<sup>38</sup> This approach, however, has not yet been used in any review and needs further testing on examples using complex health interventions (Montgomery *et al*, forthcoming).

### Indirectness

Many of the reported challenges of judging indirectness can be addressed by specifying appropriate review questions. Important differences in the PICO elements beyond what is expected and specified in the review may weaken inferences regarding the directness of evidence. As outlined in the original GRADE guidance on indirectness, “it is however rare and usually unnecessary, for the intended populations and interventions to be identical to those in the studies, and evidence should be downgraded only if the differences are considered sufficient to make a difference in the outcome likely”.<sup>52</sup>

Another potential challenge relates to incompleteness of available evidence with regard to the review question. As questions for the global health audience might necessarily be broad (ie, often follow a ‘lumping’ approach), the available evidence might not address all elements in the PICO framework (eg, while the question may ask for evidence in both LMICs and high-income countries, the evidence may only be available for high-income countries). If authors suspect major differences in effect across the locations, rather than downgrading all evidence for indirectness, an alternative approach is to split the question to be able to provide direct evidence for a subset of conditions (eg, make separate certainty of evidence ratings for LMICs and high-income countries). In this case, authors may report a lack of evidence for the remaining subset of conditions (eg, LMICs) or extrapolate based on available data. In the latter case, however, reviewers may need to downgrade evidence for indirectness. Again, it is highly recommended that reviewers think about the factors that may modify intervention effects at the beginning of the review process, when scoping the review and formulating specific questions.

### Publication bias

Many evaluations of global health interventions are published as reports, working papers or programme evaluations. If review authors suspect that eligible studies are likely published in this format (rather than in indexed scientific literature), a comprehensive multicomponent search that includes grey literature and contacting of experts is critical. In addition, authors should assess whether a substantial number of studies are sponsored by any vested industries (eg, intervention developers and representatives from industries benefiting from the status quo) or run by researchers with a potential ‘allegiance bias’ to warrant downgrading for publication bias.<sup>53</sup>

### Upgrading

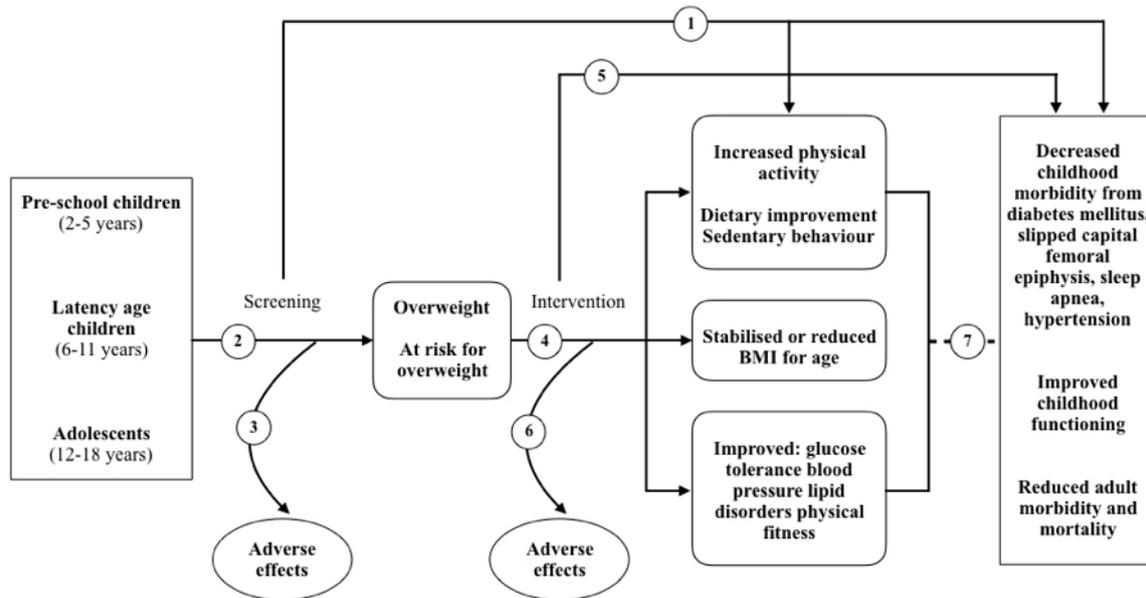
The criteria for *upgrading evidence* of complex interventions should follow the guidance of the GRADE Working Group,<sup>54</sup> including guidance on upgrading when all types of study design are initially rated at ‘high’ certainty.<sup>40</sup> In line with the GRADE guidance, upgrading criteria commonly apply when there are no major limitations in the body of evidence (such as risk of bias, inconsistency

or imprecision).<sup>54</sup> One special case for complex interventions involves intervention fidelity: authors may upgrade their certainty rating if (1) larger effects are found in studies with better implementation (criterion of dose-response effect) or (2) positive results are found among studies with low implementation fidelity (counteracting plausible residual confounding).

