Appendices to

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

Ani Movsisyan

Department of Social Policy and Intervention
University of Oxford
Wolfson College

Hilary Term, 2018

Table of Contents

Appendix 1. Search strategy

Scientific databases

Search conducted on June 2, 2016

1. MEDLINE (Ovid Interface, 1946 to May Week 4, 2016)

Search Number	Search String	Results
#1	exp Research/	515677
#2	exp Research Design/	371918
#3	exp Guideline/	27792
#4	exp Epidemiologic Studies/	1909422
#5	exp Study Characteristics/	4473942
#6	Feasibility Studies/	50925
#7	Program Evaluation/	51432
#8	Evidence Based Medicine/	61064
#9	Human Experimentation/	10229
#10	Meta Analysis/	66825
#11	(systematic adj3 review$).mp.	70050
#12	(meta-analys#s).mp.	101003
#13	(evidence adj3 synthes#s).mp.	3379
#14	or/#1 – #13	6395228
#15	((approach or approaches or standard*1 or system*1 or instrument*1 or framework*1 or level*1 or hierarchy or hierarchies or method* or profile*1 or guideline*1 or guide*1) adj3 evidence).ti.[1]	4498
#16	(quality or confidence or certainty or strength).tw.	1028928
#17	((rate or rates or rating or grade or grades or grading or measure or measuring or score or scoring or assess or assessing or evaluate or evaluating or tool*1 or checklist*1 or system*1 or instrument*1) adj3 ((level*1 or hierarchy or hierarchies) adj3 evidence)).tw.	506
#18	#15 and #16	827

[1]The search was limited to titles, because text word search retrieved 52385 publications

#19	#16 and #17	264
#20	((grade or grades or grading or rate or rates or rating or evaluate or evaluating or evaluation or assess*) adj3 evidence).ti.	1528
#21	((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) adj3 recommendation*1).ti.	307
#22	((framework*1 or system*1 or instrument*1) adj2 evidence).ti.	419
#23	#18 or #19 or #20 or #21 or #22	3199
#24	((grade or grades or grading or rate or rates or rating or assess* or evaluate or evaluating or appraisal or appraising) adj2 (quality or strength) adj5 (evidence or recommendation*1)).ti.	108
#25	((method or methods or methodology or approach or approaches) adj3 (assess or assessing or assessment or evaluate or evaluating) adj3 (recommendation*1 or evidence)).ti.	43
#26	(good adj3 practi#e*1 adj3 recommendation*1).tw.	229
#27	(#14 and #23) or #24 or #25 or 26	2385
#28	Limit #27 to English, Human and Publication year (1995 to Current)	1923

2. PsycINFO (Ovid Interface, 1987 to May Week 4, 2016)

Search Number	Search String	Results
#1	exp Meta Analysis/	3534
#2	exp Methodology/	79331
#3	Treatment Effectiveness Evaluation/	17621
#4	exp Experimentation/	58524
#5	Evidence Based Practice/	13410
#6	exp Experimental Design/	41170
#7	Clinical Trial/	9520
#8	or/#1 – #7	183003
#9	((approach or approaches or standard*1 or system*1 or instrument*1 or framework*1 or level*1 or hierarchy or	1415

	hierarchies or method* or profile*1 or guideline*1 or guide*1) adj3 evidence).ti.	
#10	(quality or confidence or certainty or strength).tw.	252400
#11	((rate or rates or rating or grade or grades or grading or measure or measuring or score or scoring or assess or assessing or evaluate or evaluating or tool*1 or checklist*1 or system*1 or instrument*1) adj3 ((level*1 or hierarchy or hierarchies) adj3 evidence)).tw.	101
#12	#9 and #10	230
#13	#10 and #11	47
#14	((grade or grades or grading or rate or rates or rating or evaluate or evaluating or evaluation or assess*) adj3 evidence).ti.	666
#15	((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) adj3 recommendation*1).ti.	44
#16	((framework*1 or system*1 or instrument*1) adj2 evidence).ti.	172
#17	#12 or #13 or #14 or #15 or #16	1123
#18	((grade or grades or grading or rate or rates or rating or assess* or evaluate or evaluating or appraisal or appraising) adj2 (quality or strength) adj5 (evidence or recommendation*1)).ti.	16
#19	((method or methods or methodology or approach or approaches) adj3 (assess or assessing or assessment or evaluate or evaluating) adj3 (recommendation*1 or evidence)).ti.	19
#20	(good adj3 practi#e*1 adj3 recommendation*1).tw.	72
#21	(#8 and #17) or #18 or #19 or #20	515
#22	Limit #21 to English, Human and Publication year (1995 to Current)	479

3. EMBASE (Ovid Interface, 1988 to 2016 Week 22)

Search Number	Search String	Results
#1	exp Research/	582597
#2	exp Meta Analysis/	109010
#3	exp Systematic Review/	106868

#4	Epidemiology/	106658
#5	Randomized Controlled Trial/	387363
#6	Evidence Based Practice/	43451
#7	Practice Guideline/	275322
#8	Intervention/	83
#9	Clinical Trial/	812713
#10	or/#1–#9	1955325
#11	((approach or approaches or standard*1 or system*1 or instrument*1 or framework*1 or level*1 or hierarchy or hierarchies or method* or profile*1 or guideline*1 or guide*1) adj3 evidence).ti.	5922
#12	(quality or confidence or certainty or strength).tw.	1427287
#13	((rate or rates or rating or grade or grades or grading or measure or measuring or score or scoring or assess or assessing or evaluate or evaluating or tool*1 or checklist*1 or system*1 or instrument*1) adj3 ((level*1 or hierarchy or hierarchies) adj3 evidence)).tw.	1036
#14	#11 and #12	1192
#15	#12 and #13	467
#16	((grade or grades or grading or rate or rates or rating or evaluate or evaluating or evaluation or assess*) adj3 evidence).ti.	1916
#17	((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) adj3 recommendation*1).ti.	493
#18	((framework*1 or system*1 or instrument*1) adj2 evidence).ti.	443
#19	#14 or #15 or #16 or #17 or #18	4297
#20	((grade or grades or grading or rate or rates or rating or assess* or evaluate or evaluating or appraisal or appraising) adj2 (quality or strength) adj5 (evidence or recommendation*1)).ti.	129
#21	((method or methods or methodology or approach or approaches) adj3 (assess or assessing or assessment or evaluate or evaluating) adj3 (recommendation*1 or evidence)).ti.	57
#22	(good adj3 practi#e*1 adj3 recommendation*1).tw.	401
#23	(#10 and #19) or #20 or #21 or #22	2776

#24	Limit #23 to English, Human and Publication year (1995 to Current)	1996

4. Cochrane Methodology Register and Cochrane Groups – Cochrane Library (1995 to 2016)

Search Number	Search String	Results
#1	Title, Abstract, Keywords: (((approach or approaches or standard* or system* or instrument* or framework* or level* or hierarchy or hierarchies or method* or profile* or guideline* or guide*) near/3 evidence) and (quality or confidence or certainty or strength))	395
#2	Title, Abstract, Keywords: (good near/3 practice* near/3 recommendation*)	2
#3	Record Title: ((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) near/3 (evidence or recommendation*))	98
#4	Record Title: ((grade or grading or rate or rating or assess or assessing or evaluate or evaluating or appraisal or appraising) near/2 (quality or strength) near/5 evidence)	22
#5	Record Title: ((framework* or system* or instrument*) near/2 evidence)	25
#6	#1 or #2 or #3 or #4 or #5	457

5. Scopus Social Sciences (1960 to Present)

Search Number	Search String	Results
#1	TITLE((approach or approaches or standard* or system* or instrument* or framework* or level* or hierarchy or hierarchies or method* or profile* or guideline* or guide*) W/3 evidence)	1578
#2	TITLE-ABS-KEY(quality or confidence or certainty or strength)	228403
#3	#1 and #2	245

#4	TITLE-ABS-KEY(good W/3 practice* W/3 recommendation*)	303
#5	TITLE((framework* or systems* or instrument*) W/2 evidence)	357
#6	TITLE((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing) W/3 evidence)	431
#7	TITLE((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) W/3 recommendation*)	52
#8	TITLE((grade or grading or rate or rating or assess* or evaluate or evaluating or appraisal or appraising) W/2 (quality or strength) W/5 evidence)	22
#9	TITLE((method or methods or methodology or approach or approaches) W/3 (assess or assessing or assessment or evaluate or evaluating) W/3 (evidence or recommendation*))	14
#10	#3 or #4 or #5 or #6 or #7 or #8 or #9	1315
#11	Limit 8 to English, Publication Year (1995 to 2016)	1205

6. Social Science Citation Index (SSCI); Web of Science™ Core Collection (1956 to 2016)

Search Number	Search String	Results
#1	TI=((approach or approaches or standard* or system* or instrument* or framework* or level* or hierarchy or hierarchies or method* or profile* or guideline* or guide*) near/3 evidence)	4072
#2	TS=(quality or confidence or certainty or strength)	383959
#3	#1 and #2	699
#4	TS=(good near/3 practice* near/3 recommendation*)	415
#5	TI=((framework* or system* or instrument*) near/2 evidence)	836

#6	TI=((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing) near/3 evidence)	1266
#7	TI=((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) near/3 recommendation*)	96
#8	TI=((grade or grading or rate or rating or assess* or evaluate or evaluating or appraisal or appraising) near/2 (quality or strength) near/5 evidence)	47
#9	TI=((methods or method or methodology or approach or approaches) near/3 (assess or evaluate or assessment or assessing or evaluating) near/3 (evidence or recommendation*))	46
#10	#3 or #4 or #5 or #6 or #7 or #8 or #9	3154
#11	Limit 8 to English and Publication Year (1995 to 2016)	2857

7. Applied Social Sciences Index and Abstracts (1987-2016)

Search Number	Search String	Results
#1	ti((approach or approaches or standard* or level* or hierarchy or hierarchies or method* or framework* or system* or instrument* or profile* or guideline* or guide*) near/3 evidence)	604
#2	ab(quality or confidence or certainty or strength)	59539
#3	1 and 2	129
#4	ti(good near/3 practice* near/3 recommendation*) OR ab(good near/3 practice* near/3 recommendation*)	147
#5	ti((framework* or system* or instrument*) near/2 evidence)	131
#6	ti((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing) near/3 evidence)	124
#7	ti((grade or grades or grading or rate or rates or rating or evaluate or evaluating or assess or assessing or quality or strength) near/3 recommendation*)	23
#8	ti((grade or grading or rate or rating or assess* or evaluate or evaluating or appraisal or appraising) near/2 (quality or strength) near/5 evidence)	14
#9	ti((methods or method or methodology or approach or approaches) near/3 (assess or evaluate or assessment or assessing or evaluating) near/3 (evidence or recommendation*))	5
#10	#3 or #4 or #5 or #6 or #7 or #8 or #9	518
	Limit #10 to Publication Year (1995 to 2016)	491

8. SCIE Social Care Online

Search Number	Search String	Results
#1	Title: framework or system or instrument or grade or rate or assess or evaluate or method or methodology or approach or level or hierarchy or appraisal or guideline or guide	19073
#2	Abstract: evidence	12957
#3	Abstract: quality or strength	13041
#4	#1 and #2 and #3	351
#5	All fields: “strength of evidence” or “strength of recommendation” or “quality of evidence” or “quality of recommendation”	131
#6	#4 or #5	452

Key stakeholder organisation website search process

Organisation Website	Search
Agency for Healthcare Research and Quality	**Date:** 11.05.2016 **Search performed:** *evidence grading* (in the Search for Research Summaries, Reviews, and Reports. Effective Health Care Program) **Number of records retrieved and screened:** 237
Appraisal of Guidelines for Research and Evaluation (AGREE)	**Date:** 24.05.2016 **Search performed:** hand searched the Resource Center and Research Projects sections of the website **Number of records retrieved and screened:** 8
Campbell Collaboration	**Date:** 23.05.2016 **Search performed:** hand searched the Methods Group, Campbell Methods Series, Campbell Policies and Guidelines and Resource Center sections of the website **Number of records retrieved and screened:** 7
Canadian Task Force on Preventive Health Care (CTFPHC)	**Date:** 23.05.2016 **Search performed:** hand searched the Methods section of the website **Number of records retrieved and screened:** 1
Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement (DECIPHer)	**Date:** 25.05.2016 **Search performed:** *evidence or grading or rating or assessing or quality or strength* (searched in the Title of the website Publications) **Number of records retrieved and screened:** 37
Centre for Diet and Activity Research (CEDAR)	**Date:** 25.05.2016 **Search performed:** *evidence grading or evidence rating or strength of evidence or quality of evidence or quality assessment* (searched for Keywords in Titles and Abstracts of website Publications); also hand searched the Resources section of the website

	Number of records retrieved and screened: 8
Centre for Evidence-Based Crime Policy	**Date:** 24.05.2016 **Search performed:** hand searched the Research Programs section of the website **Number of records retrieved and screened:** 7
Centre for Evidence-Based Intervention (CEBI)	**Date:** 23.05.2016 **Search performed:** hand searched the Methodology section of the website **Number of records retrieved and screened:** 1
Centre for Evidence-Based Medicine (CEBM)	**Date:** 23.05.2016 **Search performed:** hand searched the EBM Resources section of the website **Number of records retrieved and screened:** 4
Centre for Reviews and Dissemination (CRD)	**Date:** 24.05.2016 **Search performed:** hand searched the Our Guidance section of the website and the CRD Database: Title: *rating or grading or (assessing adj5 quality) or (strength and evidence)* **Number of records retrieved and screened:** 34
Centre for Translational Research in Public Health (Fuse)	**Date:** 25.05.2016 **Search performed:** hand searched the Research Section of the website **Number of records retrieved and screened:** 13
Centre of Excellence for Public Health Northern Ireland	**Date:** 25.05.2016 **Search performed:** hand searched the Research and Publications sections of the website **Number of records retrieved and screened:** 8
Cochrane Collaboration	**Date:** 27.05.2016 **Search performed:** hand searched Cochrane Methods Group, Cochrane and EPOC Resources for Authors section of the website **Number of records retrieved and screened:** 51

Critical Appraisal Skills Programme (CASP)	**Date:** 24.05.2016 **Search performed:** hand searched the CASP Tools and Checklists Section of the website **Number of records retrieved and screened:** 40
Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE)	**Date:** 25.05.2016 **Search performed:** hand searched the Publications and Other Dissemination Activities section of the website **Number of records retrieved and screened:** 6
Department for Education	**Date:** 20.05.2016 **Search performed:** *strength of evidence or evidence grading or evidence rating or quality of evidence* **Number of records retrieved and screened:** 676
Department of Health	**Date:** 20.05.2016 **Search performed:** *strength of evidence or evidence grading or evidence rating or quality of evidence* **Number of records retrieved and screened:** 374
Department for International Development (DFID)	**Date:** 20.05.2016 **Search performed:** *strength of evidence or evidence grading or evidence rating or quality of evidence* **Number of records retrieved and screened:** 184
ESRC UK Centre for Evidence Based Policy and Practice	**Date:** 07.06.2016 **Search performed:** hand searched the outputs of the project **Number of records retrieved and screened:** 32
EQUATOR Network	**Date:** 25.05.2016 **Search performed:** Reporting Guidelines for Systematic Reviews **Number of records retrieved and screened:** 24
European Centre for Disease Prevention and Control	**Date:** 25.05.2016 **Search performed:** *evidence grading or evidence rating or strength of evidence* (searched conducted in Publications, All sites, News and Events sections of the website)

	Number of records retrieved and screened: 156
Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)	**Date:** 16.05.2016 **Search performed:** hand searched publications on systematic review/evidence synthesis methodology: appraising and synthesising evidence **Number of records retrieved and screened:** 19
GRADE Working Group	**Date:** 26.05.2016 **Search performed:** hand searched the Publication section of the website **Number of records retrieved and screened:** 19
Guidelines International Network (G-I-N)	**Date:** 26.05.2016 **Search performed:** hand searched the Working Groups and Resources sections of the website **Number of records retrieved and screened:** 37
International Initiative for Impact Evaluation (3ie)	**Date:** 23.05.2016 **Search performed:** hand searched the Resources and Systematic Reviews sections of the website **Number of records retrieved and screened:** 1
Joanna Briggs Institute (JBI)	**Date:** 23.05.2016 **Search performed:** hand searched the website **Number of records retrieved and screened:** 6
Ministry of Justice	**Date:** 20.05.2016 **Search performed:** *strength of evidence or evidence grading or evidence rating or quality of evidence* **Number of records retrieved and screened:** 119
National Foundation for Educational Research (NFER)	**Date:** 20.05.2016 **Search performed:** *assessing evidence or strength of evidence or evidence grading or evidence rating* (search in the publications) **Number of records retrieved and screened:** 88
National Guideline Clearinghouse	**Date:** 31.05.2016

	Search performed: hand searched the website (Guideline Matrix and Guideline Resources sections) **Number of records retrieved and screened:** 2
National Health and Medical Research Council (NHMRC)	**Date:** 12.05.2016 **Search performed:** hand searched the resources for guideline developers section of the website **Number of records retrieved and screened:** 15
National Institute for Health and Care Excellence	**Date:** 11.05.2016 **Search performed:** hand searched the process and methods guides sections of the website: *evidence grading* **Number of records retrieved and screened:** 42
NHS Health Development Agency	**Date:** 12.05.2016 **Search performed:** searched all the titles of HDA publications **Number of records retrieved and screened:** 606
Norwegian Institute of Public Health	**Date:** 25.05.2016 **Search performed:** hand searched Research Projects section of the website **Number of records retrieved and screened:** 27
Public Health Agency of Canada	**Date:** 07.06.2016 **Search performed:** quality of strength (in the title of documents) **Number of records retrieved and screened:** 15
Scottish Intercollegiate Guidelines Network	**Date:** 23.05.2016 **Search performed:** hand searched the Methodology section of the website **Number of records retrieved and screened:** 4
Social Care Institute for Excellence	**Date:** 23.05.2016 **Search performed:** hand searched the Research & Knowledge section of the website: Knowledge review, Guide and Research resource **Number of records retrieved and screened:** 86

Specialist Unit for Review Evidence (SURE)	**Date:** 25.05.2016 **Search performed:** hand searched the Projects, Publications, Resources for Systematic Reviewers and Critical Appraisal Checklists sections of the website **Number of records retrieved and screened:** 151
The Public Health Agency of Sweden	**Date:** 13.10.2016 **Search performed:** contacted the website for an English version guidance **Number of records retrieved and screened: 0**
The National Board of Health and Welfare (Socialstyrelsen)'s MetodGuiden: Sweden	**Date:** 13.10.2016 **Search performed:** contacted the website for an English version guidance **Number of records retrieved and screened: 1**
US Community Preventive Services Task Force	**Date:** 12.05.2016 **Search performed:** hand searched the "Methods" of the Publication section of the website **Number of records retrieved and screened:** 3
USAID Development Experience Clearinghouse	**Date:** 27.05.2016 **Search performed: evidence** *grading or evidence rating or strength or quality of evidence* (Titles of the Documents) **Number of records retrieved and screened:** 78
Vanderbilt University Evidence-Based Practice Center	**Date:** 27.05.2016 **Search performed:** *grading or rating or strength* (Titles and Abstracts of the Projects and Publications) **Number of records retrieved and screened:** 13
WHO evidence-informed policy-making: Health Evidence Network (HEN)	**Date:** 31.05.2016 **Search performed:** "evidence grading" or "evidence rating" or "quality of evidence" or "strength of evidence" (searched HEN Sources of Evidence Database) **Number of records retrieved and screened:** 44

WHO evidence-informed policy-making: Evidence-informed Policy Network (EVIPNet)	**Date:** 31.05.2016 **Search performed:** hand searched the website (including the Resources for Evidence-Based Policy section of EVIPNet Global website) **Number of records retrieved and screened:** 151
WHO Guidelines	**Date:** 31.05.2016 **Search performed:** searched the text of all the guidelines for "GRADE" or "strength" or "quality" **Number of records retrieved and screened:** 154
World Bank Impact Evaluation: Open Knowledge Repository	**Date:** 26.05.2016 **Search performed:** evidence or grading or rating or quality or strength or confidence (search limited to Titles only) **Number of records retrieved and screened:** 223

US Clearinghouses

Best Evidence Encyclopedia	http://www.bestevidence.org/index.cfm
Best Practices Registry for Suicide Prevention	http://www.sprc.org/strategic-planning/finding-programs-practices
California Evidence-Based Clearinghouse for Child Welfare (CEBC)	http://www.cebc4cw.org
California Healthy Kids Resource Center	http://www.californiahealthykids.org/index
Center for Knowledge Translation for Disability and Rehabilitation Research	http://ktdrr.org
CrimeSolutions.gov	http://CrimeSolutions.gov
Evidence-Based Practices for Substance Use	http://lib.adai.washington.edu/ebpsearch.htm
FindYouthInfo.gov	http://www.findyouthinfo.gov/
Home Visiting Evidence on Effectiveness	http://homvee.acf.hhs.gov
My Brother's Keeper	http://mbk.ed.gov/
National Guideline Clearinghouse	See above
National Registry of Evidence-based Programs and Practices (NREPP)	http://nrepp.samhsa.gov/01_landing.aspx

Office of Adolescent Health: Teen Pregnancy Prevention Evidence-Based Programs	http://www.hhs.gov/ash/oah/
Office of Juvenile Justice and Delinquency Prevention (OJJDP). Model Programs Guide	http://www.ojjdp.gov
OJJDP's Strategic Planning Tool	https://www.nationalgangcenter.gov/SPT/
Promise Neighborhoods Research Consortium	http://promiseneighborhoods.org/index.html
Research-tested Intervention Programs	http://rtips.cancer.gov/rtips/index.do
Strengthening Families Evidence Reviews	http://familyreview.acf.hhs.gov
The Community Guide	See above
The Clearinghouse for Labor Evaluation and Research (CLEAR)	http://clear.dol.gov
United States Interagency Council on Homelessness' Solutions Database	https://www.usich.gov/solutions
Washington State Institute for Public Policy	http://www.wsipp.wa.gov
What Works Clearinghouse (WWC, Department of Education)	http://ies.ed.gov/ncee/wwc/
What Works in Reentry Clearinghouse (WWR)	https://whatworks.csgjusticecenter.org

UK What Works Network

Affiliate: Public Policy Institute for Wales	http://ppiw.org.uk
Affiliate: What Works Scotland	http://whatworksscotland.ac.uk
Centre for Ageing Better	http://www.centreforageingbetter.com/
Early Intervention Foundation	http://guidebook.eif.org.uk/
Education Endowment Foundation	http://educationendowmentfoundation.org.uk/toolkit/
National Institute for Health and Care Excellence	See above
What Works Centre for Crime Reduction	http://www.college.police.uk/en/20018.htm

What Works Centre for Local Economic Growth	http://whatworksgrowth.org/
What Works Centre for Wellbeing	https://whatworkswellbeing.org

Other Agencies

Centre for Excellence and Outcomes in Children and Young People's Services	http://www.c4eo.org.uk/home.aspx
Child family Communities Australia	https://aifs.gov.au/cfca/
Commissioning Toolkit – Parenting Programmes	https://www.education.gov.uk/commissioning-toolkit/Programme/Index
Investing In Children	http://investinginchildren.eu/
Nesta Standards of Evidence	http://www.nesta.org.uk
Project Oracle - Children & Youth Evidence Hub: UK	http://project-oracle.com/
The Edna McConnell Clark Foundation	http://www.emcf.org/our-strategy/selection-process/evidence/

Appendix 2. Data extraction template

1. Descriptive information

Instructions: in this spreadsheet extract data on the requested aspects of the document and its publication.

1.1. Authors of the document	
1.2. Title of the document	
1.3. Year of publication	
1.4. Title of the system	
1.5. Publication source	☐ Journal ☐ Agency/website
1.6. Name of the source	
1.7. Type of the system	☐ Generic ☐ Specific
1.8. Specification (if specific)	☐ Public health ☐ Social work ☐ Psychology ☐ Interventional development ☐ Criminology ☐ Clinical medicine ☐ Other
1.9. Practice domain	
1.10. Purpose of the system	Guideline development Evidence grading
1.11. General comments	

2. Eligibility

2.1. Is the document eligible for inclusion in the review?

☐ Yes
☐ No
☐ Uncertain

2.2. Please, provide reasons, if you answered “No” to 2.1.

3. General content

Instructions: in this spreadsheet extract data on requested items of general information from the system.

3.1. Evidence hierarchy *Does the system adopt an evidence hierarchy approach of any sort when grading evidence? If yes, describe the evidence hierarchy?*	
3.2. Evidence synthesis *Does the system support a specific approach to evidence synthesis? If yes, describe that approach.*	
3.3. Definition of the construct of the "certainty of evidence" *How does the system suggest to define the construct of the "certainty of evidence"?*	
3.4. General comments	

4. Domains of evidence on intervention effectiveness

Instructions: in this spreadsheet extract data on the domains of evidence that the system uses to rate the certainty of evidence on intervention effectiveness. Repeat the process for each separate domain of evidence describe in each system.

4.1. Name of the domain	
4.2. Definition of the domain *How does the system define and justify inclusion of this domain?*	
4.3. Criterion 1 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
4.4. Definition of criterion 1 *How does the system define the criterion (guidance) for rating the specified domain?*	
4.5. Criterion 2 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
4.6. Definition of criterion 2 *How does the system define the criterion (guidance) for rating the specified domain?*	
4.7. Criterion 3 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
4.8. Definition of criterion 3	

How does the system define the criterion (guidance) for rating the specified domain?	
4.9. Criterion 4 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
4.10. Definition of criterion 4 *How does the system define the criterion (guidance) for rating the specified domain?*	
4.11. Criterion 5 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
4.12. Definition of criterion 5 *How does the system define the criterion (guidance) for rating the specified domain?*	
4.13. Process of rating the domain *How does the system describe rating the specified domain?*	
4.14. General comments	

5. Other domains of evidence

Instructions: in this spreadsheet extract data on the domains of evidence that go beyond rating the certainty of evidence on intervention effectiveness. Repeat the process for each separate domain of evidence describe in each system.