### Future work for rating certainty in reviews using a complexity perspective

Based on our project findings, we suggest several areas of future work. Many interventions and complex technologies may have long and variable causal pathways. Our consultation with stakeholders suggests a strong interest in developing a robust domain for an approach to rating certainty that is based on the ‘coherence of the causal pathway’ or ‘chain of evidence’.<sup>55 56</sup> A similar approach is currently used by the US Preventive Services Task Force to describe different links in the causal chain of an intervention and inform what types of evidence should be searched for and synthesised (figure 1).<sup>57</sup> If review authors manage to populate different links in the causal chain of an intervention with rigorous evidence (eg, links 4 and 7 in figure 1), then this may increase their certainty in the effects of its distal outcomes (eg, link 5 in figure 1). As discussed in the previous example on deworming interventions, logic models (also known as analytical frameworks) that visually depict the links in these causal pathways can be useful in identifying the important items of evidence that should be searched for and synthesised in a systematic review.<sup>33</sup> Authors could revisit their initial logic models at predefined stages of the review process, in particular, at the end of the review, using the evidence collected and synthesised for each individual link in the pathway, to assess coherence in the causal pathway originally proposed.<sup>58</sup> This chain of evidence approach might be particularly informative in circumstances where direct evidence linking the intervention with the distal outcomes is unavailable. The work is ongoing both within the GRADE Working Group and beyond, for example, on how to conduct model-driven synthesis of evidence.<sup>59</sup>

Second, several stakeholders are interested in more systematically examining whether there are specific NRS designs that are consistently sufficiently robust to start as ‘moderate’ rather than ‘low’ certainty in the traditional GRADE approach. Such an approach would allow for a ‘quick and dirty’ sorting of study designs, where stronger NRS designs would start off as ‘moderate’ certainty and weaker designs as ‘low’ certainty. Advantages would be a quicker and more user-friendly distinction between different levels of certainty, which is likely to be applicable by many Cochrane and non-Cochrane systematic reviewers; the initial rating could then be refined by risk of bias assessments, but would be less dependent on the use of a very sophisticated risk of bias tool, such as the ROBINS-I tool. The significant challenge, however, relates to which study designs and associated features merit starting in one category versus the other.<sup>60 61</sup> Based



**Figure 1** Example chain of evidence approach: screening and interventions for overweight in childhood. Arrow 1: Is there direct evidence that screening (and intervention) for overweight in childhood improves age-appropriate behavioural or physiological measures or health outcomes? Arrow 2: (1) What are appropriate standards for overweight in childhood, and what is prevalence of overweight based on these? (2) What clinical screening tests for overweight in childhood are reliable and valid in predicting obesity in adulthood? (3) 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: (1) Do weight control interventions lead to improved intermediate outcomes? (2) What are common behavioural and health system elements of efficacious interventions? (3) Are there differences in efficacy between patient subgroups? Arrow 5: Do weight control interventions lead to improved health outcome 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). BMI, body mass index. Taken from Whitlock *et al*, 2005.<sup>57</sup>

on the results of our Delphi process and the consensus meeting, this option was generally supported by those concerned with inconsistent implementation of NRS risk of bias tools leading to overestimation of certainty, especially given the novelty of the tools and varying levels of expertise in using them.<sup>62</sup> However, we do not recommend this approach until future research conclusively establishes such a set of acceptable NRS designs.