5.1. Name of the domain	
5.2. Definition of the domain *How does the system define and justify inclusion of this domain?*	
5.3. Criterion 1 to assess the domain *What specific criterion (guidance) does the system describe to rate the specified domain?*	
5.4. Definition of criterion 1 *How does the system define the criterion (guidance) for rating the specified domain?*	
5.5. Criterion 2 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
5.6. Definition of criterion 2 *How does the system define the criterion (guidance) for rating the specified domain?*	
5.7. Criterion 3 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
5.8. Definition of criterion 3	

How does the system define the criterion (guidance) for rating the specified domain?	
5.9. Criterion 4 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
5.10. Definition of criterion 4 *How does the system define the criterion (guidance) for rating the specified domain?*	
5.11. Criterion 5 to assess the domain *What specific criteria (guidance) does the system describe to rate the specified domain?*	
5.12. Definition of criterion 5 *How does the system define the criterion (guidance) for rating the specified domain?*	
5.13. Process of rating the domain *How does the system describe rating the specified domain?*	
5.14. General comments	

6. Categories of certainty of evidence ratings

Instructions: in this spreadsheet extract data on the categories of ratings that the system uses to rate the certainty of evidence on intervention effectiveness. In addition, extract data on the process that the system employs to rate the certainty of evidence, as well as the specific examples that the systems uses to illustrate the rating process.

6.1. Rating 1	
6.2. Rating 1 explanation	
6.3. Rating 2	
6.4. Rating 2 explanation	
6.5. Rating 3	
6.6. Rating 3 explanation	
6.7. Rating 4	
6.8. Rating 4 explanation	
6.9. Rating 5	
6.10. Rating 5 explanation	
6.11. Worked examples	
6.12. General comments	

7. Development and dissemination of the included systems

Instructions: in this spreadsheet extract data on how the system was developed and disseminated using the questions outlined below.

Preliminary activities and development process

7.1. Review of literature *Did the authors describe whether they identified previous relevant domains of evidence and/or identified key limitations of these? If yes, describe how.*	
7.2. Participants involved *Did the authors report any participants involved in the development of the system? If yes, describe how.*	
7.3. Funding obtained *Did the authors report obtaining any funding for the development of the system? If yes, describe how.*	
7.4. Delphi process *Did the authors report conducting a Delphi exercise? If yes, describe how.*	
7.5. Expert meeting *Did the authors report holding an expert meeting to develop the system? If yes, describe how.*	
7.6. Meeting description *Did the authors describe the process of reaching consensus on the system? If yes, describe how.*	
Write-up and dissemination activities	
7.7. Publication development *Did the authors discuss how the document describing the system was written? If yes, describe how.*	
7.8. Explanation/instructions *Did the authors provide an explanatory document and instructions for using the system? If yes, describe how.*	
7.9. Feedback & criticism *Did the authors describe how they planned to deal with criticism and feedback for the system? If yes, describe how.*	
7.10. Availability on a website *Is the system placed and available in an open-access website? If yes, describe how.*	
7.11. Adherence to the system *Did the authors report any processes for seeking adherence to the system? If yes, describe how.*	
7.12. Translation of the system *Has the system been translated into other languages? If yes, describe how.*	

Appendix 3. Evidence domains in the excluded studies

Specification of the evidence domains for classifying interventions into Tiers

California Evidence-Based Clearinghouse (CEBC) for Child Welfare					
Well-supported by research evidence	**Supported by research evidence**	**Promising research evidence**	**Evidence fails to demonstrate effect**	**Concerning practice**	**Not able to be rated**
• At least two rigorous RCTs in different usual care or practice settings have found the practice to be superior to an appropriate comparison practice • In at least one of these RCTs, the practice has shown to have a sustained effect of at least one year • The RCTs have been published in	• At least one rigorous RCT in usual care or a practice setting has found the practice to be superior to an appropriate comparison practice • In that same RCT, the practice has shown to have a sustained effect of at least six months • That same RCT has been	• At least one study utilising some form of control (e.g., untreated group, placebo group, matched wait list study) has established the practice's benefit • The study has been published in peer-reviewed literature • If multiple studies are conducted, then the majority support the benefit of the practice • Valid and reliable outcome measures • No evidence for risk of harm	• Two or more RCTs have found the practice has not resulted in improved outcomes • The studies have been published in peer-reviewed literature • If multiple studies are conducted, then the majority do not support the	• If multiple studies are conducted, then the majority do not support the negative effects of the practice • There is evidence for risk of harm • The practice has a manual	• The practice does not have any published, peer-reviewed study utilising some form of control • The practice is generally accepted in clinical practice as appropriate for use with children • The practice does not meet criteria for any

peer-reviewed literature • If multiple studies are conducted, then the majority support the benefit of the practice • Valid and reliable outcome measures • No evidence for risk of harm • The practice has a manual	published in peer-reviewed literature • If multiple studies are conducted, then the majority support the benefit of the practice • Valid and reliable outcome measures • No evidence for risk of harm • The practice has a manual	• The practice has a manual	benefit of the practice • Valid and reliable outcome measures • No evidence for risk of harm • The practice has a manual		other level on the CEBC Scientific Rating Scale • No evidence for risk of harm • The practice has a manual
California Healthy Kids Resource Center					
Research-validated programmes					
• Research-Validated programs have empirically demonstrated reductions in health-risk behaviors and/or increases in health-promoting behaviors at least six months after the completion of the program • Research that provides evidence of effectiveness for Research-Validated programs is published in scholarly peer-reviewed journals • Research-Validated program materials are complete, available, and ready to be implemented at school sites in California					

Centre for Excellence and Outcomes In Children and Young People's Services (C4EO)				
Validated Local Practice	**Promising Practice**	**Emerging Practice**		
• A clear description, strong rationale and strong evidence of impact and outcomes for children, young people and their families	• A clear description, good rational and some evidence of impact and outcomes for children, young people and their families	• A clear description and rationale with clearly defined steps identified towards service redesign and/or transformation. There may be little or no evidence yet of impact and outcomes for children, young people and their families		
Center on Knowledge Translation for Disability and Rehabilitation Research (KTDRR)				
Scale 5	**Scale 4**	**Scale 3**	**Scale 2**	**Scale 1**
• Supporting evidence is based on a large, long-term RCT or systematic reviews RCTs of smaller sample sizes that yield sufficient power to confidently predict cause and effect or answer the questions "what works." Technologies/devices are supported by user data demonstrating measured effectiveness and benefit sufficient to support predictable widespread use or technology transfer	• Supporting evidence is based on mixed method research designs and/or RCTs with small sample sizes showing statistically non-significant trends and may show false negative results. Technologies and devices are tested rigorously by substantial number of consumers demonstrating similar consistent results when used	• Supporting evidence is based on qualitative or quantitative research study designs that are non-randomized, controlled, cohort based, case series, case-controlled, cross sectional, survey or descriptive/ exploratory in nature. Innovative technologies/devices are supported by consumer efficacy reviews and evaluations	• Supporting evidence is based on expert opinions of expert panels, committees, professional associations, or consumer organizations or other stakeholder groups	• Supporting evidence is based on opinion of the author(s)

Child Family Community Australia: Promising Practice Profiles		
Promising		
• The practice is effective (the evidence may be obtained in a number of ways, including: stakeholder/client interviews; feedback forms; or communication with other (non-client) stakeholders who have observed change, for example, teachers at the kindergarten or school) • The practice should draw on the available and accepted evidence base about what works to improve outcomes for children, families and communities. The evidence must show that the practice will deliver positive results in at least some situations and contexts • The practice should contribute to our knowledge of 'what works' in the area of early childhood. Consider, in particular, how the practice relates to the National Agenda for Early Childhood priority areas • The practice should be able to be replicated in some other situations and contexts • The practice may be considered promising if it uses a new approach that improves upon, or changes, existing practice • A practice is sustainable when it has the capacity to continue in some form after the initial program has finished		
Clearinghouse for Labor Evaluation and Research (CLEAR)		
High causal evidence	**Moderate causal evidence**	**Low causal evidence**
• RCTs (if they meet all criteria for RCTs) • ITSs (if they meet all criteria for ITSs, including ITS Criteria 2b, 4 and 5 for analysis of a group of people; see below)	• RCTs and ITSs (if they don't meet criteria for high causal evidence) • Non-experimental (regression analyses) designs (if they meet Regression Criteria 1 to 4, studies in which group-level effects are estimated must also meet Regression Criterion 4 and studies involving use of random effects must also meet Regression Criterion 5)	• Causal designs that do not meet criteria for high or moderate causal evidence ratings

Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Study Limitations for RCTs	Confidence that the estimated effects are solely attributable to the intervention examined	• Were there no confounding effects? • Was sample attrition low? • Was the probability of assignment into the research groups consistent over time, or if not, was the change accounted for?

Study Limitations for ITS	Confidence that the estimated effects are solely attributable to the intervention examined	• Was there no selection into the intervention based on pre-intervention trends in the outcomes of interest or the characteristics of participants? • Did the study include multiple demonstrations and multiple observations per demonstration? - Did the study use data from at least three points in time before and three points in time after the demonstration? - Did the study use data from at least five points in time before and five points in time after the demonstration? • Was sample members' anticipation of the intervention either unlikely or appropriately controlled for? • Was the intervention introduced at a predetermined time and in a predetermined manner? • Were there no changes in group comparison (special criterion for ITS designs with group-level analyses)?
Study Limitations for Regression Analyses (e.g. matched comparison group; difference-in-difference; fixed-effects; instrumental variables; other methods)	Confidence that the estimated effects are solely attributable to the intervention examined	• Were the intervention and comparison groups similar before the intervention? • Were there no confounding factors? • Was sample members' anticipation of the intervention either unlikely or appropriately controlled for? • Were there no changes in group composition (special criterion for estimates for group-level effects)? • Was random effects used over fixed (special criterion for random effects)? • Did the instrument have sufficient strength (special criterion for instrumental variables)? • Did the instrument satisfy the exclusion restriction (or instrument exogeneity; special criterion for instrumental variables)? • Was the rank condition satisfied (special criterion for instrumental variables)?

The Parenting Programme Evaluation Tool (PPET)		
Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Specification	Who is the programme designed for and for what level of need?	• Who is the programme/approach designed for (i.e., description of the targeted population)? • What is the process for checking if parents are suitable for participation in the programme/approach (i.e., needs analysis or assessment)? • What change is likely when parents participate in the programme (i.e., expected outcomes)? • What is the classification of the programme (programme classification)?
Content and Processes	What is the content of the programme and how do you deliver it?	• What is the theoretical framework or assumption/s that the programme is based on (i.e., validity of the theoretical framework)? • What do parents learn during the course of the programme (i.e., content of the programme)? • How do parents learn during the course of the programme (i.e., format of the programme)? • What resources are available to enable other practitioners to deliver the programme (i.e., appropriateness and adequacy of the resources)?
Implementation Processes	How do you train and support others to use the programme most effectively and consistently in new settings?	• What level of experience and qualification do practitioners need to run the programme? • What training is available to instruct practitioners to be able to deliver the programme? • What mechanism is available to support and supervise practitioners to deliver the programme? • What mechanisms are available to support organisations wishing to implement the programme in their area (i.e. capacity for dissemination)?
Evaluation Quality	What is the effect of the programme on the targeted outcomes?	• What mechanisms are used to evaluate the outcomes of the programme? Examples of the information which will be considered: - Sample: sufficient size, description of who/what and recruitment methods

		- Design: strongest design possible to allow for causal inferences, assignment procedure (random, matched-sample) - Outcome measures: multiple methods – direct observation, self-report, parent report, reports by other significant people; psychometrically sound (reliable and valid) - Data collection procedures: multiple measurement periods to detect durability of results; collection independent of practitioner delivering the intervention - Analyses and results: attrition, appropriateness of analysis strategy, statistically significant effects, clinically significant effects - Replication studies
Dartington Social Research Unit: The "What Works" Standards of Evidence		
Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Intervention Specificity	This standard is concerned with whether an intervention is focused, practical and logical	Good enough • Is the intended population of focus clearly described? • Are the outcomes of the intervention clearly specified and do they meet one of the key developmental outcomes or outcome domains? • Is it possible to identify the risk and protective factors that the intervention seeks to change, using the intervention's logic model or theory explaining why the intervention may lead to better outcomes? • Is there clarity and documentation about what the intervention comprises? Best • Is there a research base summarising the prior empirical evidence to support the causal mechanisms (risk and protective factors) that underlie the change in outcome being sought?

Evaluation Quality	This standard measures the level of confidence we can have in the evaluation	Good enough • Has the intervention been evaluated by at least one randomised controlled trial (RCT) OR two quasi-experimental (QED) evaluations (initial QED and a replication) with the following characteristics: - Was the assignment to the intervention at a level appropriate to the intervention, i.e. individual, school, etc. (to reduce spill-over effects, etc.)? - Was there use of measurement instruments that are appropriate for the intervention population of focus and desired outcomes? - Was analysis based on "intent-to-treat"? - Were there appropriate statistical analyses? - Did the analysis of baseline differences indicate equivalence between - intervention and comparison groups? • Was there a clear statement of the demographic characteristics of the population with whom the intervention was tested? • Was there documentation regarding what participants received in the intervention and counterfactual conditions? • Was there no evidence of significant differential attrition? • Were outcomes measure not dependent on the unique content of the intervention? • Did outcome measures reflect relevant key developmental outcomes or outcome domains? • Were outcome measures rated not solely by the person or people delivering the intervention? Best • Were there two RCTs OR one RCT and one QED evaluation (in which analysis and controls rule out plausible threats to internal validity)?

		• Was there a minimum of one long-term follow-up (at least 12 months following completion of the intervention) on at least one outcome measure indicating whether results are sustained over time? • Did the evaluation results indicate the extent to which fidelity of implementation affects the impact of the intervention? • Was dose-response analysis reported? • Where possible and appropriate was there analysis of the impact on sub-groups (e.g. do the results hold up for different age groups, boys and girls, ethnic minority groups)? • Was there verification of the theoretical rationale underpinning the intervention, provided by mediator analysis showing that effects were taking place for the reasons expected?
Intervention Impact	This standard measures how much difference the intervention makes	Good enough • Was there a positive impact on a relevant key developmental outcome or outcome domain? • Was there a positive and statistically significant effect size, with analysis done at the level of assignment (or, if not, with appropriate correction made)? • Was there a reported sample size weighted mean effect size of .2, with a sample size of more than 500 individuals across all studies? • Was there an absence of iatrogenic effects for intervention participants (this includes all sub-groups and important outcomes)? Best • If two or more RCTs or at least one RCT and one QED had been conducted, and they met the "good enough: methodological criteria stipulated for Evaluation Quality, was there evidence of a positive effect and an absence of iatrogenic effects from a majority of the studies? • Was there evidence of a positive dose-response relationship that meets the "best" methodological standard for identifying this in Evaluation Quality?

System Readiness	This standards is concerned with whether the intervention can be implemented within service systems	Good enough • Were there explicit processes for ensuring that the intervention gets to the right people? • Were there training materials and implementation procedures? • Were there one or more manuals detailing the intervention? • Was there reported information on the *financial* resources required to deliver the intervention? • Is the intervention that was evaluated still available? Best • Is the intervention currently being widely disseminated? • Has the intervention been tested in "real world" conditions? • Is technical support available to help implement the intervention in new settings? • Is there a fidelity protocol of assessment checklist to accompany the intervention?
Dutch Recognition System for Health Promotion Interventions		
Effectiveness	**-**	**General criteria for all levels of effectiveness** • Were the outcomes most relevant given the objective and the target group for the intervention? • Did the changes relate the objective and the target group of the intervention? - Did the studies reveal that the intended target group has been effectively achieved? - Did the employed instruments provide a reliable and valid operationalisation to measure the realisation of the objectives of the intervention? - Did the study use satisfactory statistical techniques? - Was the size of the effect indicated in terms of Cohen's D or the data to calculate Cohen's D is specific

		• Is the size of the effects reasonably convincing and does it match the objective and the target group of the intervention? • Were possible negative effects stated? • Was research documented such that replication of the study was possible? • Had the intervention been implemented as intended? Had it been demonstrated that the elements of the intervention had actually been applied? • Were there enough studies to conclude that during the implementation of the intervention changes occurred in accordance with the intervention's objectives? **Strong indications for effectiveness** • Did the design of the empirical research provide for at least a strong causal level of evidence the research has a quasi-experimental/experimental or, if that is not possible, another design (for example, repeated case studies, a study into the correlation between the extent to which the intervention is applied and the extent to which the intended outcomes have occurred, or a cohort study) of high quality? • Were the studies carried out in everyday practice with a follow-up period of at least six months? • Were there at least two Dutch studies into the intervention in question with a strong or very strong level of evidence or one Dutch study into the intervention in question in combination with at least one national or international study into this or a comparable intervention with a strong or very strong level of evidence? • In case of repeated case studies were there at least ten cases carried out by different treating practitioners under different conditions? **Good indications for effectiveness** • Did the design of the empirical research provides for at least a moderate causal level of evidence the research has a quasi-experimental/experimental

		or, if that is not possible, another design (for example, repeated case studies, a study into the correlation between the extent to which the intervention is applied and the extent to which the intended outcomes have occurred, or a cohort study) of high quality? • Were there at least two Dutch studies into the intervention in question with a moderate to fairly strong level of evidence or one Dutch study into the intervention in question in combination with at least one national or international study into this or a comparable intervention with at least a moderate level of evidence? • In case of repeated case studies were there at least six cases carried out by different treating practitioners under different conditions? **First indications for effectiveness** • Did the design of the empirical research provide for at least a weak causal level of evidence and a baseline measurement and a follow-up measurement, without a control condition? • Were there at least two Dutch studies into the intervention in question with a weak level of evidence or one Dutch study into the intervention in question in combination with at least one national or international study into this or a comparable intervention with at least a weak level of evidence?
Theoretically-sound	**-**	**Description** • Were the nature, size, spread and possible consequences of the problem or theme clearly described? • Were the target group for the intervention clearly described on the basis of relevant characteristics; were possible inclusion and exclusion criteria stated? • If the target group was involved in the development of the intervention then how this happened was described? • Had the objectives been formulated as tangibly as possible and if relevant were they distinguished in main objective and sub-objectives?

		• Were the sequence, frequency, intensity, duration, timing of activities, recruitment method and location of the intervention described? • Was the method of the intervention described as completely as possible in concrete activities? • Was a description given of the parties involved in the implementation and how these parties collaborated? • Were the materials needed and their availability clearly described? **Theoretical underpinning/intervention logic** • Are the problem, risk or theme completely and clearly described with data about, for example, the nature, severity, size, spread, perception of those involved, costs and other possible consequences? • Had the analysis been made of how the problem has arisen in which the possible causal, risk, maintenance, mitigating or protective factors are described? • Were the factors that will be tackled with the intervention stated and linked to the objectives and sub-objectives of the intervention (justifying objectives)? • Were the effective elements (or techniques or principles) in the approach stated and justified, in the framework of a change model or an intervention theory, or based on the results of research carried out previously. • Do target groups, objectives and working method fit together? Is justification given of how the approach chosen will be able to effectively achieve the objectives for this target group? • Where relevant, were sources stated with respect to the theoretical underpinning? **Implementation conditions/feasibility** • Is the intervention transferable? - Is there a manual or protocol for transfer? - Is there support for the introduction of the intervention (training the trainer, supervision, helpdesk, etc.)?

		- Is there a system for implementation or an implementation plan? • Were data about maintenance, quality care and safeguarding specified? • Were the boundary conditions essential for the implementation specified? (e.g. intervention: use of personnel, use of time, costs; implementing professionals: training, experience, competencies; organisation: internal and external support, possibilities for internal and external collaboration) • Is it likely that the objective can be realised within the boundary conditions and costs stated? • In case if intervention has not been developed in the Netherlands then is the original context briefly described and the modifications made to adapt the intervention to the Dutch situation are explained? • If relevant to the problem or the area of implementation, does the intervention offer space for flexibility: i.e. does the manual contain information about the effective principles or elements that must be adhered to? • Had a pre-test or process evaluation been carried out? - Was the study design described? - Were data available about, e.g., the scope, success and failure and the assessment of implementers? - Were the results positive? - Had the intervention been modified (insofar as necessary) on the basis of these results? • Does the research reveal the relevant context factors that influence the effect and implementation of the intervention?
Well-described	**-**	**Description** • Were the nature, size, spread and possible consequences of the problem or theme clearly described? • Were the target group for the intervention clearly described on the basis of relevant characteristics; were possible inclusion and exclusion criteria stated?

		• If the target group was involved in the development of the intervention then how this happened was described? • Had the objectives been formulated as tangibly as possible and if relevant were they distinguished in main objective and sub-objectives? • Were the sequence, frequency, intensity, duration, timing of activities, recruitment method and location of the intervention described? • Was the method of the intervention described as completely as possible in concrete activities? • Was a description given of the parties involved in the implementation and how these parties collaborated? • Were the materials needed and their availability clearly described? **Consistency** • Were the relationship between background, objectives, target groups and approach clearly described? **Implementation** • Were the necessary costs of and/or hours needed for the intervention stated? • Were the specific skills and vocational training of the professionals who would implement the intervention described? • Were people needed to support the intervention stated and was the ways this support could described? • Does the manual contain a description of the objectives, target group and materials as well as the content of the various activities? • Was the support offered for implementing and realising the intervention described? • Were the means for the quality of the intervention monitoring described?

Early Intervention Foundation (EIF)					
Consistently effective	**Effective**	**Potentially effective**	**Theory-based**	**Unspecified**	**Ineffective/harmful**
• Multiple high-quality evaluations (RCT/QED) with consistently positive impact across populations and environments	• Single high-quality evaluation (RCT/QED) with positive impact	• Lower-quality evaluation (not RCT or QED) showing better outcomes for programme participants	• Logic model and testable features, but not current evidence of outcomes or impact	• No logic model, testable features, or current evidence of outcomes or impact	• Evidence from at least one high-quality evaluation (RCT/QED) indicating null or negative impact

Education Endowment Foundation				
Very extensive	**Extensive**	**Moderate**	**Limited**	**Very limited**
• Consistent high quality evidence from at least five robust and recent meta- analyses where the majority of the included studies have good ecological validity and where the outcome measures include curriculum measures or standardised tests in school subject areas	• Three or more meta-analyses from well controlled experiments mainly undertaken in schools using pupil attainment data with some exploration of causes of any identified heterogeneity	• Two or more rigorous meta-analyses of experimental studies of school age students with cognitive or curriculum outcome measures	• At least one meta-analysis or systematic review with quantitative evidence of impact on attainment or cognitive or curriculum outcome measures	• Quantitative evidence of impact from single studies, but with effect size data reported or calculable. No systematic reviews with quantitative data or meta-analyses located

National Institute of Justice: CrimeSolutions.gov (practices)		
Effective	**Promising**	**No effects**
• If one meta-analysis: Outcomes in Class 1[2] • Ratings for the same outcome from multiple meta-analyses are summed and averaged: If the averaged points ≥ 1.50, then final outcome rating is Effective	• If one meta-analysis: Outcomes in Class 2[3] • Ratings for the same outcome from multiple meta-analyses are summed and averaged: If the averaged points are ≥ 0.50 and ≤ 1.49, the final outcome rating is Promising	• If one meta-analysis: Outcomes in Class 3[4] or Class 4[5] • Ratings for the same outcome from multiple meta-analyses are summed and averaged: If the averaged points ≤ 0.49, the final outcome rating is No Effects.

Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Study Limitations for Meta-Analysis	The Study Reviewers make this assessment based on information about the methods and procedures used in each meta-analysis. Multiple items are scored to determine overall quality and these items are weighted differently based on their importance	Methodological Quality • To what extent were meta-analyses authors were attentive to the methodological quality of the primary studies included in the meta-analysis? Main Analyses • To what extent did the authors use appropriate methods for calculate and report effect size estimates (e.g. handling dependent effect sizes; effect size reporting; weighting of results; analysis model; heterogeneity attentiveness)? Eligibility and Search • To what extent does the meta-analysis provide a clear statement of the inclusion and exclusion criteria for selecting

[2] Statistically significant mean effect size favouring the intervention and a summative score for 2 dimensions of effectiveness combined

[3] Statistically significant mean effect size favouring the intervention and a summative score of 4 or 5

[4] Statistically significant mean effect size favouring the comparison condition and a summative score of 4, 5 or 6

[5] The mean effect size is not statistically significant and a summative score of 2 or 3

		studies to be included, and was the literature search comprehensive and not limited to commercial publishers (e.g. comprehensive literature search; grey literature searches)? Reliability, Outliers, Publication Bias • To what extent did the authors use appropriate methods to extract data from primary studies, account for extreme scores and biases towards large and statistically significant effects in published findings (e.g. coder reliability, outlier analysis, publication bias)?
Study Limitations: Internal Validity	Internal validity refers to the extent to which the research design is free from threats that potentially bias the effect estimate. Randomized control trials (RCTs) have the strongest inherent internal validity and this item uses the extent to which the mean effect size is based on results from randomized controlled trials to assess the overall internal validity of the mean effect size being coded	• Were at least 60% (30% or fewer for medium and low internal validity ratings, respectively) of the studies included in the mean effect size RCTs? • Was the mean effect size covariate adjusted to estimate the effect size expected if all studies were RCTs? • Were there at least 5 RCTs, the mean effect sizes for the RCTs and non-RCTs reported separately, and at least one of the following applied? - A statistical test was reported that showed no statistically significant difference between the mean effect size for the RCT and non-RCT studies - The mean effect sizes for the RCTs and non-RCTs both fell within an approximate fixed effect 95% confidence interval around the mean effect size for both combined. - The mean effect size for the RCTs is tested for statistical significance, the - The mean effect size for the combined RCTs and non-RCTs was in the same direction and was also tested for statistical significance, and the results were the same in both cases (both significant or both non-significant)

National Institute of Justice: CrimeSolutions.gov (programmes)		
Effective	**Promising**	**No effects**
• Must have at least 1 study in Class 1[6] • May have up to two studies in Class 2[7] • Must have 0 studies in Class 3[8] • May have up to 1 study in Class 4[9]	• Must have 0 studies in class 1 • Must have at least 1 study in class 2 • Must have 0 studies in class 3 • May have up to 1 study in class 4	• Must have 0 studies in class 1 • Must have 0 studies in class 2 • Must have at least 1 study in either class 3 or class 4

Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Conceptual framework	Conceptual framework assesses the degree to which the program is grounded in the research literature	Prior Research • To what extent does previous evidence support the conceptual framework of comparable programmes? Theoretical Base • To what extent is the programme based on a well-articulated, conceptually sound program theory? Program Description • To what extent are the programme details described (i.e. the logic of the program; the details of all key components; the frequency and duration of the programme activities; the targeted population; the targeted behaviours; the setting)?
Design quality	Design quality assesses the quality of the research design	Research Design • What is the ability of the design to infer a causal relationship between program treatment and outcome? Sample Size (Power)

[6] Strong evidence of positive effect: Must have at least 2.0 scores in all four dimensions of programme effectiveness

[7] Some evidence of positive effect: Must have at least 1.5 scores in the design and outcome evidence dimensions

[8] Strong evidence of negative effect: Must have less than 0 score in the outcome evidence dimension, yet at least 2.0 in design and fidelity in other dimensions of programme effectiveness

[9] Strong evidence of null effect: Must have a neutral score (from 0 to 1.4) in the outcome evidence dimension, yet at least 2.0 in design and fidelity in other dimensions of programme effectiveness

		• How adequate was the sample size to detect meaningful programme effects? Statistical adjustment • Did the analysis employ necessary statistical adjustments to control for group differences? Instrumentation • What were the reliability and validity of the measures used in the study? Internal validity • To what extent can the observed changes be attributed to the programme (i.e. to what extent are they free from threats to internal validity)? Follow-up period • What was the length of follow-up (e.g. less than or equal to 6 months; more than 6 months; but less than a year; more than a year)? Displacement/diffusion/anticipatory benefits • To what extent did the evaluation examine for the presence of any crime displacement, diffusion of benefits, or anticipatory benefits surrounding the program implementation
Outcome evidence	Outcome evidence assess the quality of the results	Substantive program effects • What is the level of confidence that the effect is the result of the programme rather than other factors? Behaviour • To what extent did the program demonstrate changes in behaviour? Outcome (directional indicator) • What is the direction of the effects based on the preponderance of the evidence (e.g. negative, no effect, positive)?
Fidelity	Fidelity assesses the degree to which the program is delivered as designed and intended	Documentation • To what extent were the core programme services or components implemented as designed via the programme description? Adherence (directional indicator)

		• To what extent were the core programme services or components implemented/delivered as designed (via the programme description)?
National Registry of Evidence-based programs and practices (NREPP)		
Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Rigour	Rigor assesses the strength of the study methodology	• Design/Assignment Were the two group equivalent at the beginning? • Intent-to-Treat Analysis Did the analysis preserve the assignment of participants to their original groups? • Statistical Precision What was the size of the study groups? (sample size) • Pretest Equivalence Were there significant differences between study groups on observed variables at pretest? • Pretest Adjustment Did authors adjust for pre-existing differences/scores? • Analysis Method Did the authors choose the method for analysis that best suits the data? • Other threats to internal validity E.g: were there factors other than the program's impact that could account for changes in the outcomes? • Measurement Reliability Did the study use consistent/reliable outcome measures? • Measurement Validity Did studies use valid outcome measures? • Attrition What was the drop-out rate in the study?

Fidelity	Reviewers examine the evaluation studies to determine if the program was delivered as intended and to the target population.	• Service utilisation Did the program reach the appropriate target population? • Service Delivery Were the core program services or components implemented as depicted in the conceptual framework? (or did the participants receive the proper amount, type, and/or quality of service or treatment?)
Effect Size	An effect size is a way to measure whether a program had an impact, how big that impact was, and whether it helped or hurt the treatment group.	• Favourable: CI lies within the favourable range (> than .10) • Possibly favourable: CI spans both the favourable (> than .10) and trivial range (from -.25 to .10) • Trivial: CI lies within the trivial range (from -.25 to .10) or spans the harmful and favourable range (from -.25 to > than .10) • Possibly harmful: CI spans both the harmful (lower than -.25) and trivial range (from -.25 to .10) • Harmful: CI lies within the harmful range (lower than -.25)
Conceptual Framework	This dimension is concerned with how clearly the components of a program are articulated.	• Program goals To what extent are the program goals clearly defined? • Program components To what extent are the program activities or components sufficient to attain the intended goal. • Theory of change To what extent is the program impact theory plausible?

Nesta Standards of Evidence for Impact Investing				
Level 5	**Level 4**	**Level 3**	**Level 2**	**Level 1**
• You can show that your intervention could be operated up by someone else, somewhere else and scaled up, whilst continuing to have positive and direct impact on the outcome, and whilst remaining a financially viable proposition. We expect to see use of methods like multiple replication evaluations; future scenario analysis; fidelity evaluation • You have manuals, systems and procedures to ensure consistent replication and positive impact	• You are able to explain why and how your intervention is having the impact you have observed and evidenced so far. An independent evaluation validates the impact • The intervention can deliver impact at a reasonable cost, suggesting that it could be replicated and purchased in multiple locations. At this stage, we are looking for a robust independent evaluation that investigates and validates the nature of the impact	• You can demonstrate that your intervention is causing the impact, by showing less impact amongst those who don't receive the product/service. We will consider robust methods using a control group (or another well justified method) that begin to isolate the impact of the product/service. Random selection of participants strengthens your evidence at this Level, you need to have a sufficiently large sample at hand (scale is important in this case). You can	• You are gathering data that shows some change amongst those receiving or using your intervention. At this stage, data can begin to show effect but it will not evidence direct causality. You could consider such methods as: pre and post– survey evaluation; cohort/panel study, regular interval surveying. You capture data that shows positive change, but you cannot confirm you caused this	• You can give an account of impact. By this we mean providing a logical reason, or set of reasons, for why your intervention could have an impact and why that would be an improvement on the current situation. You should be able to do this yourself, and draw upon existing data and research from other sources. You can describe what you do and why it matters, logically, coherently and convincingly

		demonstrate causality using a control or comparison group		
Office of Juvenile Justice and Delinquency Prevention				
Level I. Effective/Exemplary	**Level II. Effective**		**Level III. Promising**	
• Must score an overall rating of at least 28 points with at least 6 points in each dimension of program effectiveness. In general, when implemented with a high degree of fidelity, these programs demonstrate robust empirical findings using a reputable conceptual framework and an evaluation design of the highest quality	• Must score an overall rating of at least 23 points with at least 5 points in each dimension of program effectiveness. In general, when implemented with sufficient fidelity, these programs demonstrate adequate empirical findings using a sound conceptual framework and an evaluation design of high quality		• Must score an overall rating of at least 18 points with at least 4 points in each dimension of program effectiveness	
Project Oracle Standards of Evidence				
Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***		
Standard 5: System Ready	A project validated at Standard 5 has undertaken multiple independent evaluations in different settings, providing evidence that the project can be replicated as a large-scale mainstream service and deliver strong results	What kinds of projects are suitable for validation at Standard 5? • Has the project been extensively replicated across multiple cities or regions in the UK? What kind of evidence is required at Standard 5? • Are there multiple independent and rigorous evaluations (at least 3 external evaluations of operations in at least 5 UK locations)? What kinds of technical support do I need to be able to provide at Standard 5?		

		• Is there systems and documentation to support large-scale implementation in place and can the running the intervention be transferred to other agencies?
Standard 4: Model Evidence	A project validated at Standard 4 has undertaken independent evaluation which demonstrates that changes in the outcomes are observed when the project is replicated in other settings. The evaluation provides insight into how these changes come about, and if doing more or less of your project, or parts of it, has better or worse results. Cost-benefit analysis of the project has also been undertaken	What kinds of projects are suitable for validation at Standard 4? • Has the project been validated at least once? What kind of evidence is required at Standard 4? • Are there two or more rigorous impact evaluations of the project, including at least one undertaken by an external evaluator and at least one involving comparison group or other appropriate comparison data? What kinds of technical support do I need to be able to provide at Standard 4? • Is there technical support and detailed information on the resources (money and people) required to deliver the project?
Standard 3: Evidence of Impact	A project that is validated at Standard 3 has undertaken evaluation that draws a consistent link between the project and the change in outcomes, indicating that the project has caused the observed changes. The project also has procedures in place to increase the likelihood of it being implemented in the future in ways faithful to its design	What kinds of projects are suitable for validation at Standard 3? • Has the project undertaken evaluation that draws a consistent link between the project and the change in outcomes, indicating that the project has caused the observed changes? What kind of evidence is required at Standard 3? • Has there been at least one rigorous evaluation using a comparison group or other appropriate comparison data, ideally with long term follow up? In cases, when impossible to set up suitable control groups or use appropriate comparison data, or where long term follow-up is not feasible, does the programme show a strength

		of the theoretical model underpinning the intervention, or the quality of the data used to assess impact? What kinds of technical support do I need to be able to provide at Standard 3? • Are there procedures in place to ensure consistent implementation of the project (e.g. manuals and staff training processes)?
Standard 2: Indication of Impact	A project that is validated at Standard 2 has undertaken an evaluation that measures relevant outcomes in an appropriate way, and has provided findings that indicate that the project has a demonstrable effect on those outcomes. The main part of the Standard 2 validation is undertaking an evaluation and writing a report that meets the criteria outlined below: Evaluation design: Your evaluation measures changes in the outcomes in an appropriate way. This can include qualitative and quantitative methods. Control and comparison groups are not a requirement. The methods you use must: - include pre and post analysis - use valid and reliable measurement tools which are appropriate for the participants	Evaluation Design • Does the evaluation measure changes in the outcomes in an appropriate way (may include qualitative and quantitative methods, control and comparison groups are not a requirement, methods muse include: pre-post analysis, use of valid and reliable measurement tools)? Evaluation Report • Is there an evaluation report, which meets the criteria for this standard?
Standard 1: Project Model & Evaluation Plan	A project that is validated at Standard 1 has provided a coherent and plausible description of the logic that lies behind it. This includes a description of the project activities, intended outcomes and aims, how these are connected	• What is the aim that you are working towards? • What are the measurable outcomes, which you can affect, that contribute to the aim? • What are the activities that contribute to the outcomes? • Which activities contribute to each outcome?

	and what assumptions are being made; therefore, providing all the elements of a Theory of Change. The project has also planned how it will evaluate the effect it has on the outcomes described, and has a timeline for implementing this.	• What assumptions have you made in determining your outcomes? • To what extent was evidence used in the design of your project? • What is your evaluation plan? • What is your theory of change? • What is your supporting evidence (attach to the programme)?

Quality and Impact of Component (QuIC) Evidence Assessment			
Best	**Promising Quality**	**Promising Impact**	**Emerging**
• At least one empirical study - Directly or indirectly linking the component to mostly positive actual outcomes relevant to health and equity, reach, and/or efficiency across two or more distinctly different regions of the United States - Published in a credible journal • Several research and practice-based studies as well as a large amount of translational evidence • Evidence from an expert group recommending the	• At least one empirical research study published in a credible journal directly or indirectly projecting a positive expected outcome relevant to health • Many narrative reviews or commentaries suggesting the component based on translation of research-based and practice-based studies and/or practice-based knowledge and experience • Evidence from an expert group suggesting the component based on narrative review	• At least one empirical study, which directly links a component to a positive actual outcome relevant to health and equity, reach, and/or efficiency - This study likely took place across two or more distinctly different regions of the United States - However, its design may not have been rigorous, the report may have not come from a source with known credibility, and there may be little to no evidence from research, translation, or practice to	• At least one commentary, narrative review, or predictive study directly or indirectly suggesting positive expected outcomes relevant health, equity, reach, and/or efficiency • However, evidence on actual outcomes is still needed

component based on narrative or systematic review		confirm positive outcomes	
Research-tested Intervention Programs (RTIPs)			
Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***	
Research Integrity	Research Integrity reflects the overall confidence reviewers can place in the findings of a program's evaluation based on its scientific rigor	Theory/hypothesis driven measure selection • Were outcome measures supported by literature related to study theories and/or hypotheses? Reliability • Did outcome measures demonstrate evidence of reliability? Validity • Did outcome measures demonstrate evidence of validity? Intervention Fidelity • Was experimental intervention implemented as intended or modified as appropriate as the study progressed? Nature of Comparison Condition • Was study's comparison an appropriate contrast to the experimental intervention? Comparison Fidelity • Was the comparison condition implemented as intended or modified as appropriate as the study progressed? Assurances to participants • Were confidentiality and assurances that participants' standard of care would not be affected by study participation likely to result in more accurate responses? Participant expectations • Was information used to recruit and inform study participants carefully crafted if possible to equalise expectations?	

		Standardised data collection • Were all outcome data collected in a standardised manner? Data collection bias • Were data collectors aware of the conditions to which study participants had been assigned? Selection bias • Did the study establish baseline equivalence across study conditions on key variables, and if not, were baseline differences appropriately controlled for in statistical analyses? Attrition • Did the study appropriately account for attrition? Missing data • Did the study appropriately account for missing data? Analysis meets data assumptions • Did the study employ appropriate statistical analyses with regards to the data? Hypothesis-driven selection of analytic methods • Were data analytic approaches consistent with study hypotheses rather than being ex post facto data-driven. Anomalous findings • Do the findings contradict the theories and hypotheses underlying an intervention?
Intervention Impact	Intervention Impact describes whether, and to what degree, a program is usable and appropriate for widespread application and dissemination	Population reach • What proportion of members of the defined target population (i.e., defined according to demographic and/or risk factor characteristics) did the study or studies exclude or probably exclude? Intervention effect size

		• What was the effect size of the intervention (Small; Medium; Large)?
Dissemination Capability	Dissemination Capability refers to the readiness of program materials for use by others as well as a program's capability to offer services and resources to facilitate dissemination	• What is the quality of implementation materials? • Does the programme provide training and technical assistance protocols? • Are there quality assurance measures available to determine whether implementation was done with high fidelity to the original model?

The Best Evidence Encyclopedia				
Strong evidence of effectiveness	**Moderate evidence of effectiveness**	**Limited evidence of effectiveness: strong evidence of modest effects**	**Limited evidence of effectiveness: weak evidence with notable effect**	**No quality studies**
• At least one large randomised or randomised quasi-experimental study and one additional large qualifying study, or multiple smaller studies, with a combined sample size of 500 and an overall weighted mean effect size of at least +.20	• Two large matched studies, or multiple smaller studies with a collective sample size of 500 students, with a weighted mean effect size of at least +.20	• Studies meet the criteria for "Moderate Evidence of Effectiveness" except that the weighted mean effect size is +.10 to +.19	• A weighted mean effect size of at least +.20 based on one or more qualifying studies insufficient in number or sample size to meet the criteria for "Moderate Evidence of Effectiveness"	• No studies met inclusion standards

The Edna McConnell Clark Foundation		
Proven Effectiveness	**Demonstrated Effectiveness**	**High Apparent Effectiveness**
• A well-designed and well-executed experimental evaluation of program outcomes, created and conducted by an independent, external evaluator, establishes the most rigorous evidence of effectiveness. Ideally, participants in the study are randomly assigned to the treatment groups. Outcome data for both groups is collected and compared in this randomised controlled trial • The study concludes there are meaningful, positive, statistically significant differences between outcomes for youth served by the programme and outcomes for youth in the control group • At the highest level of proven effectiveness, a programme has evidence of impact from multiple sites • Under some circumstances, a well-implemented program that has been proven effective elsewhere, or a third-party quasi-experimental evaluation that compares participants to a comparison group that has not been randomly assigned, may represent the highest proof point a program is capable of reaching	• A well-designed and well-executed quasi-experimental evaluation of programme outcomes, created and conducted by an independent, external evaluator, measures outcomes for programme participants against outcomes for a carefully chosen comparison group. People in both groups are at the same baseline on measured characteristics such as demographics and variables relevant to the study, and likely to be similar when it comes to unmeasured characteristics such as motivation at the start of the study • This study, also called a comparison group evaluation, concludes there are meaningful, positive, statistically significant differences between outcomes for youth served by the programme and outcomes for youth in the comparison group	• Every programme participant is given a unique identifier (such as a tracking or identification number) • The organisation collects basic demographic data from programme participants, such as address and contact information, age, gender, race/ethnicity, primary language, and socioeconomic status • Initial data about programme participants includes baseline data for measuring changes over time (outcomes) • The outcomes the organisation intends for programme participants are specified in a theory of change • Outcomes are tracked for all programme participants (or at least for a sample), and show meaningful, positive results, comparable to the results from similar well-implemented programmes.

The weight of evidence: A method for assessing the strength of evidence on the effectiveness of HIV prevention interventions among young people			
Go	**Ready**	**Steady**	**Do not go**
• Identified mechanism of action • Experiential base • Adequate intensity, duration and completeness • Probability evidence • Moderate size of positive effect (statistical effect plus the reach of intervention) • Positive effect is in the cultural context being proposed • Consistency in findings in >1 study	• Experiential base • Adequate intensity, duration and completeness • Plausibility evidence • At least small size of positive effect (statistical effect plus the reach of intervention) • Consistency in findings in >1 study	• Experiential base • Careful pilot or informed judgement	• Experiential base • Adequate intensity, duration and completeness • Sufficient condition to recommend "Do not go" if lack of effectiveness or harmful effects found in several studies for this type of intervention • Negative (harmful) effect is sufficient condition for "Do not go"

U.S. Department of Health and Human Services: Strengthening Families Evidence Reviews			
High	**Moderate**	**Low**	**Unrated**
• High quality ratings are only applicable to RCTs if they meet all the criteria (see below)	• Moderate quality ratings are applicable to RCTs (if all criteria are met except for SL5 and SL6 or SL2 or SL3) and quasi-experimental studies if they criteria SL4, SL5 and SL6 below	• Studies received a low rating if it includes participant outcomes but does not use a comparison group, as well as studies that have methodological problems, i.e. cannot be rated as high or moderate	• Studies that do not include participant outcomes

Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Study Limitations	This rating reflects the level of confidence that should be applied when assessing how well the research design can determine whether the	• Was the sample randomly assigned to at least two conditions (e.g. treatment and comparison groups)

	program, rather than other factors, caused the reported outcomes.	• Did the sample meet the WWC standards for low levels of overall and differential attrition? • Were the sample members reassigned after random assignment was conducted? • Were there confounding factors (when on part of the design lined up exactly with either the treatment of comparison groups)? • Was there baseline equivalence of the treatment or comparison groups on selected measures? • Did the analysis include statistical adjustments for selected measures?
U.S. Department of Health and Human Services: Home Visiting Evidence of Effectiveness (HomVEE)		
Meets DHHS criteria for evidence-based program model	**Does not meet DHSS criteria for evidence-based program model**	
• At least one high- or moderate-quality impact study of the model finds favorable, statistically significant impacts in two or more of the eight outcome domains OR • At least two high- or moderate-quality impact studies of the model using non-overlapping analytic study samples with one or more favorable, statistically significant impacts in the same domain OR • At least five studies examining the intervention meet the WWC's pilot single-case design standards without reservations or standards with reservations (equivalent to a "high" or "moderate" rating in HomVEE, respectively)	• The programme does not meet the criteria	

• The single-case designs are conducted by at least three research teams with no overlapping authorship at three institutions. • The combined number of cases is at least 20							
RCT		**Matched Comparison**		**Single Case Design**		**Regression Discontinuity**	
High	**Moderate**	**High**	**Moderate**	**High**	**Moderate**	**High**	**Moderate**
Meets criteria SLR1-SRL6 below	Meets criteria SLR1-SLR6 except for either SLR2 or SLR3 below	Not applicable	Meets criteria SLR4-SLR6 below	Meets criteria SLC1-SLC4 below	Meets criteria SLC1-SLC4 below	Meets SLR2 and SLRD1-SLRD3 criteria below	Meets SLR2 SLRD1,and SLRD3 criteria below

Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Study Limitations	-	• Did the study randomly assign participants? • Did the sample meet the WWC standards for low levels of overall and differential attrition? • Were the sample members reassigned after random assignment was conducted? • Were there confounding factors (when on part of the design lined up exactly with either the treatment of comparison groups)? • Was there baseline equivalence of the treatment or comparison groups on selected measures)? • Did the study control for baseline imbalances?

		• Was timing of intervention systematically manipulated? • Did outcomes meet WWC standards for inter-assessor agreement? • Were there at least three attempts to demonstrate an effect? • Did the study have at least five (at least three for moderate rating) data points in relevant phases? • Was integrity of forcing variable maintained? • Was the relationship between the outcome and the forcing variable continuous? • Did the study meet WWC standards for functional form and bandwidth?

U.S. Department of Health and Human Services: Office of Adolescent Health	
Meets DHHS criteria for Effective Program	**Doesn't meet DHHS criteria for Effective Program**
• To meet this criterion, the program's supporting research study must show evidence of a positive, statistically significant impact on at least one priority outcome measure for either the full analytic sample or a subgroup defined by (1) gender or (2) sexual experience at baseline	• The programme does not meet the criteria

High quality study	**Moderate quality study**	**Low quality study**
Meets criteria SLR1 –SLR6 below	Meets criteria SLR4-SLR6	Doesn't meet criteria for high or moderate quality

Domain	***Definition***	***Criteria for the Domain (converted into signalling questions)***
Study Limitations	Quality of individual studies are defined in terms of the risk of bias in the study's impact estimates	• Did the study use random or functionally random assignment? • Did the sample meet the WWC standards for low levels of overall and differential attrition?

		• Were the sample members reassigned after random assignment was conducted? • Were there confounding factors (when on part of the design lined up exactly with either the treatment of comparison groups)? • Was there baseline equivalence of the treatment or comparison groups on selected measures)? • Did the study control for baseline imbalances?
Washington State Institute for Public Policy (WSIPP)		
Evidence-based practice	**Research-based practice**	**Promising practice**
• A program or practice that has been tested in heterogeneous or intended populations with multiple randomized, or statistically controlled evaluations, or both; or one large multiple site randomized, or statistically controlled evaluation, or both, where the weight of the evidence from a systemic review demonstrates sustained improvements in at least one outcome (p-value < 0.20). "Evidence-based" also means a program or practice that can be implemented with a set of procedures to allow successful replication in Washington and, when possible, is determined to be cost-beneficial (at least a 75%	• A program or practice that has been tested with a single randomized, or statistically controlled evaluation, or both, demonstrating sustained desirable outcomes; or where the weight of the evidence from a systemic review supports sustained outcomes (p-value < 0.20), but does not meet the full criteria for evidence-based	• A practice that, based on statistical analyses or a well-established theory of change, shows potential for meeting the evidence-based or research-based criteria, which may include the use of a program that is evidence-based for outcomes other than those of interest to WSIPPP (defining "evidence-based").

chance of a positive net present value)				
What Works Centre for Local Economic Growth				
Level 5	**Level 4**	**Level 3**	**Level 2**	**Level 1**
• Reserved for designs involving explicit randomisation into comparable treatment group • Randomisation is successful • Attrition carefully addressed or not an issue • Contamination not an issue	• Entails use of an instrument or discontinuity in treatment • Discontinuity in treatment is sharp (e.g. strict eligibility requirement) or fuzzy discontinuity method used (for RDD design) • Only treatment changes at boundary (for RDD design) • Behaviour is not manipulated to make the cut-off (for RDD design) • Instrument is relevant (explains treatment; for Instrumental Variables) • Instrument is exogenous (not explained by outcome; for Instrumental Variables)	• Control group would have followed same trend and treatment group (for DfD design) • There is known time period for treatment (for DfD design) • Fixed effect is at the unit of observation (for Panel Fixed Effects) • Year effects are included (for Panel Fixed Effects) • Appropriate time-varying controls are used (for Panel Fixed Effects) • Year effects are included (for First Differences)	• Good matching variables (i.e. relevant to selection; for Propensity Score Matching with Cross-Sectional method) • Significant common support (for Propensity Score Matching with Cross-Sectional method) • Adequate control variables are used (for Cross-Sectional Regression or Before-and-After method)	• Either (a) a Cross-Sectional comparison of treated groups with untreated groups, or (b) a Before-and-After comparison of treated group, without an untreated comparison group. No use of control variables in statistical analysis to adjust for differences between treated and untreated groups or periods

	• Instrument is excludable (does not directly affect outcome; for Instrumental Variable)	• Year effects are included (for Fixed Differences) • Appropriate time-varying controls are used (for Fixed Differences) • Matching criteria are satisfied (for Propensity Score Matching with DiD or Panel method) • DiD or panel criteria are satisfied (for Propensity Score Matching with DiD or Panel method)		

What Works Clearinghouse					
Positive	**Potentially positive**	**No discernible effects**	**Mixed**	**Potentially negative**	**Negative**
• Two or more studies* show statistically significant* positive effects • At least one of which meets WWC group design standards without reservations*	• At least one study shows statistically significant or substantively important positive effects • Fewer or the same number of studies show indeterminate effects than show	• None of the studies shows statistically significant or substantively important effects, either positive or negative	• At least one study shows statistically significant or substantively important positive effects • At least one study shows statistically	• Significant or substantively important negative effects • No studies show statistically significant or substantively important positive effects	• Two or more studies show statistically significant negative effects, at least one of which

• No studies show statistically significant or substantively important negative effects	statistically significant or substantively important positive effects • No studies show statistically significant or substantively important negative effects		significant or substantively important negative effects, BUT no more such studies than the number showing statistically significant or substantively important positive effects **OR** • At least one study shows statistically significant or substantively important effects • More studies show an indeterminate effect than show statistically significant or substantively	**OR** • Two or more studies show statistically significant or substantively important negative effects, at least one study shows statistically significant or substantively important positive effects • More studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects	meets WWC group design standards without reservations • No studies show statistically significant or substantively important positive effects

			important effects		
Domain	***Definition***		***Criteria for the Domain (converted into signalling questions)***		
Design	To be included in the review, studies need to have either an RCT or a quasi-experimental design.		• What was the study design? - randomised controlled trials (RCT) - quasi-experimental design		
Evidence of effect	A statistically significant estimate of effect		• Statistically significant: one for which the probability of observing such a results by chance is < than one in 20 (using a two-tailed-t-test with p=0.05) • Substantively important: an effect size of 0.25 standard deviations or larger • Indeterminate effect: one for which the single or mean effect is neither statistically significant nor substantively important		
Study limitations (referred to as "WWC group design standards" in the system)	• Assesses credibility of evidence from a single study. The three possible ratings are: • Meets WWC group design standards without reservations • Meets WWC group design standards with reservations • Does not meet WWC group design standards		• Was group assignment determined by a random process (design)? • In what range did the combination of overall and differential attrition fall (attrition)? • Were the baseline equivalence requirements met (is there evidence of baseline non-equivalence)? • Were the outcome data collected in the same way for both the treatment and experimental groups? • Did outcome measures demonstrate face validity and reliability? • Was outcome over-aligned with their intervention? • Did the study account for confounding factors?		