**CONCLUSION**

This primer provides a concise discussion of how to incorporate a complexity perspective when applying the GRADE approach in systematic reviews estimating the effects of interventions in global health. Key considerations include: sources of complexity when framing the review questions, such as important dimensions of context and implementation and other potential mediators and moderators of effect; a choice of a threshold or a range that matches the needs of intended users of their review, assessment of evidence from NRS designs and the criteria within each GRADE domain for rating certainty (see table 3). Suggested future work involves investigating the feasibility of (1) a domain on the coherence of evidence across the hypothesised causal pathway of an intervention, which may not need to be integrated into

the GRADE ratings as it will apply at a higher, systematic review level beyond assessment of certainty of evidence in specific outcomes and (2) the identification of specific NRS designs that could start as ‘moderate’ rather than ‘low’ certainty. Researchers, including systematic reviewers, authors of HTA and guideline developers in global health should continue to report their experience using GRADE and this primer in reviews aiming to address sources of complexity. More examples of using GRADE are particularly needed for social interventions, interventions in LMIC contexts, assessments where meta-analysis may not be possible, in bodies of evidence with rigorous NRSs (as well as mixed bodies of evidence), and by researchers with varying degrees of experience in systematic reviewing and evidence assessment.

There is wide interest in finding ways to assess which interventions are effective and, equally, which ones are relevant and appropriate in diverse contexts. Using a complexity perspective can contribute to this. This discussion on considering the ‘complexity perspective’ in GRADE ratings provides a primer for systematic reviewers, authors of HTA and guideline developers to better assess evidence relating to complex interventions and systems, which could ultimately enhance the use of such evidence in global health policy and practice decisions.

**Acknowledgements** The authors would like to acknowledge the contribution of WHO and in particular the members of the group working on this linked series of papers. We also want to thank Elaine Kidney for her editing assistance.

**Contributors** Conceptualisation: all authors. Data curation: AM. Formal analysis: PM, AM, SPG and EAR. Funding acquisition: PM. Investigation: PM, AM and SPG. Methodology: all authors. Project administration: PM. Supervision: PM and EAR. Validation: all authors. Visualisation: PM, AM and SPG. Writing and original draft preparation: PM and AM. Writing, review and editing: all authors.

**Funding** This project was prepared as part of the GRADE Guidance for Complex Interventions, funded by the Economic and Social Research Council (ES/N012267/1). Additional funding was also provided by the World Health Organization Department of Maternal, Newborn, Child and Adolescent Health through grants received from the United States Agency for International Development and the Norwegian Agency for Development Cooperation.

**Competing interests** All authors are current members of the GRADE Working Group. SPG's spouse is a salaried employee of Eli Lilly and Company and owns stock. SPG has accompanied his spouse on company-sponsored travels.

**Patient consent** Not required.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** Data will be made available when the entire study is completed on the University of Birmingham repository.

**Open access** This is an open access article distributed under the terms of the Creative Commons Attribution-Non commercial IGO License (CC BY 3.0 IGO), which permits use, distribution, and reproduction for non-commercial purposes in any medium, provided the original work is properly cited. In any reproduction of this article there should not be any suggestion that WHO or this article endorse any specific organization or products. The use of the WHO logo is not permitted. This notice should be preserved along with the article's original URL.