What Works Reentry Clearinghouse				
Strong beneficial	**Modest beneficial**	**No evidence of effect**	**Modest harmful**	**Strong harmful**
• The study findings show a consistent pattern indicating benefits in the area of interest in relation to the comparison group, and most or all of these findings are statistically significant	• The study findings are mixed: some findings indicate significant benefits in the area of interest than in relation to the comparison group, but other findings show no significant differences; or findings indicate that the treatment group experienced better outcomes on several measures, but these differences generally failed to reach statistical significance	• The study findings show very few or no significant differences between groups	• The study findings are mixed: some findings indicate significant benefits in the area of interest, but other findings show no significant differences; or findings indicate that the comparison group experienced better outcomes on several measures, but these differences generally failed to reach statistical significance	• The study findings show a consistent pattern indicating that the comparison group experienced better post-release outcomes in the area of interest than the treatment group, and most or all of these findings are statistically significant

Appendix 4. Evidence domains in the included systems

Specification of the evidence domains in the evidence rating systems included in the systematic review

Baral et al. (2012) **The Highest Attainable Standard of Evidence (HASTE)**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Efficacy data	N/D	N/D
Berkman et al. (2013) **Agency for Healthcare Research and Quality (AHRQ)**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Design	N/D	• What was the study design? - Randomised controlled trial (high quality) - Observational study (low quality)
Study Limitations in RCTs/CCTs	It refers to the judgment that the findings from included studies of a treatment (or treatment comparison) for a given outcome are adequately protected against bias (i.e., have good internal validity), based on the design and conduct of those studies.	• Was the allocation sequence generated adequately (e.g., random number table, computer- generated randomisation)? (i.e., no potential for selection bias) • Was the allocation of treatment adequately concealed (e.g., pharmacy-controlled randomisation or use of sequentially numbered sealed envelopes)? (i.e., no potential for selection bias) • Were participants analysed within the groups they were originally assigned to? (i.e., no potential for selection bias) • Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches? (i.e., no potential for selection bias)

		• Did researchers rule out any impact from a concurrent intervention or an unintended exposure bias that might bias results? (i.e., no potential for performance bias) • Did the study maintain fidelity to the intervention protocol? (i.e., no potential for performance bias) • If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)? (i.e., no potential for attrition bias) • In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome the same for cases and controls? (i.e., no potential for detection bias) • Were the outcome assessors blinded to the intervention or exposure status of participants? (i.e., no potential for detection bias)
		• Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants? (for CCTs only; i.e., no potential for detection bias) • Were the potential outcomes pre-specified by the researchers? Are all pre-specified outcomes reported? (i.e., no potential for reporting bias)
Study Limitations in Case-Control Studies	The same as for RCTs/CCTs	• Were cases and controls selected appropriately (e.g. appropriate diagnostic criteria or definitions, equal application of exclusion criteria to case and

		controls, sampling not influenced by exposure status)? (i.e., no potential for selection bias) • Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches? (i.e., no potential for selection bias) • Did researchers rule out any impact from a concurrent intervention or an unintended exposure bias that might bias results? (i.e., no potential for performance bias) • Did the study maintain fidelity to the intervention protocol? (i.e., no potential for performance bias) • If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)? (i.e., no potential for attrition bias) • In prospective studies, was the length of follow-up different between the groups, or in case-control studies, was the time period between the intervention/exposure and outcome the same for cases and controls? (i.e., no potential for detection bias) • Were the outcome assessors blinded to the intervention or exposure status of participants? (i.e., no potential for detection bias) • Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias)

		• Were the potential outcomes pre-specified by the researchers? Are all pre-specified outcomes reported? (i.e., no potential for reporting bias)
Study Limitations in Cross-Sectional Studies	The same as for RCTs/CCTs	• Did the study apply inclusion/exclusion criteria uniformly to all comparison groups? (i.e., no potential for selection bias) • Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches? (i.e., no potential for selection bias) • Did researchers rule out any impact from a concurrent intervention or an unintended exposure bias that might bias results? (i.e., no potential for performance bias) • If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)? (i.e., no potential for attrition bias) • Were outcome assessors blinded to the intervention or exposure status of participants? • Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were the potential outcomes pre-specified by the researchers? Are all pre-specified outcomes reported? (i.e., no potential for reporting bias)

Study Limitation in Case Series	The same as for RCTs/CCTs	• Does the design or analysis control account for important confounding and modifying variables through matching, stratification, multivariable analysis, or other approaches? (i.e., no potential for selection bias) • Did researchers rule out any impact from a concurrent intervention or an unintended exposure bias that might bias results? (i.e., no potential for performance bias) • Did the study maintain fidelity to the intervention protocol? (i.e., no potential for performance bias) • If attrition (overall or differential nonresponse, dropout, loss to follow-up, or exclusion of participants) was a concern, were missing data handled appropriately (e.g., intention-to-treat analysis and imputation)? (i.e., no potential for attrition bias) • Were outcome assessors blinded to the intervention or exposure status of participants? • Were interventions/exposures assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were outcomes assessed/defined using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were confounding variables assessed using valid and reliable measures, implemented consistently across all study participants? (i.e., no potential for detection bias) • Were the potential outcomes pre-specified by the researchers? Are all pre-specified outcomes reported? (i.e., no potential for reporting bias)

Consistency	Consistency refers to the degree of similarity in the direction of effects or the degree of similarity in the effect sizes (magnitudes of effect) across individual studies within an evidence base. EPCs may choose which of these two notions of consistency (direction or magnitude) they are scoring; they should be explicit about this choice.	• What was the consistency in direction of effect estimates in relation to the line that distinguishes superiority from inferiority (or minimally important difference for non-inferiority or equivalence)? - To what extent did confidence intervals overlap? • What was the consistency in the magnitude of effect? - Was the statistical test for heterogeneity (Cochran's Q test) significant? - Was the magnitude of statistical heterogeneity (as measured by I^2) large? - Did point estimates vary widely? Note: The consistency of a single-study evidence base is judged as unknown.
Directness	Directness of evidence expresses how closely available evidence measures an outcome of interest. Assessing directness has two parts: directness of outcomes and directness of comparisons. Applicability of evidence (external validity) is considered explicitly but separately from strength of evidence.	Was the included outcome an intermediate or a proxy of an ultimate health outcomes? • Did investigators use proxy respondents to stand in for certain kinds of patients or subjects in measuring the outcome of interest? • Were the conclusions based on direct (head-to-head) comparisons?
Precision	Precision is the degree of certainty surrounding an estimate of effect with respect to an outcome. It is based on the potential for random error evaluated through the sufficiency of sample size and, in the case of	• What was the width of the confidence interval around the pooled effect estimate? • Was the optimal information size criterion met? • What was the potential for random error in individual studies (specifically when a quantitative synthesis is not possible?

	dichotomous outcomes, the number of events.	
Reporting Bias	Reporting bias occurs when authors, journals, or both decide to publish or report research findings based on their direction or magnitude of effect. There are three main types of reporting bias that authors or journals can introduce: • Publication bias • Selective outcome reporting bias • Selective analysis reporting	• Did the authors conduct a quantitative assessment of the "missingness" of outcome data from small studies? - Tests of funnel plot asymmetry; trim and fill method; selection modeling • Did the authors conduct a qualitative assessment of the risk of reporting bias (considers 7 factors below)? - Estimated number of studies affected by reporting biases - Total sample size affected by reporting biases - Total number of studies in evidence base - Total number of participants in evidence base - Consistency of effect estimates across contributing studies - Study limitation of the evidence base - Comprehensiveness of study retrieval and identification
Large Magnitude of Effect	Strength of association refers to the likelihood that the observed effect is large enough that it cannot have occurred solely as a result of bias from potential confound factors. This additional domain should be considered when the effect size is particularly large.	• Was there large magnitude of effect?
Dose-Response Relationship	This association, either across or within studies, refers to a pattern of a larger effect with greater exposure (dose, duration, adherence). This domain should be considered when studies in the	• Was there a dose-response relationship between the intervention and the outcome?

	evidence base have noted levels of exposure.	
Plausible confounding	Occasionally, in an observational study, plausible confounding would work in the direction opposite that of the observed. This additional domain should be considered when plausible confounding exists that would decrease the observed effect.	• Would plausible confounding decrease the observed effect?
Briss et al. (2000) **The Guide to Community Preventive Services**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Design Suitability	Suitability of study design is characterised based on several characteristics that help to protect against a variety of potential threats to validity.	• Is the study design suitable to protect against a variety of potential threats to validity? - Concurrent comparison (greatest suitability) - Comparison, but not concurrent (moderate suitability) • Single group (least suitability)
Study Execution	Reviewers assess quality of study execution by considering six categories of threats to validity (see the criteria below).	**Description** • Was the study population (i.e. the intervention and comparison population) well described? • Was the intervention well described? **Sampling** • Did the authors specify (i.e. describe characteristics and size of) the sampling frame or universe of selection for the study population? • Did the authors specify the screening criteria for study eligibility (if applicable)? • Was the population that served as the unit of analysis the entire eligible population or a probability sample at the point of observation?

		• Are there other selection bias issues not identified above? **Measurement** • Was there an attempt to measure exposure to the intervention? • Were the exposure variables valid measures of the intervention under study? - Clear definition of the exposure variable - Measurement of exposure in different ways - Citations of discussion as to why the use of these measures is valid • Were the exposure variables reliable (consistent and reproducible) measures of the intervention under study? - Measures of internal consistency - Measurement of exposure in different ways - Inter-rater reliability checks - Citations or discussion as to why the use of these measures is reliable • Were the outcome and other independent (or predictor) variables valid measures of the outcome of interest? - Clear definition of the outcome variable - Measurement of the outcome in different ways - Citations or discussion as to why the use of these measures is valid • Were the outcome and other independent (or predictor) variables reliable (consistent and reproducible) measures of the outcome of interest? - Measures of internal validity - Measurement of the outcome in different ways - Considered consistency of coding, scoring or categorization between observers or between different outcome measures - Considered how setting and sampling of study population might affect reliability **Analysis** • Did the authors conduct appropriate analysis?

		- Conducting statistical testing? - Reporting which statistical tests were used? - Controlling for design effects in the statistical model? - Controlling for repeated measures in the analysis, for study designs in which the same population was followed with repeated measurements over time? - Accounting for different levels of exposure in segments of the study population in the analysis? - If the authors analysed group-level and individual-level covariates in the same statistical model, was the model designed to handle multi-level data? • Were there other problems with data analysis that limit interpretation of the results of the study? **Interpretation of results** • Did at least 80% of enrolled participants (i.e. intervention and comparison groups) complete the study? • Did the authors assess whether the units of analyses were comparable prior to exposure to the intervention? • Considering the study design, were appropriate methods for controlling confounding variables and limiting potential biases used? • Did the authors identify and discuss potential biases or unmeasured/contextual confounder that may account for or influence the observed results and explicitly state how they assessed these potential confounders and biases? **Other** • Are there other issues that limit your ability to interpret the results of the study that were not identified handled in one of the other categories?
Number of Studies	N/D	• How many studies contribute to the evidence base?
Consistency	N/D	• Were findings generally consistent in direction and size?

Magnitude of Effect	Sufficient and large effect sizes are defined on a case-by-case basis and are based on Task Force	• Was there large effect size?
Bruce et al. (2014) **Grading of Evidence for Public Health Interventions (GEPHI)**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Design	N/D	• What was the study design? - Randomised controlled trial (high quality) - Quasi-experimental [with controls] and before and after [uncontrolled] studies: this would be the case so long as the evidence from these studies can clearly be shown to be stronger in terms of minimising selection bias and confounding than other observational designs (moderate quality) • Observational study (low quality)
Analogy	Coherent evidence on the effect of similar environmental health interventions or exposures that operate through the same or a similar mechanism.	• Is there coherence evidence on the effect of similar environmental health intervention or exposures that operate through the same or a similar mechanism?
Consistency	Consistent evidence is found across a large number of settings, geographical locations and/or over time, and across diverse epidemiological study designs and/or gathered by different researchers.	• Is consistent evidence found across a large number of settings, geographical locations and/or over time, and across diverse epidemiological study designs and/or gathered by different researchers?
Coherence	Coherence of evidence contributing to the causal chain.	• Is there coherence in the body of evidence contributing to the causal chain?

Clark et al. (2009) **Let Evidence Guide Every New Decision (LEGEND)**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Study Quality for RCTs	The aggregate quality ratings for individual studies (including their design)	***Validity: are the results of the RCT valid or credible?*** • Were the patients randomly assigned to treatment and control groups? • Was that randomisation conducted appropriately? - Was the randomization concealed from those responsible for recruiting subjects? - Were patients, parents, clinicians, and analysts masked to which treatment was being received? • Were the groups similar at the start of the trial, with respect to known prognostic factors? • Aside from the experimental treatment, were the groups treated equally? • Were all patients who entered the trial accounted for at its conclusion? - Was there a low rate of attrition? (Note: if greater than 20% lost to follow up, bias may be of greater concern) • Were patients accounted for (and analysed) in the groups to which they were randomized (i.e. ITT analysis)? • Was the study long enough to fully study effects of the intervention? • Were instruments used to measure the outcomes valid and reliable? • Was there freedom from conflict of interest? ***Reliability: are these valid study results important?*** • Did the study have a sufficiently large sample size? - Was there a power analysis? - Did the sample size achieve or exceed that resulting from the power analysis? - Did each subgroup also have sufficient sample size (e.g. at least 6 to 12 participants)?

		• What were the main results of the RCT?
		  - What was the effect size? (How large was the treatment effect?)
		  - What were the measures of statistical uncertainty (e.g. precision)?
		• Were the results statistically significant?
		• Were the results clinically significant?
		  - If potential confounders were identified, were they discussed in relationship to the results?
		• Were adverse events assessed?
		Applicability: can I apply these valid, important study results to treating my patients?
		• Can the results be applied to my population of interest?
		  - Is the treatment feasible in my care setting?
		  - Do the patient outcomes apply to my population or question of interest?
		  - Are the likely benefits worth the potential harm and costs?
		  - Were the patients in this study similar to my population of interest?
		• Are my patient's and family's values and preferences satisfied by the treatment and its consequences?
		Would you include this study/article in development of a care recommendation?

Study Quality for Systematic Reviews and Meta-Analyses	The same as for RCTs	***Validity: are the results of the systematic review/meta-analyses valid or credible?***
		• Did the overview address a focused clinical question?
		• Was the search for relevant studies detailed and exhaustive?
		  - Was it unlikely that important, relevant studies were missed?
		• Did the systematic review use RCTs?
		  - Were the criteria used to select articles for inclusion appropriate?
		  - Was the assignment of patients to treatments randomized?

		• Were the included studies appraised and assigned a high level of quality? • Were the methods consistent from study to study? - Were populations among the included studies comparable and appropriate? - Were the outcomes, interventions, and exposures measured in the same way? • Was there freedom from conflict of interest? - Sponsor/Funding Agency or Investigators? ***Reliability: are these valid study results important?*** • What were the main results of the systematic review/meta-analysis? - What was the effect size? (how large was the treatment effect?) - What were the measures of statistical uncertainty (e.g. precision)? • Were the results statistically significant? • Were the results clinically significant? - If potential confounders were identified, were they discussed in relationship to the results? • Were adverse events discussed? ***Applicability: can I apply these valid, important study results to treating my patients?*** • Can the results be applied to my population of interest? - Is the treatment feasible in my care setting? - Do the patient outcomes apply to my population or question of interest? - Are the likely benefits worth the potential harm and costs? - Are the patients in this study similar to my population of interest? • Are my patient's and family's values and preferences satisfied by the treatment and its consequences?

		• Would you include this study/ article in development of a care recommendation?
Study Quality for Longitudinal Studies (e.g. time series)	Same as for RCTs	***Validity: are the results of the longitudinal study valid or credible?*** • Were the study methods appropriate for the question? - Were the study methods clearly described (e.g. setting, sample population)? - Were data collected at more than one point in time (i.e. before/after, pretest/posttest, time series)? • Were instruments used to measure the outcomes valid and reliable? - Were the instruments tested to be reliable? • Were all appropriate variables (e.g. potential confounders, exposures, predictors) and interventions clearly described? • Were all appropriate outcomes clearly described? • Was there freedom from conflict of interest? - Sponsor/Funding Agency or Investigators. ***Reliability: are these valid study results important?*** • Were the statistical analysis methods appropriate? - Were the statistical analysis methods clearly described? • Did the study have a sufficiently large sample size? - Was a power analysis described? - Did the sample size achieve or exceed that resulting from the power analysis? - Did each subgroup also have sufficient sample size (e.g., at least 6-12 participants)? • What were the main results of the study? - What was the effect size? - What were the measures of statistical uncertainty (e.g. precision? • Were the results statistically significant?

		• Were the results clinically significant?
		  - If potential confounders were identified, were they discussed in relationship to the results?
		• Were any adverse events assessed?
		Applicability: can I apply these valid, important study results to treating my patients?
		• Can the results be applied to my population of interest?
		  - Is the treatment feasible in my care setting?
		  - Do the patient outcomes apply to my population or question of interest?
		  - Are the likely benefits worth the potential harm and costs?
		  - Were the patients in this study similar to my population of interest?
		• Are my patient's and family's values and preferences satisfied by the treatment and its consequences?
		• Would you include this study/article in development of a care recommendation?

Study Quality for Cohort Studies	The same as for RCTs	***Validity: are the results of the cohort study valid or credible?***
		• Were the study methods appropriate for the question?
		  - Were the study methods clearly described (e.g. setting, sample population)?
		  - Were the instruments clearly described?
		  - Were the interventions clearly described?
		• Were the participants recruited prospectively with a comparison group?
		• Were instruments used to measure the outcomes valid and reliable?
		  - Were the instruments tested to be valid and reliable?
		• Were all appropriate variables (e.g. potential confounders, exposures, predictors) and interventions clearly described?
		• Were all appropriate outcomes clearly described?

		• Was the follow-up process described and complete?
		  - Was the follow-up long enough to fully study the effects of the intervention?
		  - Was there a low rate of attrition? (20%)
		• Was there freedom from conflict of interest?
		  - Sponsor/Funding Agency or Investigators.
		Reliability: are these valid study results important?
		• Were the statistical analysis methods appropriate?
		  - Were the statistical analysis methods clearly described?
		• Did the study have a sufficiently large sample size?
		  - Was a power analysis described?
		  - Did the sample size achieve or exceed that resulting from the power analysis?
		  - Did each subgroup also have sufficient sample size (e.g., at least 6-12 participants)?
		• What were the main results of the study?
		  - What was the effect size?
		  - What were the measures of statistical uncertainty (e.g. precision)?
		• Were the results statistically significant?
		• Were the results clinically significant?
		  - If potential confounders were identified, were they discussed in relationship to the results?
		• Were any adverse events assessed?
		Applicability: can I apply these valid, important study results to treating my patients?
		• Can the results be applied to my population of interest?
		  - Is the treatment feasible in my care setting?

		- Do the patient outcomes apply to my population or question of interest? - Are the likely benefits worth the potential harm and costs? - Were the patients in this study similar to my population of interest? • Are my patient's and family's values and preferences satisfied by the treatment and its consequences? • Would you include this study/article in development of a care recommendation?
Study Quality for Case-Control Studies	The same as for RCTs	***Validity: are the results of the case-control study valid or credible?*** • Were the study methods appropriate for the question? - Were the study methods clearly described (e.g. setting, sample population)? - Were cases and controls matched appropriately for confounders or comorbidities? - : Were appropriate numbers of control participants matched to the case participants? • Were instruments used to measure the outcomes valid and reliable? - Were the instruments tested to be reliable? • Were all appropriate variables (e.g. potential confounders, exposures, predictors) and interventions clearly described? • Were all appropriate outcomes clearly described? • Were all participants accounted for at the conclusion of the study? - Were missing data explained? • Was there freedom from conflict of interest? - Sponsor/Funding Agency or Investigators. ***Reliability: are these valid study results important?*** • Were the statistical analysis methods appropriate? - Were the statistical analysis methods clearly described?

		• Did the study have a sufficiently large sample size?
		  - Was a power analysis described?
		  - Did the sample size achieve or exceed that resulting from the power analysis?
		  - Did each subgroup also have sufficient sample size (e.g., at least 6-12 participants)?
		• What were the main results of the study?
		  - What was the effect size?
		  - What were the measures of statistical uncertainty (e.g. precision)?
		• Were the results statistically significant?
		• Were the results clinically significant?
		  - If potential confounders were identified, were they discussed in relationship to the results?
		• Were any adverse events assessed?
		Applicability: can I apply these valid, important study results to treating my patients?
		• Can the results be applied to my population of interest?
		  - Is the treatment feasible in my care setting?
		  - Do the patient outcomes apply to my population or question of interest?
		  - Are the likely benefits worth the potential harm and costs?
		  - Were the patients in this study similar to my population of interest?
		• Are my patient's and family's values and preferences satisfied by the treatment and its consequences?
		• Would you include this study/article in development of a care recommendation?

Study Quality for Cross-Sectional Studies	The same as for RCTs	***Validity: are the results of the case-control study valid or credible?*** • Were the study methods appropriate for the question? - Were the study methods clearly described (e.g. setting, sample population)? - Were the instruments clearly described? - Were the data collected at one point in time? • Were instruments used to measure the outcomes valid and reliable? - Were the instruments tested to be valid and reliable? • Were all appropriate variables (e.g. potential confounders, exposures, predictors) and interventions clearly described? • Were all appropriate outcomes clearly described? • Were all participants accounted for at the conclusion of the study? - Were withdrawals from the study explained? - Was the rate of attrition acceptable? • Was there freedom from conflict of interest? - Sponsor/Funding Agency or Investigators. ***Reliability: are these valid study results important?*** • Were the statistical analysis methods appropriate? - Were the statistical analysis methods clearly described? • Did the study have a sufficiently large sample size? - Was a power analysis described? - Did the sample size achieve or exceed that resulting from the power analysis? - Did each subgroup also have sufficient sample size (e.g., at least 6-12 participants)? • What were the main results of the study? - What was the effect size? - What were the measures of statistical uncertainty (e.g. precision)?

		• Were the results statistically significant?
		• Were the results clinically significant?
		  - If potential confounders were identified, were they discussed in relationship to the results?
		• Were any adverse events assessed?
		Applicability: can I apply these valid, important study results to treating my patients?
		• Can the results be applied to my population of interest?
		  - Is the treatment feasible in my care setting?
		  - Do the patient outcomes apply to my population or question of interest?
		  - Are the likely benefits worth the potential harm and costs?
		  - Were the patients in this study similar to my population of interest?
		• Are my patient's and family's values and preferences satisfied by the treatment and its consequences?
		• Would you include this study/article in development of a care recommendation?

Study Quality for Case Series		***Basic elements of a case report***
		• Does the case report fit into one of the categories expected?
		  - New associations or variations in disease processes
		  - Innovative approaches/interventions for treatment in disease processes
		  - Findings that shed light on the possible pathogenesis of a disease or an adverse event
		  - Presentations, diagnoses, or management of new and emerging diseases
		  - Unreported or unusual side effects or adverse interactions involving medications or treatment
		  - Unexpected or unusual presentations of a disease

		- Unexpected association between diseases and symptoms - Unexpected event in the course of observing or treating a patient - Does not fit one of the categories expected for case reports • Does the case report include a background of the issue on which the case report focuses? • Does the case report include an up-to-date review of the previous cases in the field? • Does the case report include details relevant to the case? - A description of the patient's demographic information - Any relevant medical history of the patient or their family - The patient's signs and symptoms - Any tests that were carried out - A description of pertinent details (e.g. disorder, medication, interventions/treatments, adverse events) - The outcome of the case • Is the importance of the case explained? - What can be learned from the case report? - How will the case report advance our clinical knowledge? - Other? • Was there freedom from conflict of interest? - Sponsor/Funding Agency or Investigators/Authors ***Applicability: can I apply these valid, important study results to treating my patients?*** • Can the results be applied to my population of interest? - Is the treatment feasible in my care setting? - Was the patient (or were the patients) in this report similar to my population of interest?

		- Do the patient outcomes apply to my population or question of interest? • Are my patient's and family's values and preferences satisfied by the treatment and its consequences? • Would you include this report in development of a care recommendation?
Consistency	The extent to which similar findings are reported using similar and different study designs.	• To what extent are similar findings reported using similar and different study designs?
Number of Studies	N/D	• How many studies contribute to the evidence base?
Magnitude of Effect	N/D	• Was there large effect size?
Department for International Development (DFID, 2014) How to Note: assessing the strength of evidence		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Quality	The (technical) quality of the studies constituting the body of evidence (or the degree to which risk of bias has been addressed).	Conceptual framing • Does the study acknowledge existing research? • Does the study construct a conceptual framework? • Does the study pose a research question or outline a hypothesis? Transparency • Does the study present or link to the raw data it analyses? • What is the geography/context in which the study was conducted? • Does the study declare sources of support/funding? Appropriateness • Does the study identify a research design? • Does the study identify a research method? • Does the study demonstrate why the chosen design and method are well suited to the research question? Cultural sensitivity

		• Does the study explicitly consider any context-specific cultural factors that may bias the analysis/findings? Validity • To what extent is the study internally valid? • To what extent is the study externally valid? • To what extent is the study ecologically valid? Reliability • To what extent are the measures used in the study stable? • To what extent are the measures used in the study internally reliable? • To what extent are the findings likely to be sensitive/changeable depending on the analytical technique used? Cogency • Does the author 'signpost' the reader throughout? • To what extent does the author consider the study's limitations and/or alternative interpretations of the analysis? Are the conclusions clearly based on the study's results?
Size	Size of the body of evidence	How many studies contribute to the evidence base?
Consistency	A range of studies pointing to identical or similar conclusions	Do a range of studies point to identical or similar conclusions?
Context	Context of the body of evidence	What is the context of the body of evidence?
Ebell et al. (2004) **Strength of Recommendation Taxonomy (SORT)**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Study Quality	The aggregate quality ratings for individual studies (including their design)	N/D
Consistency	N/D	• To what extent to the studies point to identical, or similar conclusions?