## REFERENCES

- Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.
- Movsisyan A, Melendez-Torres GJ, Montgomery P. Users identified challenges in applying GRADE to complex interventions and suggested an extension to GRADE. *J Clin Epidemiol* 2016;70:191–9.
- Rehfuess EA, Akl EA. Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health* 2013;13:9.
- Petticrew M. Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Syst Rev* 2015;4:36.
- Koplan JP, Bond TC, Merson MH, *et al.* Towards a common definition of global health. *Lancet* 2009;373:1993–5.
- Hoffmann TC, Glasziou PP, Boutron I, *et al.* Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;348:g1687.
- WHO. *World Health Organization handbook for guideline development*. 2nd ed. Geneva, Switzerland: WHO Press, 2014.
- Doyle J, Waters E, Yach D, *et al.* Global priority setting for Cochrane systematic reviews of health promotion and public health research. *J Epidemiol Community Health* 2005;59:193–7.
- Peters DH, Bennett S. Better guidance is welcome, but without blinders. *PLoS Med* 2012;9:e1001188.
- Petticrew M, Knai C, Thomas J, *et al.* Implications of a complexity perspective for systematic reviews and guideline development in health decision making. *BMJ Glob Health* 2018;0:e000899.
- Anderson LM, Petticrew M, Chandler J, *et al.* Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol* 2013;66:1205–8.
- Craig P, Dieppe P, Macintyre S, *et al.* *Developing and evaluating complex interventions: new guidance: Medical Research Council (MRC)*, 2008.
- Hawe P, Shiell A, Riley T. Theorising interventions as events in systems. *Am J Community Psychol* 2009;43(3-4):267–76.
- Booth A, Moore G, Flemming K, *et al.* Taking account of context in systematic reviews and guidelines considering a complexity perspective. *BMJ Glob Health* 2019;0:e000840. doi:10.1136/bmjgh-2018-000840.
- Rehfuess EA, Stratil JM, Scheel IB, *et al.* The WHO-INTEGRATE evidence to decision framework version 1.0: integrating WHO norms and values and a complexity perspective. *BMJ Glob Health* 2019;4:e000844.
- Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. 2017 <http://gradeworkinggroup.org/>
- Higgins J, Lasserson T, Chandler J, *et al.* *Methodological expectations of Cochrane interventions reviews*. London: Cochrane, 2016.
- Movsisyan A, Melendez-Torres GJ, Montgomery P. Outcomes in systematic reviews of complex interventions never reached "high" GRADE ratings when compared with those of simple interventions. *J Clin Epidemiol* 2016;78:22–33.
- Ogilvie D, Egan M, Hamilton V, *et al.* Systematic reviews of health effects of social interventions: 2. *Best available evidence: how low should you go?* *J Epidemiol Community Health* 2005;59:886–92.
- Parkhurst JO, Abeysinghe S. What constitutes 'good' evidence for public health and social policy making? From hierarchies to appropriateness. *SERRC* 2014;3:34–46.
- Nasser SM, Cooke G, Kranzer K, *et al.* Strength of recommendations in WHO guidelines using GRADE was associated with uptake in national policy. *J Clin Epidemiol* 2015;68:703–7.
- Lewin S, Bosch-Capblanch X, Oliver S, *et al.* Guidance for evidence-informed policies about health systems: assessing how much confidence to place in the research evidence. *PLoS Med* 2012;9:e1001187.
- Lewin S, Booth A, Glenton C, *et al.* Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. *Implement Sci* 2018;13(Suppl 1):2.
- Moher D, Schulz KF, Simeria I, *et al.* Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7:e1000217.
- Movsisyan A, Dennis J, Rehfuess E, *et al.* 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* 2018;9:224–42.
- Murad MH, Mustafa RA, Schünemann HJ, *et al.* Rating the certainty in evidence in the absence of a single estimate of effect. *Evid Based Med* 2017;22:85–7.
- von Philipsborn P, Stratil JM, Burns J, *et al.* Environmental interventions to reduce the consumption of sugar-sweetened beverages and their effects on health. *Cochrane Database Syst Rev* 2016;135.
- Caldwell DM, Welton NJ. Approaches for synthesising complex mental health interventions in meta-analysis. *Evid Based Ment Health* 2016;19:16–21.
- Guise JM, Chang C, Viswanathan M, *et al.* Agency for Healthcare Research and Quality Evidence-based Practice Center methods for systematically reviewing complex multicomponent health care interventions. *J Clin Epidemiol* 2014;67:1181–91.
- Turley R, Saith R, Bhan N, *et al.* Slum upgrading strategies involving physical environment and infrastructure interventions and their effects on health and socio-economic outcomes. *Cochrane Database Syst Rev* 2013;1:CD010067.
- Weir MC, Grimshaw JM, Mayhew A, *et al.* Decisions about lumping vs. splitting of the scope of systematic reviews of complex interventions are not well justified: a case study in systematic reviews of health care professional reminders. *J Clin Epidemiol* 2012;65:756–63.
- Squires JE, Valentine JC, Grimshaw JM. Systematic reviews of complex interventions: framing the review question. *J Clin Epidemiol* 2013;66:1215–22.
- Rohwer A, Pfadenhauer L, Burns J, *et al.* 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* 2017;83:37–47.
- Welch VA, Ghogomu E, Hossain A, *et al.* Deworming and adjuvant interventions for improving the developmental health and well-being of children in low- and middle-income countries: a systematic review and network meta-analysis. *Campbell Syst Rev* 2016;7.
- Hilton J, Tovey D. Debating the evidence for deworming programmes. *Cochrane Database Syst Rev* 2012;8:ED000045.
- Andrews JR, Bogoch II, Utzinger J. The benefits of mass deworming on health outcomes: new evidence synthesis, the debate persists. *Lancet Glob Health* 2017;5:e4–e5.
- Taylor-Robinson DC, Jones AP, Garner P. Deworming drugs for treating soil-transmitted intestinal worms in children: effects on growth and school performance. *Cochrane Database Syst Rev* 2007;4:CD000371.
- Hultcrantz M, Rind D, Akl EA, *et al.* The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;87:4–13.
- Guyatt GH, Oxman AD, Vist G, *et al.* GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol* 2011;64:407–15.