Gough et al. (2007) **Weight of Evidence: a framework for the appraisal of the quality and relevance of evidence**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Study Execution	This is a generic and thus non review specific judgement about the coherence and integrity of the evidence in its own terms.	N/D
Relevance of Design	This is a review specific judgement about the appropriateness of that form of evidence for answering the review question, that is the fitness for purpose of that form of evidence.	• Is the study design appropriate for answering the review question?
Relevance of Context/Focus of Evidence	This is a review specific judgement about the relevance of the focus of the evidence for the review question.	• Is the focus of evidence relevant for answering the review question (e.g. sample, analysis, context of the evidence, ethics of the research behind the evidence)?
Hillier et al. (2011) **FORM: an Australian method for formulating and grading recommendations in evidence-based guidelines**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Evidence base (Design)	The level of evidence indicates the study design used by the investigators to assess the effectiveness of an intervention. The level assigned to a study reflects the degree to which bias has been eliminated by the study design. Level of evidence reflects	• What level of study type (design) was used to answer the question? - Systematic review of RCTs (level I) - RCT (level II) - Pseudorandomised controlled trial (level III-1) - Comparative study with concurrent controls (level III-2) - Comparative study without concurrent controls (level III-3) • Case series with either post-test or pre-test/post-test outcomes (level-IV)

	the best study types for the specific type of question.	
Evidence base (Study Quality)	The quality of the evidence refers to the methods used by the investigators during the study to minimise bias and control confounding within a study type (i.e. how well the investigators conducted the study). Study quality relates to an assessment of the risk of bias inherent in the conduct, design and reporting of results in the included studies.	Method of treatment assignment • Did the study describe a correct, blinded randomisation method and document the group similarity? Control of selection bias after treatment assignment • Did the study adhere to intention-to-treat analysis and full follow-up of the sample? Blinding • Did the study report blinding of outcome assessors and patient and care givers? Outcome assessment (if blinding was not possible) • Did the study use standardised measures of assessment?
Evidence base (Quantity)	Quantity of evidence reflects the number of the studies that have been included as the evidence base for each guideline.	• How many studies contribute to the evidence base?
Consistency	The consistency component of the "body of evidence" assesses the extent to which the findings are consistent across the included studies (including across a range of study populations and study designs). This allows users to assess whether the results are likely to be replicable or only likely to occur under certain conditions.	• To what extent are the findings consistent across the included studies? • Was the magnitude of statistical heterogeneity (as measured by e.g. I^2) large? • Was the direction of effect across multiple studies consistent?

Clinical Impact	Clinical impact is a measure of the likely benefit that application of the guideline would have across the target population, and involves a clinical judgement.	• Was evidence base relevant to the clinical question? • Was the treatment effect statistically significant (low *P*-value)? • Was the size (magnitude) of the treatment effect clinically important (did the confidence interval include a clinically important effect)? • What is the relevance of the effect to patients, compared to other management options? • What is the duration of therapy required to achieve the effect? • What is the balance of risks of benefits to the patient group, including potential harms?
Applicability	Applicability addresses whether the evidence base is relevant to the Australian health care system generally, or to more local settings for specific recommendations (such as rural areas or cities).	• Is the evidence base applicable in terms of organizational factors (e.g. availability of trained stud, clinic time, specialized equipment, tests or other resources)? • Is the evidence base applicable in terms of cultural factors (e.g. attitudes to health issues, including those that may effect compliance with the recommendation)?
Generalisability	The assessment of generalisability involves determining how precisely the available body of evidence answers the clinical question that was asked.	• How well do the participants of the included studies match the patient population being targeted by the guideline? • How well do the clinical settings of the included studies match the settings where the recommendation will be implemented? • How well do the stages of disease considered in the included studies match those being targeted by the guideline? • How well does the duration of illness considered in the included studies match that being targeted by the guideline? • How well does the prevalence of the disease considered in the population of included studies match that being targeted by the guideline?

Joanna Brigs Institute (2014) **Levels of evidence and grades of recommendations**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Design	N/D	• What level of study type (design) was in the evidence base? - Experimental Designs (level 1) - Quasi-experimental Designs (level 2) - Observational – analytic designs (level 3) - Observational – descriptive studies (level 2) • Expert opinion and bench research (level 1)
Johnson et al. (2015) **Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Methodological adequacy of evidence in terms of estimating effect sizes	Assessing the level of bias in the estimates of mean effect sizes	• Did the review use transparent well-designed search strategy? • Did the authors assess the influence of study design by means of moderator analysis? • Did the authors pay sufficient attention to the validity of the constructs, with only comparable outcomes combined, and/or exploration of the implications of combining outcome constructs? • Did the authors assess the influence of unanticipated outcomes or spin-offs on the size of the effect? • Did the authors assess publication bias? • Did the authors consider inter-coder reliability? • Did the authors consider the influence of statistical outliers?
National Institute for Health and Care Excellence (NICE, 2012) **Methods for the development of NICE public health guidance**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Study Quality	Internal (that is, to check if potential sources of bias have	Population

	been minimised and to determine if its conclusions are open to any degree of doubt) ; External (i.e. the extent to which the findings for the study participants are generalisable to the whole 'source population' (that is, the population they were chosen from).	• Is the source population or source area well described? Was the country, setting, location, population demographics, etc. adequately described? • Is the eligible population or area representative of the source population or area? Was the recruitment of individuals, clusters or areas well defined? Was the eligible population representative of the source? Were important groups under-represented • Do the selected participants or areas represent the eligible population or area? Was the method of selection of participants from the eligible population well described What % of selected individuals or clusters agreed to participate? Were there any sources of bias? Were the inclusion or exclusion criteria explicit and appropriate? Method of allocation • How was selection bias minimised? Was allocation to exposure and comparison randomised? Was it truly random, or pseudo-randomised (e.g. consecutive admissions)? If not randomised, was significant confounding likely or not? If a cross-over, was order of intervention randomised? • Were interventions (and comparisons) well described and appropriate? Were interventions and comparisons described in sufficient detail (i.e. enough for study to be replicated)? Was comparisons appropriate (e.g. usual practice rather than no intervention)? • Was the allocation concealed? Could the person(s) determining allocation of participants or clusters to intervention or comparison groups have influenced the allocation? • Were participants or investigators blind to exposure and comparison? Were participants and investigators – those delivering or assessing the intervention kept blind to intervention allocation? • Was the exposure to the intervention and comparison adequate? Is reduced exposure to intervention or control related to the intervention (e.g. adverse effects leading to reduced compliance) or fidelity of

		implementation (e.g. reduced adherence to protocol)? Was lack of exposure sufficient to cause important bias? • Was contamination acceptably low? Did any in the comparison group receive the intervention or vice versa? If so, was it sufficient to cause important bias? If a cross-over trial, was there a sufficient wash-out period between interventions? • Were other interventions similar in both groups? Did either group receive additional interventions or have services provided in a different manner? Were the groups treated equally by researchers or other professionals? Was this sufficient to cause important bias? • Were all participants accounted for at study conclusion? Were those lost-to-follow-up (i.e. dropped or lost pre-, during or post- intervention) acceptably low (i.e. typically <20%)? Did the proportion dropped differ by group? • Did the setting reflect usual UK practice? Did the setting in which the intervention or comparison was delivered differ significantly from usual practice in the UK? • Did the intervention or control comparison reflect usual UK practice? Did the intervention or comparison differ significantly from usual practice in the UK? Outcomes • Were outcome measures reliable? Were outcome measures subjective or objective? How reliable were outcome measure? Was there any indication that measures had been validated? • Were all outcome measurements complete? Were all or most study participants who met the defined study outcome definitions likely to have been identified?

		• Were all important outcomes assessed? Were all important benefits and harms assessed? Was it possible to determine the overall balance of benefits and harms of the intervention versus comparison? • Were outcomes relevant? Where surrogate outcome measures were used, did they measure what they set out to measure? • Were there similar follow-up times in exposure and comparison groups? • Was follow-up time meaningful? Was follow-up long enough to assess long-term benefits or harms? Was it too long, e.g. participants lost to follow-up? Analyses • Were exposure and comparison groups similar at baseline? If not, were these adjusted? Were there any differences between groups in important confounders at baseline? If so, were these adjusted for in the analyses (e.g. multivariate analyses or stratification). Were there likely to be any residual differences of relevance? • Was intention to treat (ITT) analysis conducted? Were all participants (including those that dropped out or did not fully complete the intervention course) analysed in the groups (i.e. intervention or comparison) to which they were originally allocated? • Was the study sufficiently powered to detect an intervention effect (if one exists)? Is a power calculation presented? If not, what is the expected effect size? Is the sample size adequate? • Were the estimates of effect size given or calculable? Were effect estimates (e.g. relative risks, absolute risks) given or possible to calculate? • Were the analytical methods appropriate? Were important differences in follow-up time and likely confounders adjusted for? If a cluster design, were analyses of sample size (and power), and effect size performed on clusters (and not individuals)? Were subgroup analyses pre-specified? • Was the precision of intervention effects given or calculable? Were they meaningful? Were confidence intervals or *P*-values for effect estimates

		given or possible to calculate? Were CI's wide or were they sufficiently precise to aid decision-making? If precision is lacking, is this because the study is under-powered? Summary • Are the study results internally valid (i.e. unbiased)? How well did the study minimise sources of bias (i.e. adjusting for potential confounders)? Were there significant flaws in the study design? • Are the findings generalisable to the source population (i.e. externally valid)? Are there sufficient details given about the study to determine if the findings are generalisable to the source population?
Quantity	N/D	• How many studies contribute to the evidence base?
Consistency	N/D	• How consistent were the findings?
Sawaya et al. (2007) **U.S. Preventive Services Task Force (USPSTF)**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Design	N/D	• Do the studies have the appropriate research design to answer the key questions (i.e., different linkages in the causal pathway/analytic framework)?
Study Quality for Systematic Reviews	To what extent are the existing studies (comprising the body of evidence) of high quality? (i.e. what is the internal validity)?	• Was the used sources/search strategy comprehensive? • Did the review appraise included studies? • Were conclusions valid? • Is review relevant and recent?
Study Quality for Randomized Controlled Trials (RCTs)	The same as for Systematic Reviews	• Did the study adequately randomise participants, including allocation concealment, and were confounders equally distributed among groups? • Did the study maintain comparable groups throughout (attrition, crossover, adherence, contamination? • Was there important differential loss to follow-up or overall high loss to follow-up?

		• Did the study employ reliable and valid measures (including masking of outcome assessment)? • Were interventions clearly described? • Were all important outcomes considered? • Did study adjust for potential confounders for cohort studies, or employ intention-to-treat analysis?
Study Quality for Case-Control Studies	The same as for Systematic Reviews	• Were cases accurately ascertained? • Did the study employ nonbiased selection of cases/controls with exclusion criteria applied equally to both? • What was the response rate in the sample? • Were diagnostic testing procedures applied equally to each group? • Did study pay appropriate attention to potential confounding variables?
Quantity (Magnitude of Effect)	N/D	• How many studies have been conducted that address the key question(s) (i.e. different linkages in the causal pathway/analytic framework)? How large are the studies? (i.e., what is the precision of the evidence?)
Generalisability	N/D	• To what extent are the results of the studies generalisable to the general U.S. primary care population? (i.e., what is the external validity?)
Consistency	N/D	• How consistent are the results of the studies?
Dose-Response		• Are there additional factors that assist us in drawing conclusions (e.g., presence or absence of dose–response effects, fit within a biologic model)?
Other	N/D	N/D
Tang et al. (2007) **Grading of evidence of the effectiveness of health promotion interventions**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***

Association	To grade evidence, it is imperative to find out whether or not an intervention works, e.g. high and presumably statistically significant association between the intervention and the outcome factors, such as indicated by a relative risk and its confidence interval.	• Was there large magnitude of effect? - RR of 2 of more • Was the effect statistically significant?
Repeatability	It is then important to find out whether the intervention is widely repeatable, e.g. in different countries and settings. This reflects the consistency of the findings in different studies.	• Was the intervention widely repeatable in different countries and settings?
How it works	It is also important to find out how it works - the theoretical basis for making an association between the intervention and the outcome factors. If the theoretical basis is not known, the strength of evidence will be less convincing.	• Does the intervention have a known theoretical basis?
Treadwell et al. (2006) **A system for rating the stability and strength of medical evidence**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Study Quality	Quality of evidence for a specific outcome. Although quality evaluation can be performed with	N/D

	a checklist or scale, any reasonable method for separating the evidence base into different categories of quality will suffice.	
Quantity	Number of studies for each outcome.	• Were there at least 3 studies for the outcome of interest? • Did a certain percentage of studies (e.g. 80% or more) have calculable effect sizes (that can be determined without imputation)?
Informativeness	This use of "informativeness" accounts for the statistical power of the evidence base	• Is the treatment beneficial? • Is the treatment effect clinically important (i.e., the lower 95% confidence interval around the meta-analytic summary statistic is greater than the effect size deemed clinically important [decided a priori])?
Homogeneity	Statistical homogeneity testing (using fixed-effects model). This is the same as statistical consistency.	• Was the *P*-value for the Q statistic and I^2 less than a priori defined thresholds (e.g. .10 and 50%, respectively)? • If heterogeneity was detected, was a meta-regression conducted to explain heterogeneity (if sufficient studies)?
Robustness	One tests robustness through sensitivity analysis. The decision on the when to stop should be decided a priori.	• Did confidence intervals of the last three cumulative random-effects meta-analyses remain fully on the same side of zero after (a) removal of the study with the smallest weight, (b) the additional removal of the study with the second smallest weight in the meta-analysis.
Turner-Stokes (2006) **Generating the evidence base for the national service framework for long-term conditions: a new research typology**		
Domain	***Definition***	***Criteria for the domain (converted into signalling questions)***
Design	N/D	• What is the type of the evidence? - Primary research-based (quantitative, qualitative, mixed-methods) - Secondary research-based (meta-analysis; secondary analysis) • Review-based (systematic review; descriptive or summary reviews of existing research)

Study Quality	Quality is assessed on the bases of five questions to reach a maximum score of 10.	• Are the research question/aims and design clearly stated? • Is the research design appropriate for the aims and objectives of the study? • Are the methods clearly described? • Does the report show a statistically significant and clinically important treatment effect or, for a negative conclusion, have high power? • Are the results generalisable?
Applicability	Population context of the study.	• What is the population context (people with long-term neurological conditions) of the evidence?

Appendix 5. Interview guide

Time of interview:

Date:

Interviewer:

Interviewee ID:

Review title:

Stage 1. Introduction and Consent

Hello. Thank you so much for agreeing to participate in this interview. Although we have already had an email correspondence, may I briefly introduce myself?

My name is Ani, and I am a DPhil (PhD) student at the University of Oxford working on the project that aims to advance the GRADE guidance for rating the certainty in the estimates of effect of social and public health interventions. We conceptualise these interventions as those that operate via psycho-social and behavioural processes at either individual or population levels targeting multiple health and social outcomes. Often-times these are referred to as complex interventions, as they may involve multiple components, outcomes and delivered in complex systems. Our project has been driven by literature on challenges of using GRADE and its specific criteria in reviews of these interventions. Specifically, we have that the use of GRADE in reviews of complex public health interventions frequently results in downgrading the best evidence possible for these interventions. At this stage we are conducting semi-structure interviews mainly with Cochrane and Campbell review authors [GRADE methodologists] regarding their experiences of using GRADE and any suggestions they might have for advancing the GRADE guidance. This will then feed into our next project phases, which include a Delphi panel and a meeting to finalise the results and operationalise the new guidance. Does this sound ok?

Before we continue, I need to go over several ethical procedures and the consent script to make sure you understand what's involved for you:

- The interview is designed to last not more than 1 hour, and it will be looking at your experiences and views of using the GRADE approach in your review. I may also contact you via email for follow-up questions and further clarifications.
- The interview covers a methodological topic and does not involve any risks.
- The collected data will be used for my doctoral thesis and project publications either in the form of analysed data or anonymised quotes.
- I would also like to confirm that an Oxford University Research Ethics committee has approved this research project and I have provided information on how to contact me (in the first instance) or the committee in case of any concerns or complaints, as well as the project's ethics reference number and relevant contact details. I have previously included this information in the participant sheet in my email, but I will be happy to resend it to you again.
- With your permission, I would like to audio record the interviews for this research, to facilitate data analysis. All collected data will be anonymised for publication purposes and mainly used to inform the questions for the subsequent Delphi panel.
- Finally, as a token of our appreciation, we would like to offer £100 amazon email voucher to you for your time and participation, in thanks to the support from the grant that we received for this project.

[*Await confirmation]* Might you have any questions at this point? So if you're happy with all of that, and have no more questions, let's start.

Stage 2: Interview

1. To start with, how familiar are you with GRADE or the work of the GRADE group? Have you used GRADE before in a guideline or a systematic review?

2. What do you generally think of GRADE (or Summary of Findings tables more specifically) and its approach for rating the certainty in the effect estimates of complex interventions for each outcome?

   PROBES:

   2.1 Do you think that this approach and the definition of quality of evidence that it suggests is relevant and useful for complex interventions? How would define complexity in systematic reviews? What sources of complexity are particularly challenging to address in systematic reviews and GRADE ratings?
   2.2 What if we don't pool data into average effect sizes? Do you think that the GRADE approach and guidance is adequate for narrative synthesis?
   2.3 Many argue that the effectiveness or effects of public health interventions are critically influenced by modes of intervention delivery and contextual issues. Do you think that this has implications on how we should define certainty in the effects / estimates of effect of these interventions?
   2.4 Can you explain how we should go about defining certainty of evidence in this case (e.g., confidence that the effects are meaningful across a range of implementation contexts)?

3. The GRADE approach suggests to start the evidence rating process by looking at the study design of a body of evidence contributing to the specific outcome. If the evidence for a specific outcome predominantly comes from studies with an RCT design, the overall judgment for study design for that evidence should be "high certainty". If the evidence for a specific outcome predominantly comes from studies with an observational (non-RCT) design, the overall judgment should be "low certainty".
   What do you think of this evidence hierarchy within GRADE?
   Might you have any challenges or concerns regarding how GRADE treats different study designs?

   PROBES:

   3.1 There are arguments that many public health interventions, especially those targeting entire communities or populations are not conducive to RCTs, and therefore use of GRADE results in downgrading the best evidence possible for

these interventions. Do you think this is a relevant concern? Could you, please elaborate on your point?

3.2 A related challenge that has been much discussed is that GRADE does not differentiate between different types of observational studies, some of which are regarded as more rigorous, such as ITS designs/ self-controlled case series. Many say that GRADE unfairly penalises quasi-experimental studies. What are your views on this?

3.3 Should the GRADE approach adopt a different evidence hierarchy approach for complex interventions? Can you give an example of how that might look like?

4. After study design consideration, the GRADE approach considers five evidence domains to downgrade evidence, and additional three domains to upgrade evidence from observational studies. The five domains include: risk of bias, inconsistency, indirectness, imprecision and publication bias.
   Have you encountered any specific challenges with regards to any of these domains in this review?

5. Do you think that the GRADE's criteria for risk of bias assessment are adequate for these complex interventions?

   PROBES:

   5.1 GRADE has its own criteria for assessing risk of bias, which currently overlap with that of the Cochrane RoB tool for RCTs.
   Do you think this Cochrane risk of bias assessment approach should reflect that of the GRADE approach?
   What if a review does not use the Cochrane RoB tool, but another approach instead? Do you think this might complicate risk of bias assessment in GRADE?

   5.2 Even if we look at the Cochrane RoB model, do you think it is adequate enough for these complex interventions?
   Much concern has also been raised with regards to the participant/personnel blinding criterion when impossible to blind in many behavioural/public health interventions? Do you think this is an issue that should be considered or addressed?
   Some also argue that it lacks important considerations, such as intervention fidelity. Do you think this should be part of RoB assessment?

   5.3. What about risk of bias assessment criteria for non-randomised studies? Do you think these are adequate? Do you have any suggestions on this? What do you think about risk of bias assessment of quasi-experimental studies, such as ITS, regression discontinuity designs, etc.?
   E.g., if review authors used a rigorous and tailored approach to assess RoB in these quasi-experimental studies, do you think in that case these study designs could start the GRADE assessment at the same level as RCTs, that is high rather than low?

6. Did you encounter any challenges with regards to inconsistency domain?
   Do you have any suggestions to improve guidance on this domain?

   PROBES:

   6.1 Many argue that there are naturally high levels of inconsistency in complex public health interventions, and therefore most frequently evidence is downgraded because of this. Do you see this as being a relevant argument?
   6.2 Do you think these concerns might be related to how we frame review questions?
   6.3 GRADE states that if we manage to explain heterogeneity through subgroup analyses, then we may not downgrade evidence. If subgroup is not possible, how do you think we can explain heterogeneity?

7. There are also arguments that indirectness does not adequately reflect applicability/generalisability of evidence? Did you encounter any challenges on this domain? Do you have any views or suggestions on this?

   PROBES:

   7.1 What about intervention fidelity issue? Do you think this consideration is lacking in GRADE when thinking about complex public health interventions, and how/in which domain should this be operationalised? (performance/detection bias or indirectness of evidence).
   7.2 If interventions are similar in design, but differ in dosage/implementation, how shall we proceed with this assessment, e.g., should we rate evidence down in these cases?

8. What about imprecision or publication bias? Did you encounter any specific challenges on these domains? Do you have any specific suggestions?

9. The GRADE approach also suggests three domains to upgrade evidence from observational studies, that is the dose-response relationship, the magnitude of effect and counteracting confounding.
   Do these seem sufficient for upgrading our confidence in the effects of complex public health interventions?
   What other considerations might be relevant in this regard for public health interventions?

PROBES:

9.1 Many argue that, while GRADE allows for the downgrading of a body of evidence for inconsistency, consistency in findings across different settings, study designs and research groups actually increases our confidence in the evidence, as already stated by Sir Austin Bradford Hill. Therefore, suggestions are made to add consistency/ or analogy (which extends the concept of consistency to parallel evidence from e.g., related interventions, population groups) as domains for potential upgrading of evidence. Do you see suggestion this as relevant?

10. Finally, do you think there is a need for any additional evidence domain in the GRADE approach?

PROBES:

10.1 Many highlight the need for further guidance for appropriate selection of a body of evidence, which would balance internal and external validities. E.g., it is unclear whether very limited RCT evidence is sufficient or not, as including non-RCT evidence will have an impact on how the quality of evidence is rated. In the extreme case, can one well-conducted and large RCT conducted in one setting and rated as high or moderate-quality be sufficient to decide on the effectiveness of a complex intervention that critically relies on context?
10.2 What about further guidance for narrative synthesis? Would you view guidance for narrative synthesis as separate from guidance or adaptations of the GRADE approach for complex interventions?

Stage 3: Ending of Interview

Do you have any finals thoughts, comments or suggestions?

Stage 4: After the Interview

I think we have covered all the questions I aimed to discuss with you. I would like to thank you once again for your time and contribution to this project. Just to brief you about our further plans, once we finish the interviews with the review authors we aim to launch a Delphi panel (currently planned for mid-March), which will be followed by a meeting of a smaller group of stakeholders in May, where we will discuss and aim to finalise the content for the guidance. If you are interested, I can add you to the list of the Delphi participants, and send you the outputs of this project when ready (perhaps early next year).

Appendix 6. COREQ checklist

Consolidated criteria for reporting qualitative studies (COREQ): 32-item checklist[10]

Topic	Item No.	Guide questions/description	Reported on Page No.
Domain 1: research team and reflexivity			
Personal characteristics			
Interviewer/facilitator	1	Which author/s conducted the interview or focus group?	P. 149
Credentials	2	What were the researcher's credentials? E.g., PhD, MD	P. 149
Occupation	3	What was their occupation at the time of the study?	P. 149
Gender	4	Was the researcher male or female?	P. 84 P. 149
Experience and training	5	What experience or training did the researcher have?	P. 148
Relationship with participants			
Relationship established	6	Was a relationship established prior to study commencement?	P. 148
Participant knowledge of the interviewer	7	What did the participants know about the researcher? E.g., personal goals, reasons for doing the research	P. 148
Interviewer characteristics	8	What characteristics were reported about the interviewer/facilitator? E.g., bias, assumptions, reasons and interests in the research topic	P. 84
Domain 2: study design			
Theoretical framework			
Methodological orientation and theory	9	What methodological orientation was stated to underpin the study? E.g., grounded theory, discourse analysis, ethnography, phenomenology	P. 150
Participant selection			
Sampling	10	How were participants selected? E.g., purposive, convenience, consecutive, snowball	P. 147-148
Method of approach	11	How were participants approached? E.g., face-to-face, telephone, mail, email	P. 149

[10]The checklist has been developed from:
Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care. 2007;19(6):349–357.

Sample size	12	How many participants were in the study?	P. 151
Non-participation	13	How many people refused to participate or dropped out? Reasons?	P. 151
Setting			
Setting of data collection	14	Where was the data collected? E.g., home, clinic, workplace	P. 149
Presence of non-participants	15	Was anyone else present besides the participants and researchers?	P. 149
Description of sample	16	What are the important characteristics of the sample? E.g., demographic data	P. 151
Data collection			
Interview guide	17	Were questions, prompts, guides provided by the authors? Was it pilot tested?	P. 149, Appendix 8
Repeat interviews	18	Were repeat interviews carried out? If yes, how many?	N/A
Audio/visual recording	19	Did the research use audio or visual recording to collect the data?	P. 150
Field notes	20	Were field notes made during and/or after the interview of focus group?	N/A
Duration	21	What was the duration of interviews or focus groups?	P. 149
Data saturation	22	Was data saturation discussed?	P. 178
Transcripts returned	23	Were transcripts returned to participants for comment and/or correction?	P. 179
Domain 3: analysis and findings			
Data analysis			
Number of data coders	24	How many data coders coded the data?	P. 179
Description of the coding tree	25	Did authors provide a description of coding tree?	P. 150
Derivation of themes	26	Were themes identified in advance or derived from the data?	P. 150
Software	27	What software, if applicable, was used to manage the date?	P. 150
Participant checking	28	Did participants provide feedback on the findings?	P. 179
Reporting			
Quotations presented	29	Were participant quotations presented to illustrate the themes/findings? Was each quotation identified? E.g., participant number	Pp. 152-170, Table 4.3
Data and findings consistent	30	Was there consistency between the data presented and the findings?	Pp. 151-170, Table 4.3
Clarity of major themes	31	Were major themes clearly presented in the findings?	Pp. 152-170, Table 4.3
Clarity of minor themes	32	Is there a description of diverse cases or discussion of minor themes?	Pp. 152-170, Table 4.3

Appendix 7. Round One online expert panel questionnaire Panel A

Page 1

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Background Information

Please find brief background information on "complex interventions" and "rating the quality of the body of evidence" below to assist you in participating in this online panel.

Complex Interventions

The "complexity" of an intervention is often understood through an assessment of one or more of the following characteristics:

- Number of interacting intervention components
- Number and difficulty of behaviours involved in the intervention delivery and receipt
- Number of groups or organisational levels targeted by the intervention
- Number and variability of outcomes
- Flexibility, tailoring or non-standardisation of intervention implementation

Complexity is also increasingly understood through an assessment of the dynamic properties of the context (or "system") into which an intervention is introduced, such as non-linear relationships, feedback loops, phase changes, emergent properties, and interdependencies.

Rating the Quality of a Body of Evidence

A "body of evidence" refers to the totality of evidence contributing to an estimate of the comparative effect of an intervention for a specific outcome in a systematic review.

A rating of the "quality" of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

"Approaches to rating the quality of a body of evidence" typically assign an initial quality rating and then consider reasons to possibly downgrade or upgrade this quality rating.

Page 2

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Round One Instructions

For a tutorial on Round One, please read the instructions below and watch this video on an example "Round One" panel in ExpertLens.

Your Task for This Round

On the following pages, you will see lists of criteria that *could* be considered when rating the quality of a body of evidence in reviews of complex interventions. As a stakeholder in complex intervention research, we want you to rate the importance of considering each criterion for *all* reviews of complex interventions.

The criteria are organised as follows:

- Criteria related to an initial rating for the quality of the body of evidence (Page 2)
- Criteria related to downgrading the initial rating for the quality of the body of evidence (Pages 3-8)
- Criteria related to upgrading the initial rating for the quality of the body of evidence (Page 9)

Using the Rating Scale and Comment Box

We will ask you to rate each criterion on a scale from "1" ("Lower Importance") to "9" ("Higher Importance"). We will interpret your ratings as follows:

- Scores of 1-to-3 indicate that you believe a criterion is *of limited importance* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 4-to-6 indicate that you believe a criterion is *important but not critical* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 7-to-9 indicate that you believe a criterion is *critically important* to consider when rating the quality of a body of evidence in reviews of complex interventions

We encourage you to consider using the full range of the rating scale (from "1" to "9") across criteria and to use the comment boxes to briefly clarify your ratings.

We ask that you respond to all questions to Round One by 10:00AM (PT) on April 5.

Page 3

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Initial Quality Rating: Study Design

The criteria below relate to the initial "quality" rating for a body of evidence based on the design of the studies included in that body of evidence.

We want you to rate the importance of each criterion for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Randomised experimental studies

An initial quality rating of "high" when the body of evidence consists of randomised controlled trials (RCTs)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Non-randomised experimental studies

An initial quality rating of "moderate" when the body of evidence consists of non-randomised experimental study designs (e.g., natural experiments, quasi-experimental studies)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Non-experimental observational studies

An initial quality rating of "low" when the body of evidence consists of non-experimental observational studies (e.g., cohort design, case-control study)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Same rating across all study designs

An initial rating of "high" for a body of evidence consisting of any type of study design

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 4

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Randomised Trials

This page lists criteria related to methodological limitations of randomised controlled trials (RCTs) included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. 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.)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Blinding intervention providers

Whether providers of the intervention are aware of the arm to which participants have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Blinding intervention recipients

Whether recipients of the intervention are aware of the arm to which they have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Blinding outcome assessors

Whether those assessing outcomes are aware of the arm to which participants have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Blinding data analysts

Whether those analysing data are aware of the arm to which participants have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. 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

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Selective outcome reporting

Whether there is incomplete or absent reporting of some outcomes and not others on the basis of the results

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Stopping study early for benefit

Whether studies have been ended early when beneficial results were found in interim analyses

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Quality of outcome measures

Whether studies used outcome measures with low validity and/or reliability to assess intervention effects

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

10. Fidelity of intervention implementation

Whether there were deviations in intervention implementation from what was intended

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 5

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Non-Randomised Studies

This page lists criteria related to methodological limitations of non-randomised studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Confounding

Whether the study used measures to adequately control for confounding

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Appropriate comparison group

Whether the study developed and applied an appropriate comparison group

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Selection of participants into the study

Whether the study used procedures to appropriately select participants into the study or into the analysis

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Classification of interventions

Whether intervention groups were clearly defined

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Deviations from intended interventions

Whether there were deviations from the intended intervention beyond what would be expected in usual practice

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Missing data

Whether the study used appropriate analytic methods for dealing with participant missing data

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Measurement of outcomes

Whether the study appropriately measured outcomes in all groups

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Selection of the reported result

Whether there is selected reporting of effect estimates based on multiple measurements and analyses

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Follow-up

Whether the study had adequate follow-up of participants

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 6

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Inconsistency of Effects in the Evidence Base

This page lists criteria related to differences in intervention effect estimates across studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Variability in point estimates

The degree to which point estimates vary across individual studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence intervals

The degree to which confidence intervals overlap across individual studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. 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

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. 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 (chance)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. 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)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Qualitative analyses exploring heterogeneity

Results of qualitative analyses of evidence exploring varying effects of interventions that help explain heterogeneity (e.g., qualitative comparative analysis)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 7

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Indirectness of the Evidence Base

This page lists criteria related to the applicability of the body of evidence to the populations, interventions, outcomes, and settings of interest.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Indirectness of study populations

Degree to which the participants in included studies compare to the population of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Indirectness of study interventions

Degree to which the interventions in included studies compare to the intervention of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Indirectness of outcomes

Degree to which the outcomes considered in included studies compare to the outcomes of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. 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

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Indirectness of comparisons

Degree to which effect estimates are from comparison groups of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 8

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Imprecision of the Effect Estimates

This page lists criteria related to the width of confidence intervals for intervention effect estimates.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Optimal information size

Whether the total number of participants in the review meets a conventional sample size for a single adequately powered trial

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence interval with line of no effect

Whether the confidence interval for the overall estimate includes effects indicating both benefit and harm

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Width of confidence interval

Whether the confidence interval includes estimates of important benefit and important harm

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 9

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Publication Bias

This page lists criteria related to the systematic under-estimation or over-estimation of underlying effects due to the selective publication of studies.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Indexed literature search

The comprehensiveness of the review authors' search of indexed literature to identify eligible studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Grey literature

The comprehensiveness of the review authors' search of grey literature to identify eligible studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Language of included manuscripts

Whether authors applied restrictions to study selection on the basis of language

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Study sponsorship

Whether developers and purveyors of the intervention had influence on studies included in the review

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Number of small studies

Degree to which the body of evidence consists of studies with small sample sizes

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Funnel plot asymmetry

Whether there was evidence of funnel plot asymmetry

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Discrepancies between published and unpublished studies

Results from any approaches to assess discrepancies in findings between published and unpublished studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 10

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Upgrading the Initial Quality Rating

This page lists criteria related to factors that can increase the initial rating of the quality of a body of evidence if that body of evidence has not been downgraded for any other reason.

We want you to rate the importance of each criterion for considering whether to upgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Large magnitude of an effect

Rating up the quality of a body of evidence from non-randomised studies that yield large or very large estimates of the magnitude of an intervention effect

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Dose-response gradient

Rating up the quality of a body of evidence from non-randomised studies when there is presence of a dose-response gradient

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Effect of plausible residual confounding

Rating up the quality of a body of evidence from non-randomised studies when all plausible residual confounding from non-randomised studies are likely to reduce the demonstrated effect or increase the effect if no effect was observed

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Consistency across diverse contexts

Rating up the quality of a body 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)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Analogous evidence

Rating up the quality of a body of evidence when there is supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanism(s)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Coherence of evidence for the causal pathway

Rating up the quality of a body of evidence when there is coherence of results in individual links in the causal pathway between intervention and distal outcomes

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 11

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Open-Ended Questions

1. Missing criteria?

Please provide any additional criteria that you believe should be considered when rating the quality of a body of evidence in reviews of complex interventions.

2. Features of complexity that pose challenges to the GRADE approach

Please provide your thoughts on whether and how aspects of "complexity" pose challenges to applying the GRADE approach for rating the quality of a body of evidence used to estimate the effects of complex interventions.

3. Mixed body of evidence

Please provide your thoughts on whether and how a mixed body of evidence comprised of different study designs (e.g., randomised controlled trials and non-randomised experimental designs) poses challenges to applying the GRADE approach in reviews of complex interventions.

4. Requirements for using the GRADE approach

Please provide your thoughts on the experience, training, and/or expertise that are required to be qualified to use the GRADE approach in reviews of complex interventions.

5. Implementing guidance on rating the quality of evidence

Please provide your thoughts on how best to disseminate and implement guidance for rating the quality of a body of evidence on complex

Appendix 8. Round One online expert panel questionnaire Panel B

Page 1

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Background Information

Please find brief background information on "complex interventions" and "rating the quality of the body of evidence" below to assist you in participating in this online panel.

Complex Interventions

The "complexity" of an intervention is often understood through an assessment of one or more of the following characteristics:

- Number of interacting intervention components
- Number and difficulty of behaviors involved in the intervention delivery and receipt
- Number of groups or organizational levels targeted by the intervention
- Number and variability of outcomes
- Flexibility, tailoring or non-standardization of intervention implementation

Complexity is also increasingly understood through an assessment of the dynamic properties of the context (or "system") into which an intervention is introduced, such as non-linear relationships, feedback loops, phase changes, emergent properties, and interdependencies.

Rating the Quality of a Body of Evidence

A "body of evidence" refers to the totality of evidence contributing to an estimate of the comparative effect of an intervention for a specific outcome in a systematic review.

A rating of the "quality" of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

"Approaches to rating the quality of a body of evidence" typically assign an initial quality rating and then consider reasons to possibly downgrade or upgrade this quality rating.

Page 2

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Round One Instructions

For a tutorial on Round One, please read the instructions below and watch this video on an example "Round One" panel in ExpertLens.

Your Task for This Round

On the following pages, you will see lists of criteria that *could* be considered when rating the quality of a body of evidence in reviews of complex interventions. As a stakeholder in complex intervention research, we want you to rate the importance of considering each criterion for *all* reviews of complex interventions.

The criteria are organized as follows:

- Criteria related to an initial rating for the quality of the body of evidence (Page 2)
- Criteria related to downgrading the initial rating for the quality of the body of evidence (Pages 3-8)
- Criteria related to upgrading the initial rating for the quality of the body of evidence (Page 9)

Using the Rating Scale and Comment Box

We will ask you to rate each criterion on a scale from "1" ("Lower Importance") to "9" ("Higher Importance"). We will interpret your ratings as follows:

- Scores of 1-to-3 indicate that you believe a criterion is *of limited importance* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 4-to-6 indicate that you believe a criterion is *important but not critical* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 7-to-9 indicate that you believe a criterion is *critically important* to consider when rating the quality of a body of evidence in reviews of complex interventions

We encourage you to consider using the full range of the rating scale (from "1" to "9") across criteria and to use the comment boxes to briefly clarify your ratings.

We ask that you respond to all questions to Round One by 10:00AM (PT) on April 5.

Page 3

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

An Initial Quality Rating based on the Design of Included Studies

Many approaches for rating the quality of a body of evidence start with an "initial quality rating" (e.g., "low", "moderate", or "high") based on the design of the studies included in that body of evidence.

We want you to rate the importance of each criterion below when assigning an initial rating for the quality of a body of evidence in *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Randomised experimental studies

An initial quality rating of "high" when the body of evidence consists of randomised controlled trials (RCTs)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Non-randomised experimental studies

An initial quality rating of "moderate" when the body of evidence consists of non-randomised experimental study designs (e.g., natural experiments, quasi-experimental studies)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Non-experimental observational studies

An initial quality rating of "low" when the body of evidence consists of non-experimental observational studies (e.g., cohort design, case-control study)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Same rating across all study designs

An initial rating of "high" for a body of evidence consisting of any type of study design

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 4

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Randomized Trials

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on methodological limitations of randomized controlled trials (RCTs) included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. 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.)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Blinding intervention providers

Whether providers of the intervention are aware of the arm to which participants have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Blinding intervention recipients

Whether recipients of the intervention are aware of the arm to which they have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Blinding outcome assessors

Whether those assessing outcomes are aware of the arm to which participants have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Blinding data analysts

Whether those analysing data are aware of the arm to which participants have been allocated

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. 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

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Selective outcome reporting

Whether there is incomplete or absent reporting of some outcomes and not others on the basis of the results

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Stopping study early for benefit

Whether studies have been ended early when beneficial results were found in interim analyses

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Quality of outcome measures

Whether studies used outcome measures with low validity and/or reliability to assess intervention effects

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

10. Fidelity of intervention implementation

Whether there were deviations in intervention implementation from what was intended

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 5

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Non-Randomized Studies

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on methodological limitations of non-randomized studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Confounding

Whether the study used measures to adequately control for confounding

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Appropriate comparison group

Whether the study developed and applied an appropriate comparison group

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Selection of participants into the study

Whether the study used procedures to appropriately select participants into the study or into the analysis

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Classification of interventions

Whether intervention groups were clearly defined

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Deviations from intended interventions

Whether there were deviations from the intended intervention beyond what would be expected in usual practice

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Missing data

Whether the study used appropriate analytic methods for dealing with participant missing data

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Measurement of outcomes

Whether the study appropriately measured outcomes in all groups

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Selection of the reported result

Whether there is selected reporting of effect estimates based on multiple measurements and analyses

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Follow-up

Whether the study had adequate follow-up of participants

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 6

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Inconsistency of Effects in the Evidence Base

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from

"high" to "low"). This page lists criteria related to downgrading the initial quality rating based on differences in intervention effect estimates across studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Variability in point estimates

The degree to which point estimates vary across individual studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence intervals

The degree to which confidence intervals overlap across individual studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. 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

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. 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 (chance)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. 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)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Qualitative analyses exploring heterogeneity

Results of qualitative analyses of evidence exploring varying effects of interventions that help explain heterogeneity (e.g., qualitative comparative analysis)

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 7

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Inapplicability of the Evidence Base

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on the applicability of the body of evidence to the populations, interventions, outcomes, and settings of interest.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Inapplicability of study populations

Degree to which the participants in included studies compare to the population of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Inapplicability of study interventions

Degree to which the interventions in included studies compare to the intervention of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Inapplicability of outcomes

Degree to which the outcomes considered in included studies compare to the outcomes of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Inapplicability of follow-up timing

Degree to which the timing of outcome assessments in included studies compares to the follow-up timepoints of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Inapplicability of comparisons

Degree to which effect estimates are from comparison groups of interest

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 8

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Imprecision of the Effect Estimates

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on the width of confidence intervals for intervention effect estimates.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Adequate sample size

Whether the total number of participants in the review meets a conventional sample size for a single adequately powered trial

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence interval with line of no effect

Whether the confidence interval for the overall estimate includes effects indicating both benefit and harm

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Width of confidence interval

Whether the confidence interval includes estimates of important benefit and important harm

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 9

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Publication Bias

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on systematic under-estimation or over-estimation of underlying effects due to the selective publication of studies.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Indexed literature search

The comprehensiveness of the review authors' search of indexed literature to identify eligible studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Grey literature

The comprehensiveness of the review authors' search of grey literature to identify eligible studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Language of included manuscripts

Whether authors applied restrictions to study selection on the basis of language

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Study sponsorship

Whether developers and purveyors of the intervention had influence on studies included in the review

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Number of small studies

Degree to which the body of evidence consists of studies with small sample sizes

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Funnel plot asymmetry

Whether there was evidence of funnel plot asymmetry

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Discrepancies between published and unpublished studies

Results from any approaches to assess discrepancies in findings between published and unpublished studies

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 10

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Upgrading the Initial Quality Rating

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "upgrade" the initial quality rating (e.g., from "low" to "high") if that body of evidence has not been downgraded for any other reason.

We want you to rate the importance of each criterion for considering whether to upgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Large magnitude of an effect

Rating up the quality of a body of evidence from non-randomised studies that yield large or very large estimates of the magnitude of an intervention effect

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Dose-response gradient

Rating up the quality of a body of evidence from non-randomised studies when there is presence of a dose-response gradient

1	2	3	4	5	6	7	8	9

Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Effect of plausible residual confounding

Rating up the quality of a body of evidence from non-randomised studies when all plausible residual confounding from non-randomised studies are likely to reduce the demonstrated effect or increase the effect if no effect was observed

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Consistency across diverse contexts

Rating up the quality of a body 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)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Analogous evidence

Rating up the quality of a body of evidence when there is supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanism(s)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Coherence of evidence for the causal pathway

Rating up the quality of a body of evidence when there is coherence of results in individual links in the causal pathway between intervention and distal outcomes

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 11

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round One

Round One ends Apr 05, 2017 10:00 AM PT

In this study, you are Client X

Open-Ended Questions

1. Missing criteria?

Please provide any additional criteria that you believe should be considered when rating the quality of a body of evidence in reviews of complex interventions.

2. Features of complexity that pose challenges to rating quality of evidence

Please provide your thoughts on whether and how aspects of "complexity" pose challenges to rating the quality of a body of evidence used to estimate the effects of complex interventions.

3. Mixed body of evidence

Please provide your thoughts on whether and how a mixed body of evidence comprised of different study designs (e.g., randomised controlled trials and non-randomised experimental designs) poses challenges to rating the quality of a body of evidence in reviews of complex interventions.

4. Requirements for assessing quality of evidence

Please provide your thoughts on the experience, training, and/or expertise that are required to be qualified to rate the quality of a body of evidence in reviews of complex interventions.

5. Implementing guidance on rating the quality of evidence

Please provide your thoughts on how best to disseminate and implement guidance for rating the quality of a body of evidence on complex

Appendix 9. Round Three online expert panel questionnaire Panel A

Page 1

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Round Three Instructions

Welcome to Round Three of this online Delphi process on rating the quality of evidence in reviews of complex interventions.

For more information on how to participate in Round Three, please read the instructions below and watch this video on an example "Round Three" panel in ExpertLens.

Your Task for This Round

On the following pages, you will be able to review Round Two discussions and revise your Round One responses, if you wish to do so. While you can read Round Two discussion topics, please note that you will no longer be able to respond to them.

As with Round One, you will see lists of criteria that *could* be considered when rating the quality of a body of evidence in reviews of complex interventions.

As a stakeholder in complex intervention research, we want you to rate the importance of considering each criterion for *all* reviews of complex interventions.

Using the Rating Scale and Comment Box

We will ask you to rate each criterion on a scale from "1" ("Lower Importance") to "9" ("Higher Importance"). We will interpret your ratings as follows:

- Scores of 1-to-3 indicate that you believe a criterion is *of limited importance* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 4-to-6 indicate that you believe a criterion is *important but not critical* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 7-to-9 indicate that you believe a criterion is *critically important* to consider when rating the quality of a body of evidence in reviews of complex interventions

We encourage you to consider using the full range of the rating scale (from "1" to "9") across criteria and to use the comment boxes to briefly clarify your ratings.

We ask that you respond to all questions to **Round Three by 8:00AM (PT) on May 1.**

Page 2

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Background Information

Please find brief background information on "complex interventions" and "rating the quality of the body of evidence" below to assist you in participating in this online panel.

Complex Interventions

The “complexity” of an intervention is often understood through an assessment of one or more of the following characteristics:

- Number of interacting intervention components
- Number and difficulty of behaviors involved in the intervention delivery and receipt
- Number of groups or organizational levels targeted by the intervention
- Number and variability of outcomes
- Flexibility, tailoring or non-standardization of intervention implementation

Complexity is also increasingly understood through an assessment of the dynamic properties of the context (or "system") into which an intervention is introduced, such as non-linear relationships, feedback loops, phase changes, emergent properties, and interdependencies.

Rating the Quality of a Body of Evidence

A "body of evidence" refers to the totality of evidence contributing to an estimate of the comparative effect of an intervention for a specific outcome in a systematic review.

A rating of the "quality" of a body of evidence indicates the reviewers’ confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

"Approaches to rating the quality of a body of evidence" typically assign an initial quality rating and then consider reasons to possibly downgrade or upgrade this quality rating.

Page 3

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Initial Quality Rating: Study Design

The criteria below relate to the initial "quality" rating for a body of evidence based on the design of the studies included in that body of evidence.

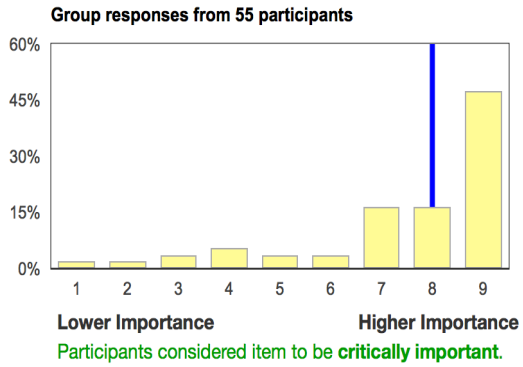
We want you to rate the importance of each criterion for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Randomised experimental studies

An initial quality rating of "high" when the body of evidence consists of randomised controlled trials (RCTs)

[11]



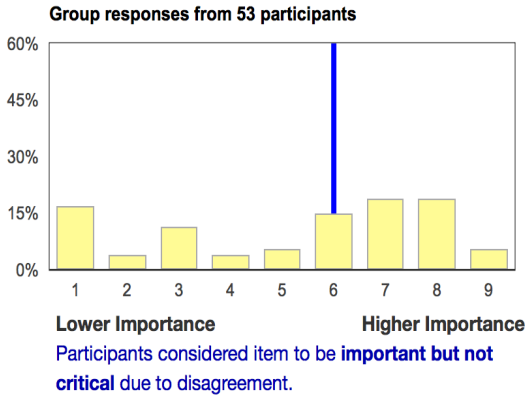
	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

[11] The diagram shows distribution of Round One answers. The height of yellow bars is determined by the number of participants choosing a particular response category. The blue line is the group median. A red dot was also present on the original questionnaire, representing a participant's response.

Please, provide the rationale behind your answer here

2. Non-randomised experimental studies

An initial quality rating of "moderate" when the body of evidence consists of non-randomised experimental study designs (e.g., natural experiments, quasi-experimental studies)

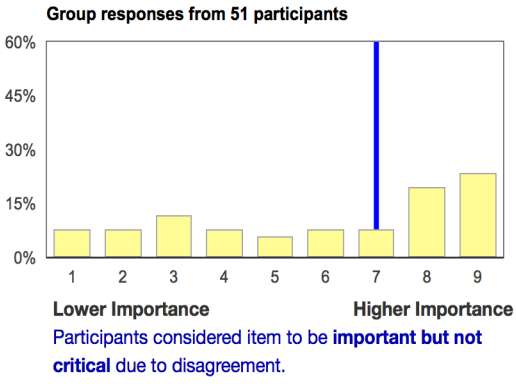


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Non-experimental observational studies

An initial quality rating of "low" when the body of evidence consists of non-experimental observational studies (e.g., cohort design, case-control study)

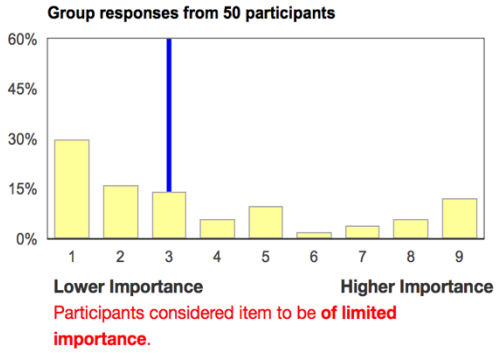


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Same rating across all study designs

An initial rating of "high" for a body of evidence consisting of any type of study design



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 4

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Randomised Trials

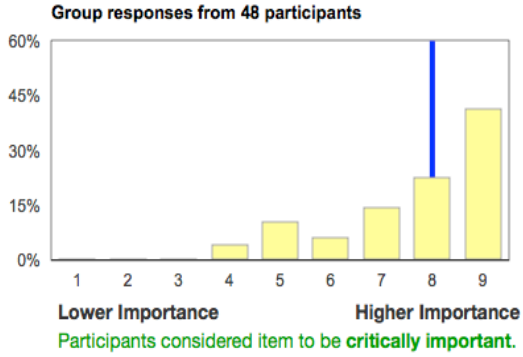
This page lists criteria related to methodological limitations of randomised controlled trials (RCTs) included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. 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.)