40. Schünemann HJ, Cuello C, Akl EA, *et al.* 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* 2018. doi: 10.1016/j.jclinepi.2018.01.012. [Epub ahead of print].
41. Sterne JA, Hernán MA, Reeves BC, *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.
42. Risk of bias tools. <http://www.riskofbias.info/>.
43. Balshem H, Helfand M, Schünemann HJ, *et al.* GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6.
44. Grant S, Pedersen ER, Osilla KC, *et al.* Reviewing and interpreting the effects of brief alcohol interventions: comment on a Cochrane review about motivational interviewing for young adults. *Addiction* 2016;111:1521–7.
45. Foxcroft DR. We cannot ignore bias, especially if effects are small, but we need better methods for evaluating prevention systems. *Addiction* 2016;111:1532–3.
46. May C. Towards a general theory of implementation. *Implement Sci* 2013;8:18.
47. Grant S, Pedersen ER, Osilla KC, *et al.* It is time to develop appropriate tools for assessing minimal clinically important differences, performance bias and quality of evidence in reviews of behavioral interventions. *Addiction* 2016;111:1533–5.
48. Schünemann HJ. Methodological idiosyncracies, frameworks and challenges of non-pharmaceutical and non-technical treatment interventions. *Z Evid Fortbild Qual Gesundheitswes* 2013;107:214–20.
49. McCarthy B, Casey D, Devane D, *et al.* Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database Syst Rev* 2015;2:CD003793.
50. Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 7. Rating the quality of evidence— inconsistency. *J Clin Epidemiol* 2011;64:1294–302.
51. O'Connor AM, Stacey D, Entwistle V, *et al.* Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2003;2:CD001431.
52. Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol* 2011;64:1303–10.
53. Dragiotis E, Dimoliatis I, Evangelou E. Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal. *BMJ Open* 2015;5:e007206.
54. Guyatt GH, Oxman AD, Sultan S, *et al.* GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol* 2011;64:1311–6.
55. Sawaya GF, Guirguis-Blake J, LeFevre M, *et al.* Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med* 2007;147:871–5.
56. Harris RP, Helfand M, Woolf SH, *et al.* Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med* 2001;20(3 Suppl):21–35.
57. Whitlock EP, Williams SB, Gold R, *et al.* Screening and Interventions for Childhood Overweight. *Rockville (MD): U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews* 2005.
58. Rehfuess EA, Booth A, Brereton L, *et al.* Towards a taxonomy of logic models in systematic reviews and health technology assessments: A priori, staged, and iterative approaches. *Res Synth Methods* 2018;9:13–24.
59. Higgins JPT, López-López JA, Becker BJ, *et al.* Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Glob Health*;2019;0:e000858. doi:10.1136/bmjgh-2018-000858.
60. Polus S, Pieper D, Burns J, *et al.* Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in Cochrane reviews. *J Clin Epidemiol* 2017;91:56–69.
61. Higgins JP, Ramsay C, Reeves BC, *et al.* 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* 2013;4:12–25.
62. Voss PH, Rehfuess EA. 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* 2013;67:98–104.
63. Pfadenhauer LM, Gerhardus A, Mozygemba K, *et al.* Making sense of complexity in context and implementation: the Context and Implementation of Complex Interventions (CICI) framework. *Implement Sci* 2017;12:21.
64. Wahlster P, Brereton L, Burns J, *et al.* Guidance on the integrated assessment of complex health technologies - the INTEGRATE-HTA Model. 2016 <http://www.integrate-hta.eu/downloads/>
65. Lewin S, Hendry M, Chandler J, *et al.* Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT\_SR). *BMC Med Res Methodol* 2017;17:76.
66. Guise JM, Butler ME, Chang C, *et al.* AHRQ series on complex intervention systematic reviews—paper 6: PRISMA-CI extension statement and checklist. *J Clin Epidemiol* 2017;90:43–50.
67. Campbell M, Katikireddi SV, Hoffmann T, *et al.* TIDieR-PHP: a reporting guideline for population health and policy interventions. *BMJ* 2018;361:k1079.
68. Alonso-Coello P, Schünemann HJ, Moberg J, *et al.* GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 2016;353:i2016.
69. Petticrew M, Anderson L, Elder R, *et al.* Complex interventions and their implications for systematic reviews: a pragmatic approach. *J Clin Epidemiol* 2013;66:1209–14.
70. Kneale D, Thomas J, Harris K. 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* 2015;10:e0142187.