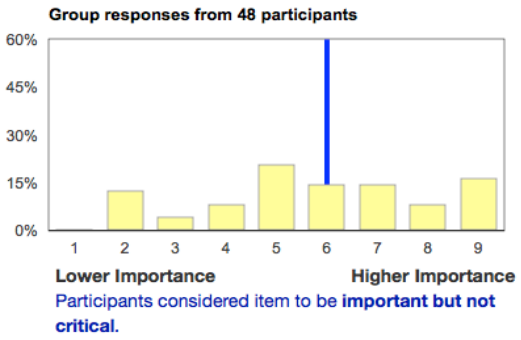


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Blinding intervention providers

Whether providers of the intervention are aware of the arm to which participants have been allocated

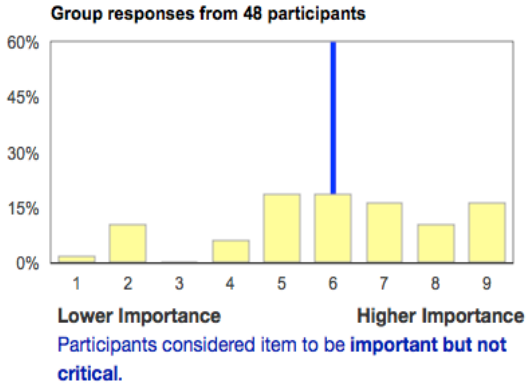


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Blinding intervention recipients

Whether recipients of the intervention are aware of the arm to which they have been allocated

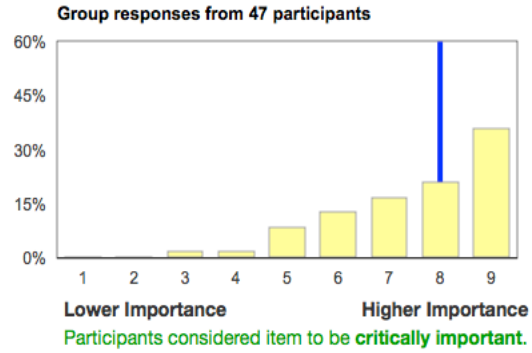


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Blinding outcome assessors

Whether those assessing outcomes are aware of the arm to which participants have been allocated

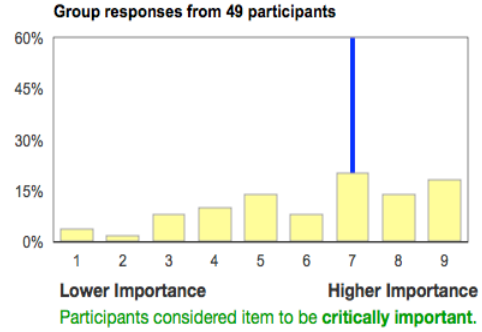


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Blinding data analysts

Whether those analysing data are aware of the arm to which participants have been allocated

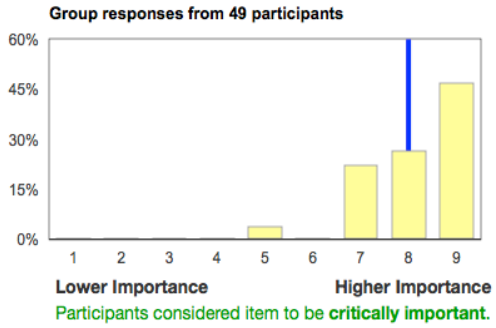


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. 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

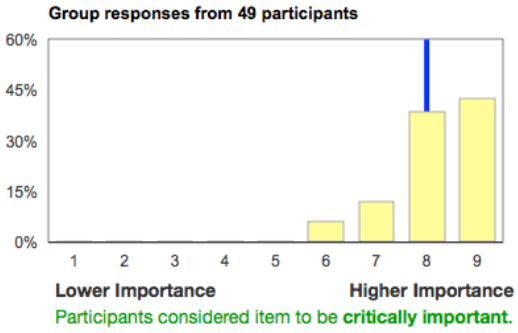


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Selective outcome reporting

Whether there is incomplete or absent reporting of some outcomes and not others on the basis of the results

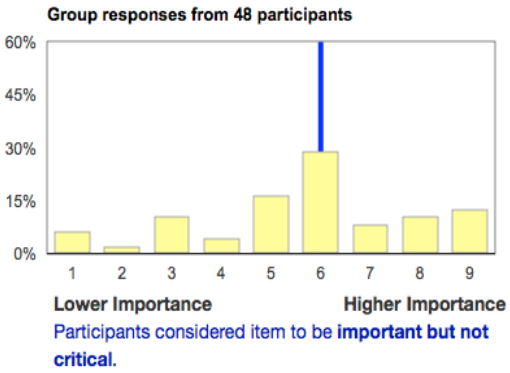


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Stopping study early for benefit

Whether studies have been ended early when beneficial results were found in interim analyses

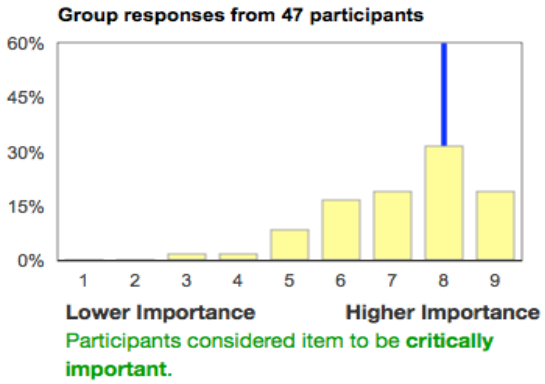


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Quality of outcome measures

Whether studies used outcome measures with low validity and/or reliability to assess intervention effects

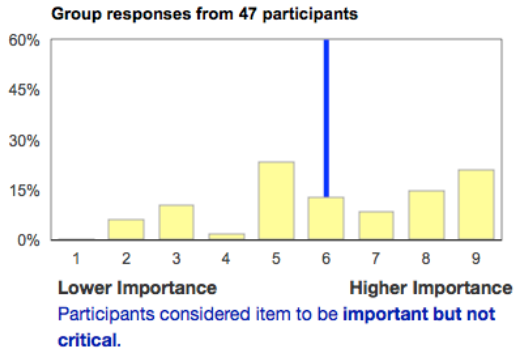


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

10. Fidelity of intervention implementation

Whether there were deviations in intervention implementation from what was intended



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 5

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Non-Randomised Studies

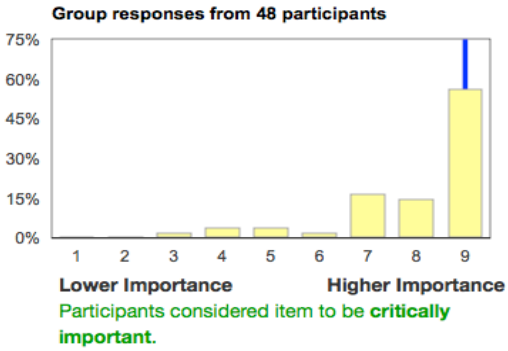
This page lists criteria related to methodological limitations of non-randomised studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Confounding

Whether the study used measures to adequately control for confounding

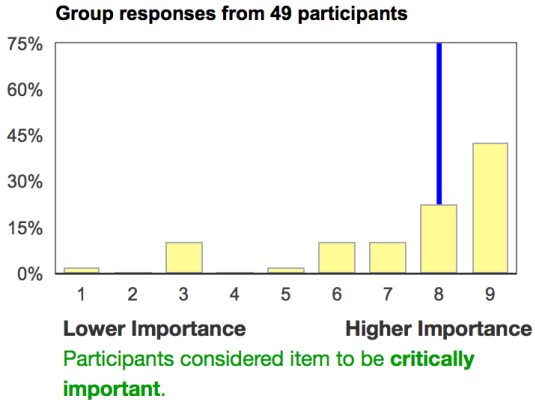


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Appropriate comparison group

Whether the study developed and applied an appropriate comparison group

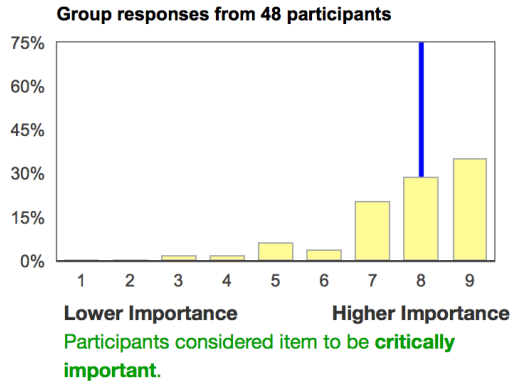


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Selection of participants into the study

Whether the study used procedures to appropriately select participants into the study or into the analysis

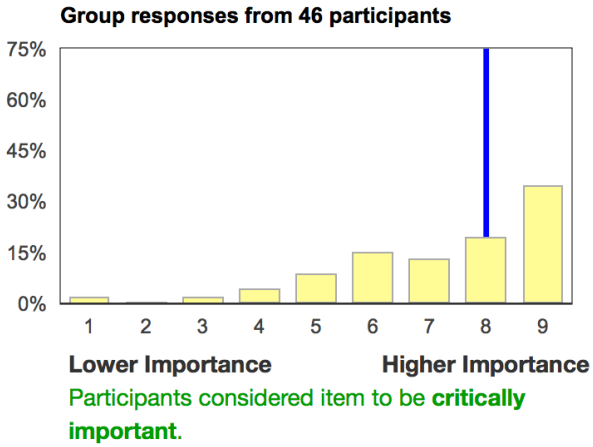


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Classification of interventions

Whether intervention groups were clearly defined

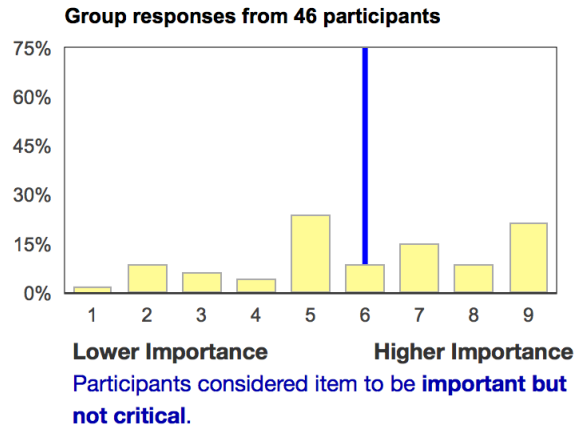


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Deviations from intended interventions

Whether there were deviations from the intended intervention beyond what would be expected in usual practice

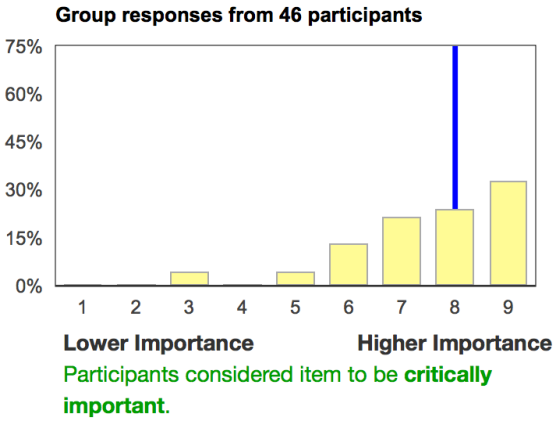


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Missing data

Whether the study used appropriate analytic methods for dealing with participant missing data

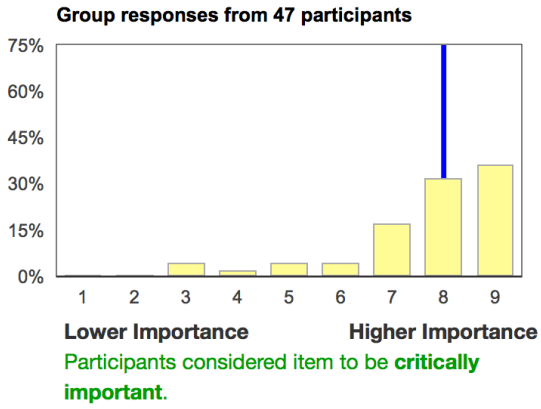


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Measurement of outcomes

Whether the study appropriately measured outcomes in all groups

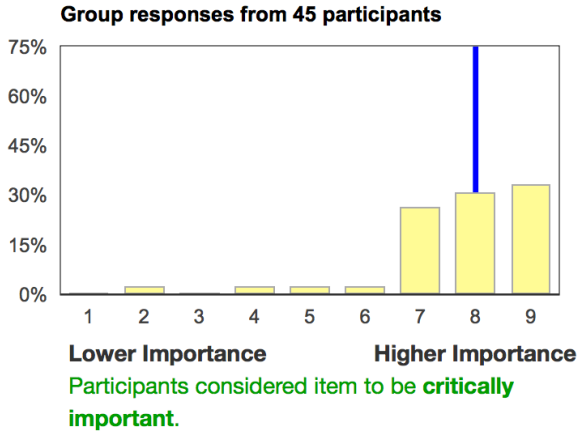


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Selection of the reported result

Whether there is selected reporting of effect estimates based on multiple measurements and analyses

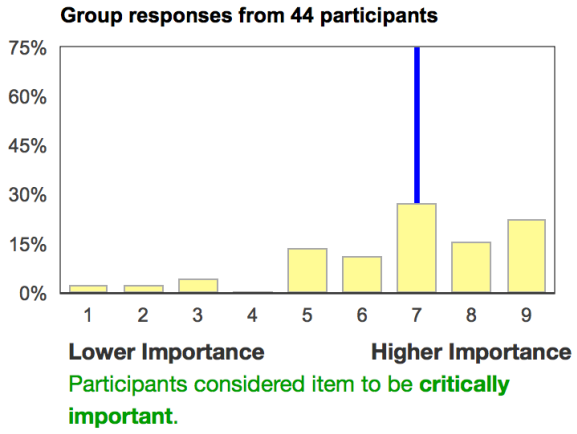


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Follow-up

Whether the study had adequate follow-up of participants



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 6

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Inconsistency of Effects in the Evidence Base

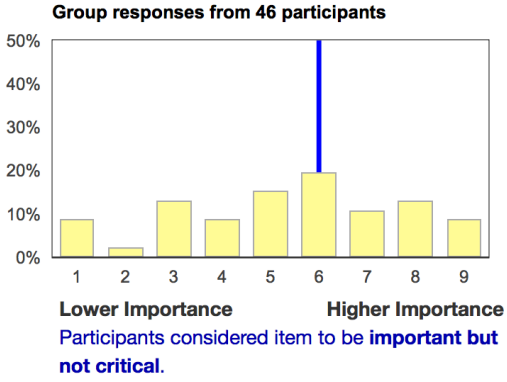
This page lists criteria related to differences in intervention effect estimates across studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Variability in point estimates

The degree to which point estimates vary across individual studies

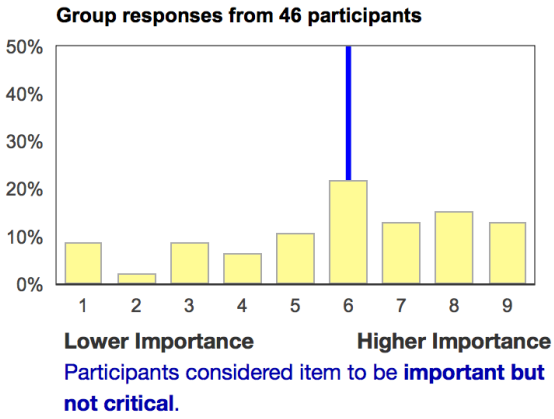


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence intervals

The degree to which confidence intervals overlap across individual studies

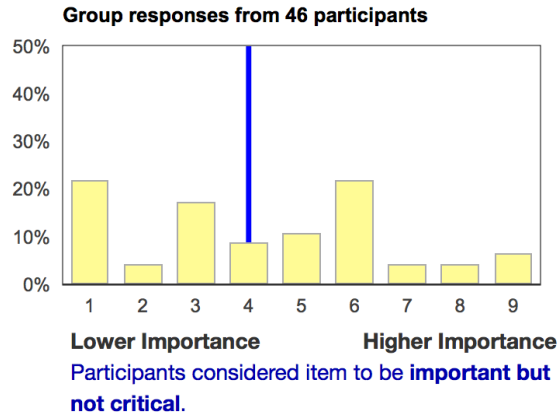


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. 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

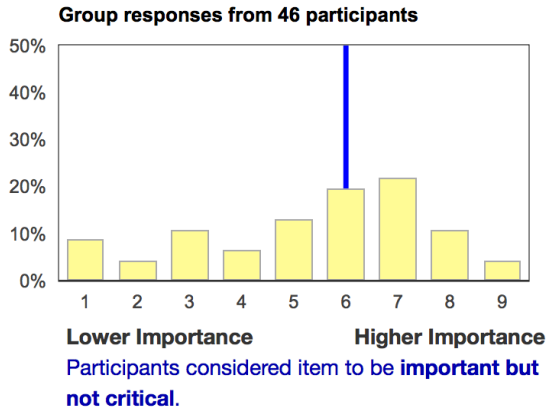


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. 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 (chance)

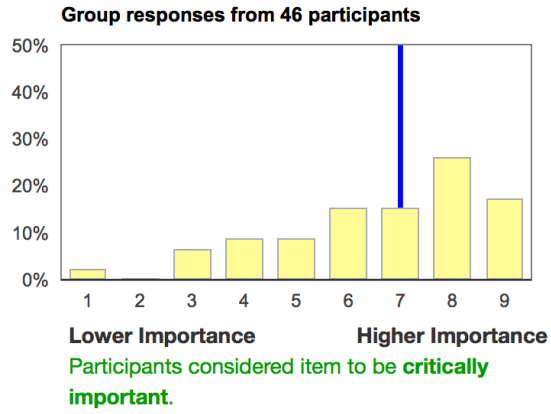


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. 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)

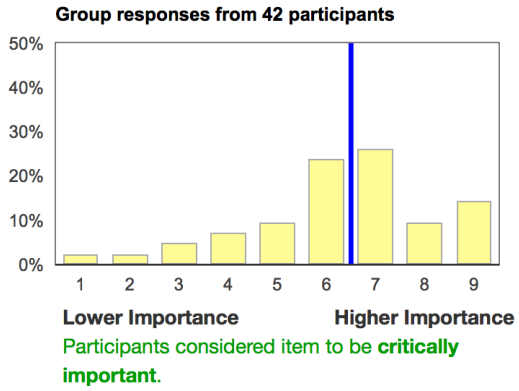


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Qualitative analyses exploring heterogeneity

Results of qualitative analyses of evidence exploring varying effects of interventions that help explain heterogeneity (e.g., qualitative comparative analysis)



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 7

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Indirectness of the Evidence Base

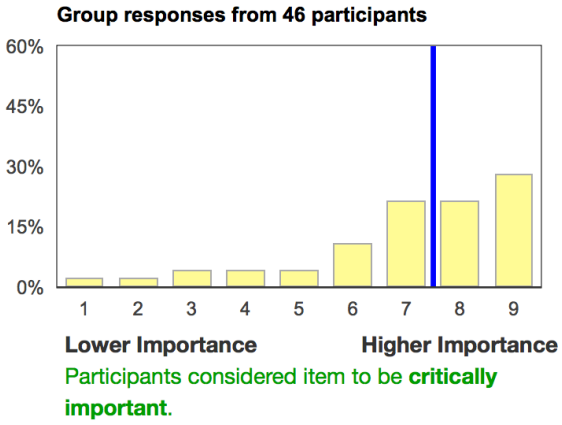
This page lists criteria related to the applicability of the body of evidence to the populations, interventions, outcomes, and settings of interest.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Indirectness of study populations

Degree to which the participants in included studies compare to the population of interest

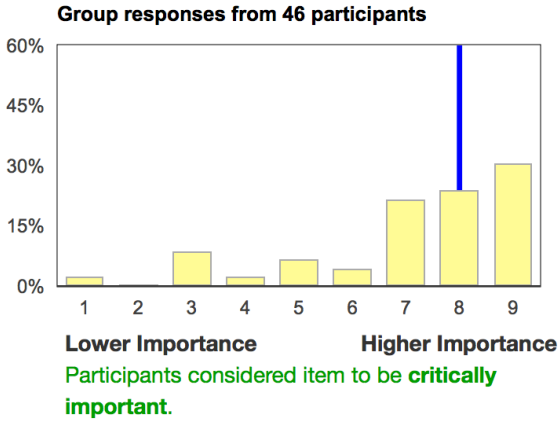


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Indirectness of study interventions

Degree to which the interventions in included studies compare to the intervention of interest

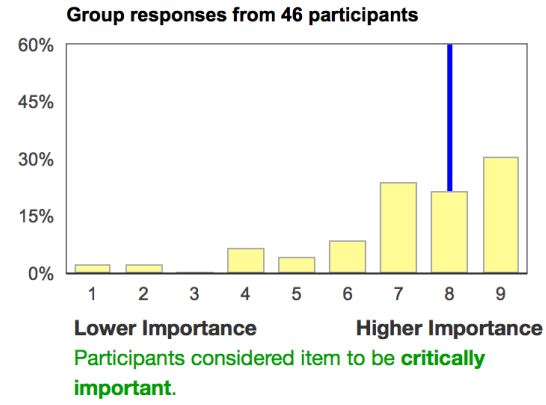


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Indirectness of outcomes

Degree to which the outcomes considered in included studies compare to the outcomes of interest

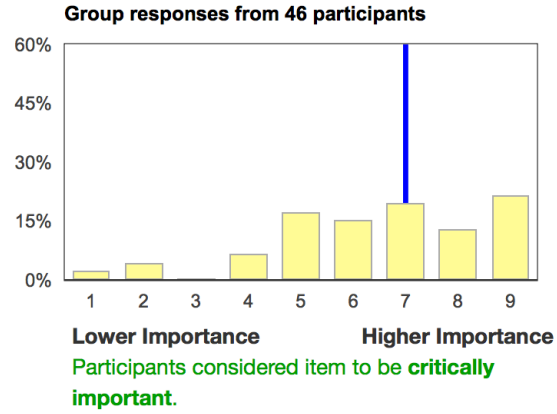


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. 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

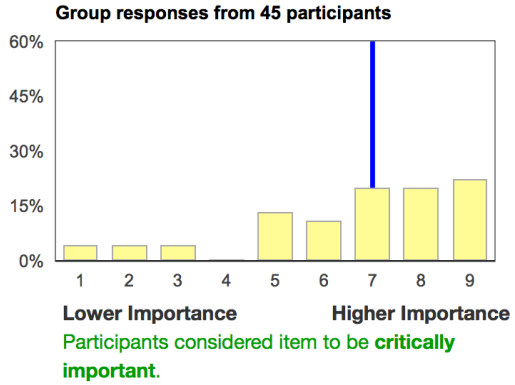


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Indirectness of comparisons

Degree to which effect estimates are from comparison groups of interest



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 8

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Imprecision of the Effect Estimates

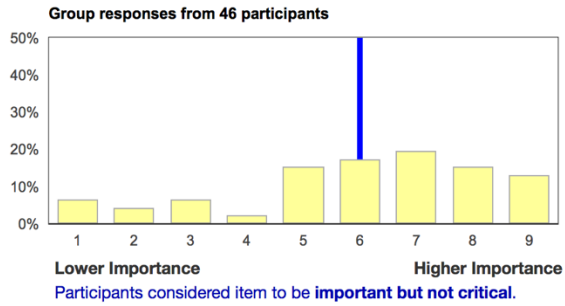
This page lists criteria related to the width of confidence intervals for intervention effect estimates.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Optimal information size

Whether the total number of participants in the review meets a conventional sample size for a single adequately powered trial

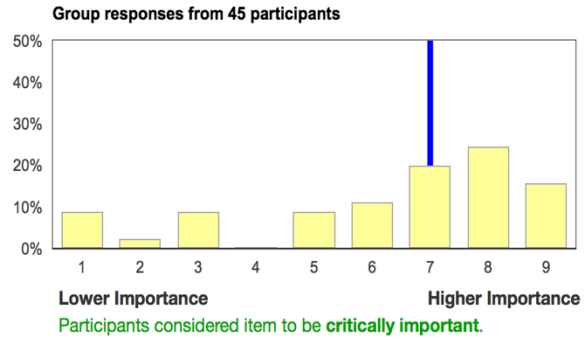


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence interval with line of no effect

Whether the confidence interval for the overall estimate includes effects indicating both benefit and harm

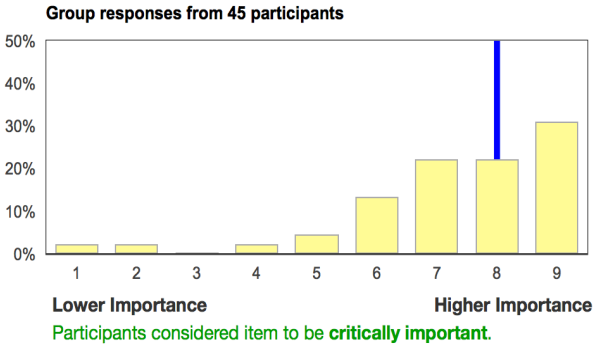


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Width of confidence interval

Whether the confidence interval includes estimates of important benefit and important harm



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 9

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Publication Bias

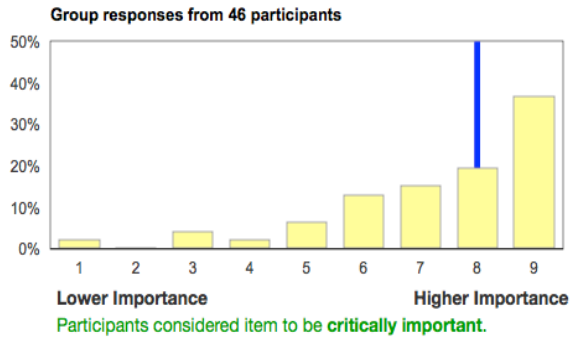
This page lists criteria related to the systematic under-estimation or over-estimation of underlying effects due to the selective publication of studies.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Indexed literature search

The comprehensiveness of the review authors' search of indexed literature to identify eligible studies



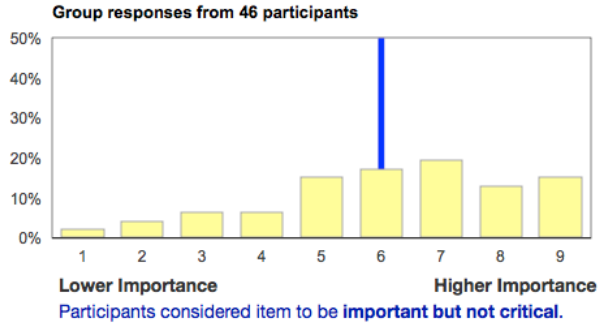
1 2 3 4 5 6 7 8 9

Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Grey literature

The comprehensiveness of the review authors' search of grey literature to identify eligible studies

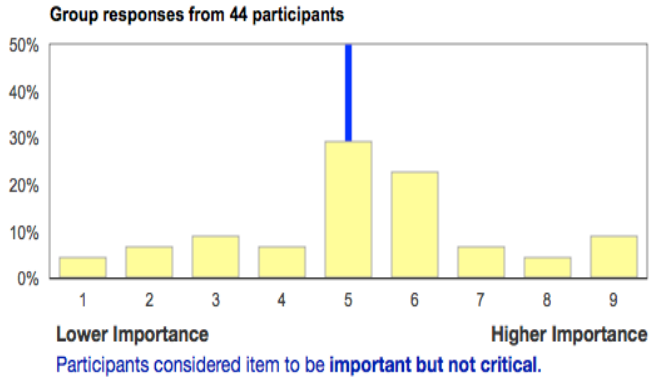


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Language of included manuscripts

Whether authors applied restrictions to study selection on the basis of language

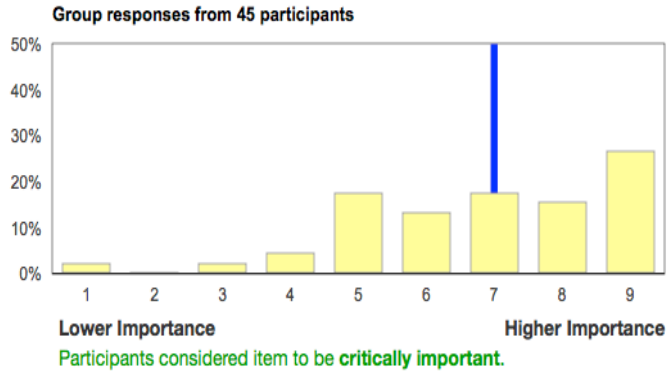


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Study sponsorship

Whether developers and purveyors of the intervention had influence on studies included in the review

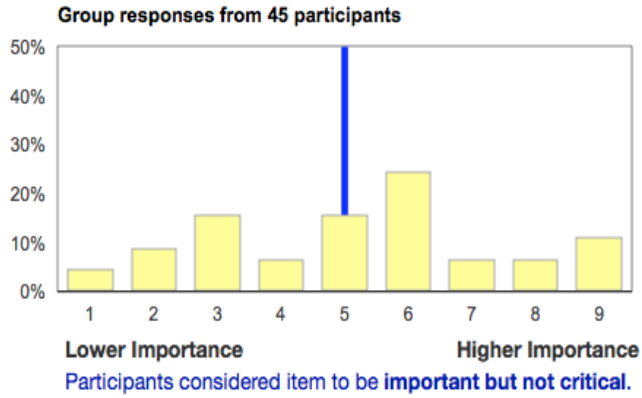


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Number of small studies

Degree to which the body of evidence consists of studies with small sample sizes

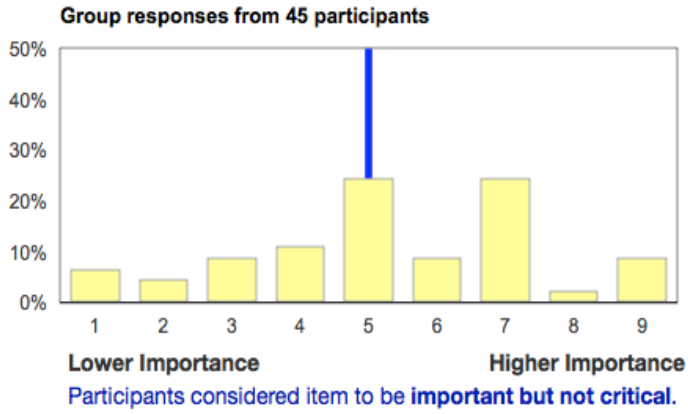


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Funnel plot asymmetry

Whether there was evidence of funnel plot asymmetry

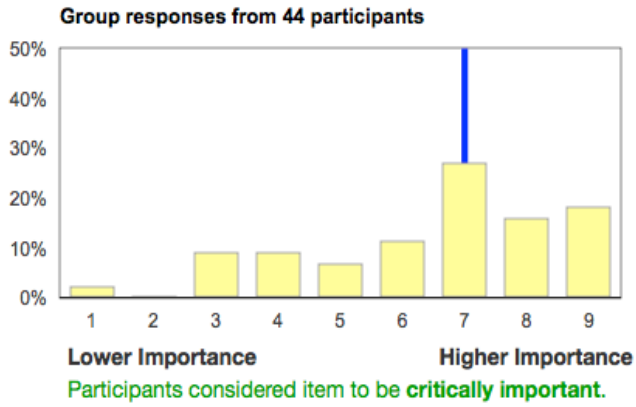


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Discrepancies between published and unpublished studies

Results from any approaches to assess discrepancies in findings between published and unpublished studies



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 10

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel A): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Upgrading the Initial Quality Rating

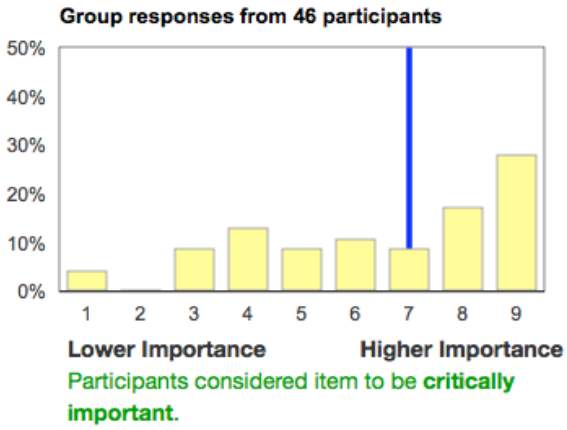
This page lists criteria related to factors that can increase the initial rating of the quality of a body of evidence if that body of evidence has not been downgraded for any other reason.

We want you to rate the importance of each criterion for considering whether to upgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Large magnitude of an effect

Rating up the quality of a body of evidence from non-randomised studies that yield large or very large estimates of the magnitude of an intervention effect

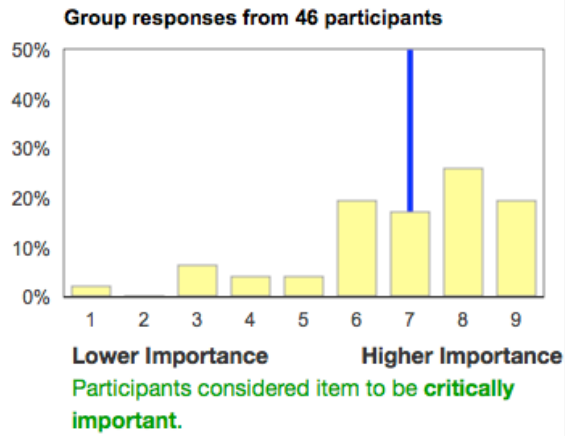


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Dose-response gradient

Rating up the quality of a body of evidence from non-randomised studies when there is presence of a dose-response gradient

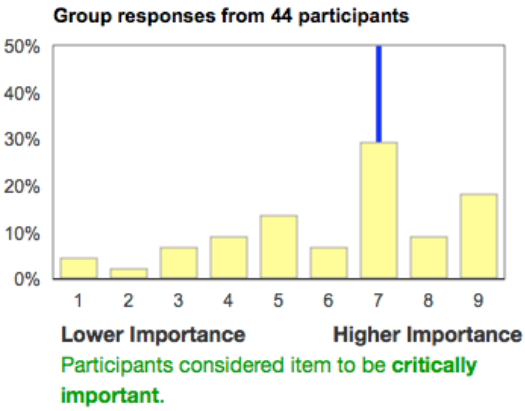


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Effect of plausible residual confounding

Rating up the quality of a body of evidence from non-randomised studies when all plausible residual confounding from non-randomised studies are likely to reduce the demonstrated effect or increase the effect if no effect was observed

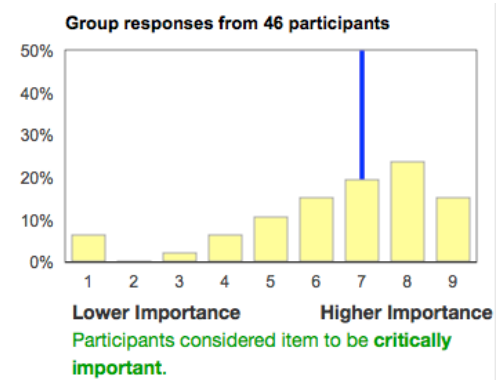


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Consistency across diverse contexts

Rating up the quality of a body 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)

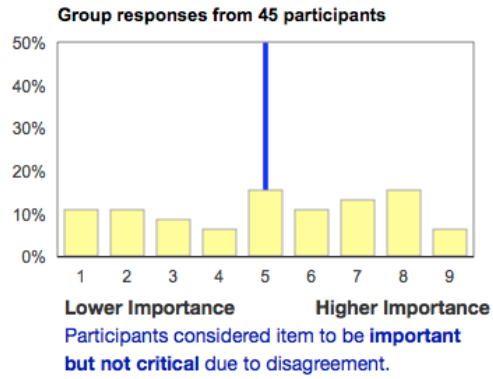


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Analogous evidence

Rating up the quality of a body of evidence when there is supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanism(s)

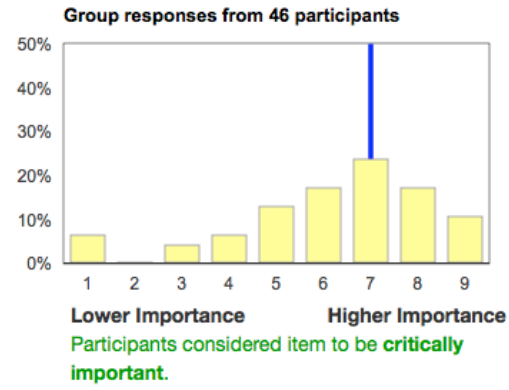


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Coherence of evidence for the causal pathway

Rating up the quality of a body of evidence when there is coherence of results in individual links in the causal pathway between intervention and distal outcomes



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Appendix 10. Round Three online expert panel questionnaire Panel B

Page 1

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Round Three Instructions

Welcome to Round Three of this online Delphi process on rating the quality of evidence in reviews of complex interventions.

For more information on how to participate in Round Three, please read the instructions below and watch this video on an example "Round Three" panel in ExpertLens.

Your Task for This Round

On the following pages, you will be able to review Round Two discussions and revise your Round One responses, if you wish to do so. While you can read Round Two discussion topics, please note that you will no longer be able to respond to them.

As with Round One, you will see lists of criteria that *could* be considered when rating the quality of a body of evidence in reviews of complex interventions.

As a stakeholder in complex intervention research, we want you to rate the importance of considering each criterion for *all* reviews of complex interventions.

Using the Rating Scale and Comment Box

We will ask you to rate each criterion on a scale from "1" ("Lower Importance") to "9" ("Higher Importance"). We will interpret your ratings as follows:

- Scores of 1-to-3 indicate that you believe a criterion is *of limited importance* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 4-to-6 indicate that you believe a criterion is *important but not critical* to consider when rating the quality of a body of evidence in reviews of complex interventions
- Scores of 7-to-9 indicate that you believe a criterion is *critically important* to consider when rating the quality of a body of evidence in reviews of complex interventions

We encourage you to consider using the full range of the rating scale (from "1" to "9") across criteria and to use the comment boxes to briefly clarify your ratings.

We ask that you respond to all questions to **Round Three by 8:00AM (PT) on May 1.**

Page 2

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Background Information

Please find brief background information on "complex interventions" and "rating the quality of the body of evidence" below to assist you in participating in this online panel.

Complex Interventions

The "complexity" of an intervention is often understood through an assessment of one or more of the following characteristics:

- Number of interacting intervention components
- Number and difficulty of behaviors involved in the intervention delivery and receipt
- Number of groups or organizational levels targeted by the intervention
- Number and variability of outcomes
- Flexibility, tailoring or non-standardization of intervention implementation

Complexity is also increasingly understood through an assessment of the dynamic properties of the context (or "system") into which an intervention is introduced, such as non-linear relationships, feedback loops, phase changes, emergent properties, and interdependencies.

Rating the Quality of a Body of Evidence

A "body of evidence" refers to the totality of evidence contributing to an estimate of the comparative effect of an intervention for a specific outcome in a systematic review.

A rating of the "quality" of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

"Approaches to rating the quality of a body of evidence" typically assign an initial quality rating and then consider reasons to possibly downgrade or upgrade this quality rating.

Page 3

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

An Initial Quality Rating based on the Design of Included Studies

Many approaches for rating the quality of a body of evidence start with an "initial quality rating" (e.g., "low", "moderate", or "high") based on the design of the studies included in that body of evidence.

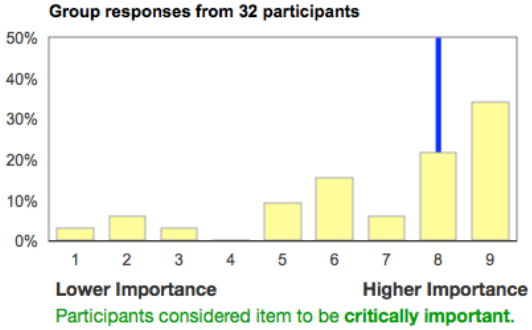
We want you to rate the importance of each criterion below when assigning an initial rating for the quality of a body of evidence in *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Randomised experimental studies

An initial quality rating of "high" when the body of evidence consists of randomised controlled trials (RCTs)

[12]



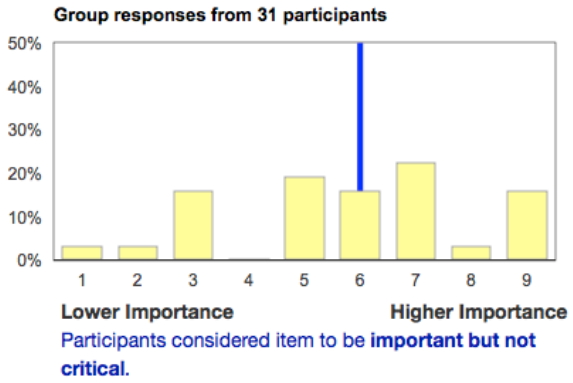
[12] The diagram shows distribution of Round One answers. The height of yellow bars is determined by the number of participants choosing a particular response category. The blue line is the group median. A red dot was also present on the original questionnaire, representing a participant's response.

	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Non-randomised experimental studies

An initial quality rating of "moderate" when the body of evidence consists of non-randomised experimental study designs (e.g., natural experiments, quasi-experimental studies)

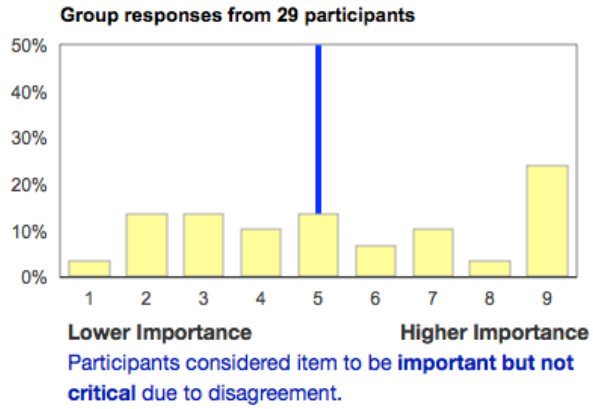


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Non-experimental observational studies

An initial quality rating of "low" when the body of evidence consists of non-experimental observational studies (e.g., cohort design, case-control study)

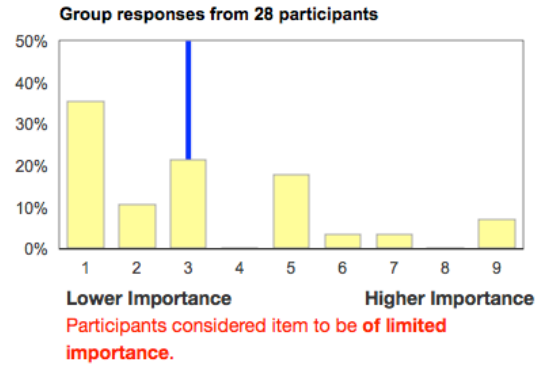


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Same rating across all study designs

An initial rating of "high" for a body of evidence consisting of any type of study design



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 4

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Randomised Trials

After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating

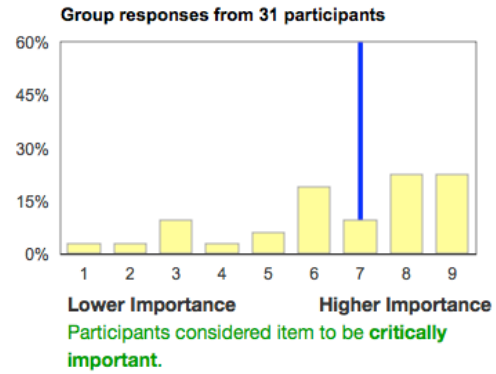
based on methodological limitations of randomized controlled trials (RCTs) included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. 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.)

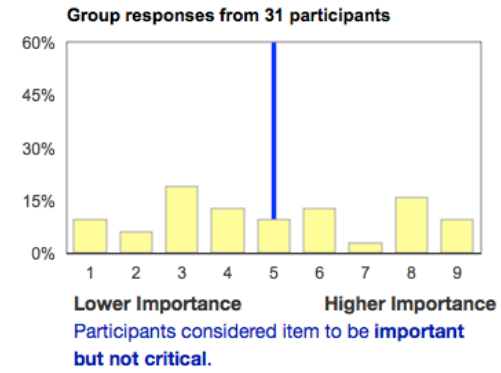


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Blinding intervention providers

Whether providers of the intervention are aware of the arm to which participants have been allocated

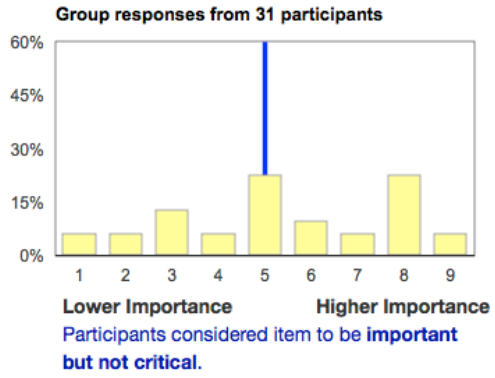


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Blinding intervention recipients

Whether recipients of the intervention are aware of the arm to which they have been allocated

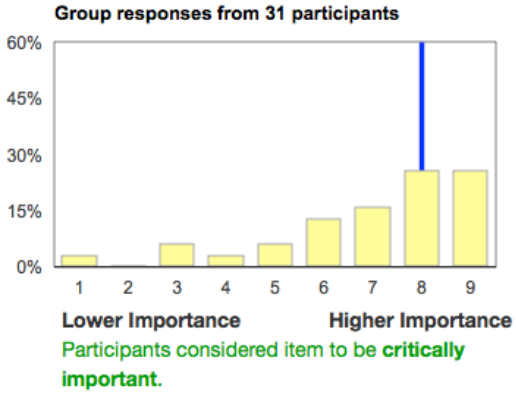


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Blinding outcome assessors

Whether those assessing outcomes are aware of the arm to which participants have been allocated

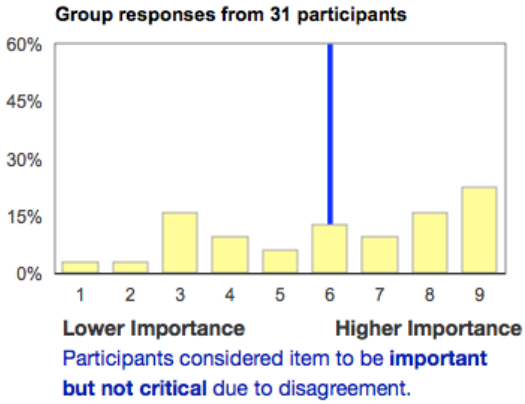


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Blinding data analysts

Whether those analysing data are aware of the arm to which participants have been allocated

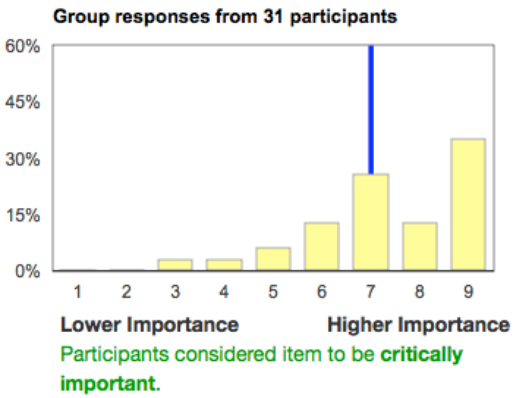


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. 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

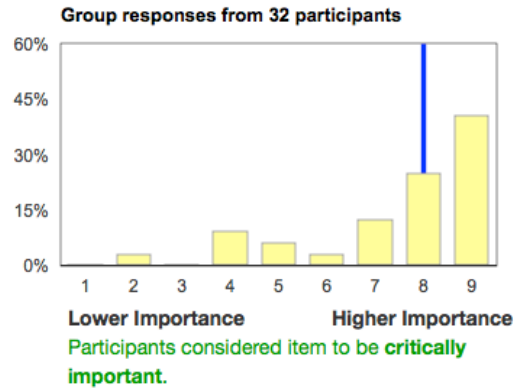


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Selective outcome reporting

Whether there is incomplete or absent reporting of some outcomes and not others on the basis of the results

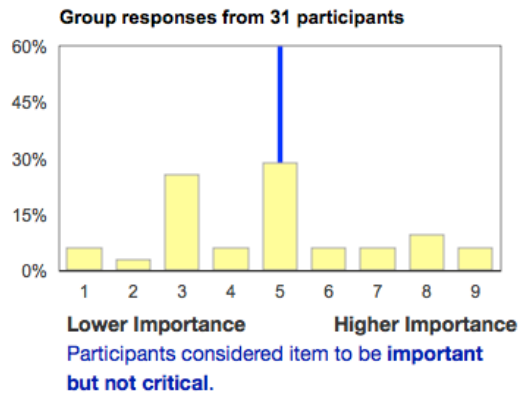


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Stopping study early for benefit

Whether studies have been ended early when beneficial results were found in interim analyses

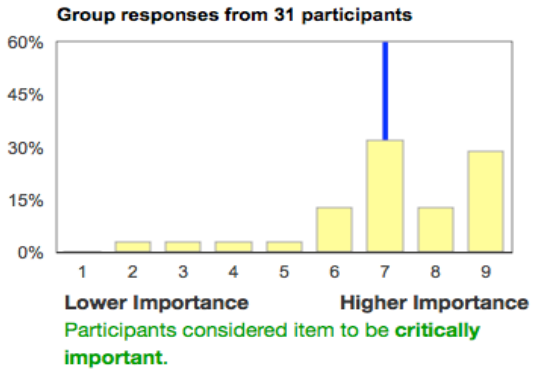


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Quality of outcome measures

Whether studies used outcome measures with low validity and/or reliability to assess intervention effects

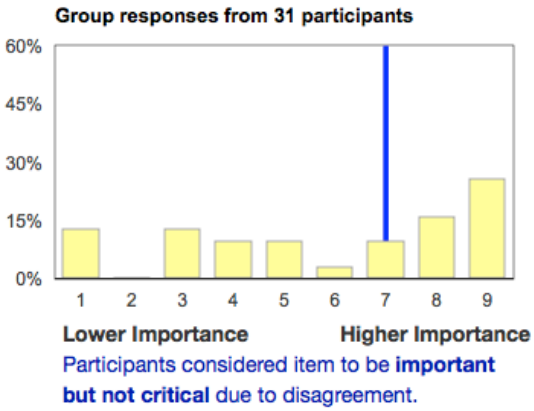


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

10. Fidelity of intervention implementation

Whether there were deviations in intervention implementation from what was intended



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 5

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Limitations of Included Non-Randomised Studies

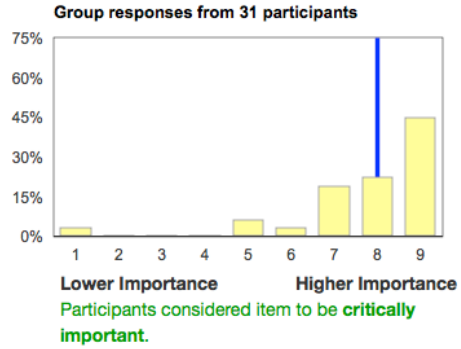
After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on methodological limitations of non-randomized studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Confounding

Whether the study used measures to adequately control for confounding

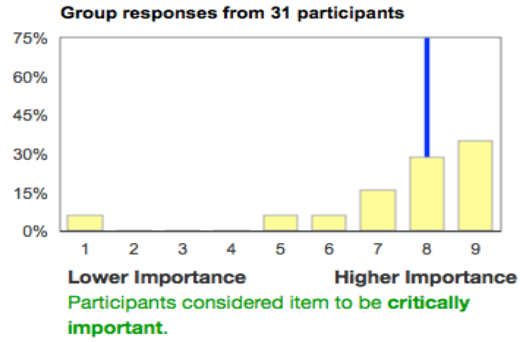


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Appropriate comparison group

Whether the study developed and applied an appropriate comparison group

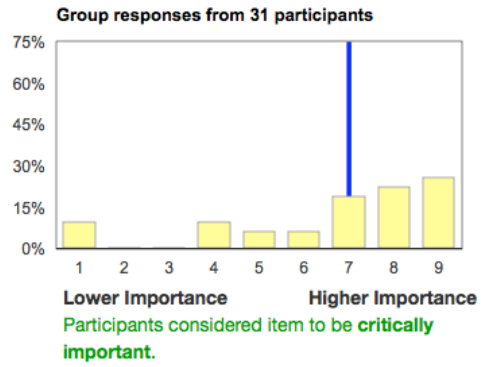


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Selection of participants into the study

Whether the study used procedures to appropriately select participants into the study or into the analysis

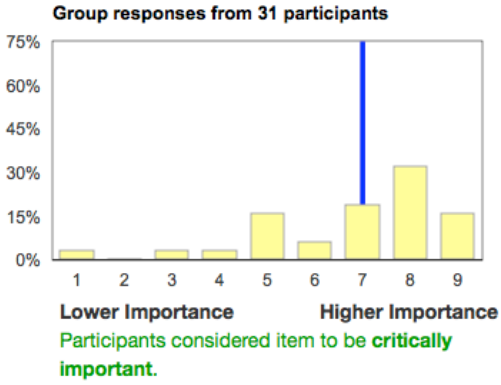


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Classification of interventions

Whether intervention groups were clearly defined

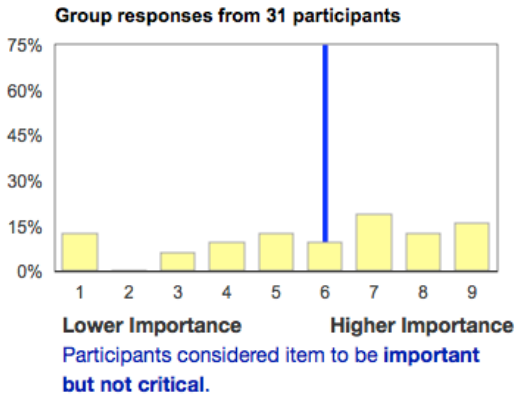


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Deviations from intended interventions

Whether there were deviations from the intended intervention beyond what would be expected in usual practice

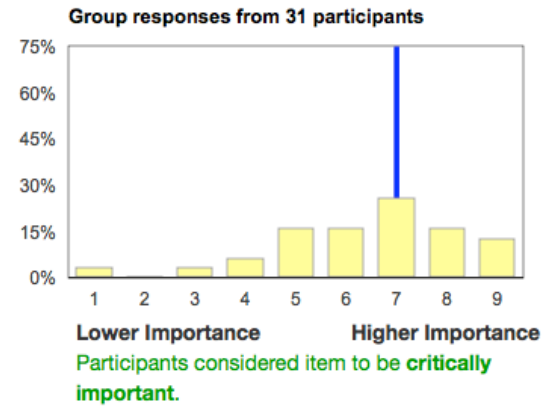


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Missing data

Whether the study used appropriate analytic methods for dealing with participant missing data

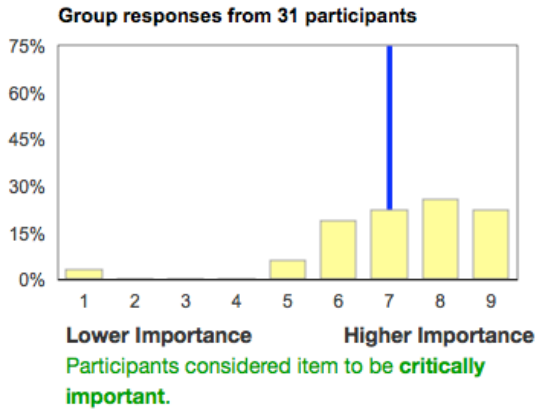


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Measurement of outcomes

Whether the study appropriately measured outcomes in all groups

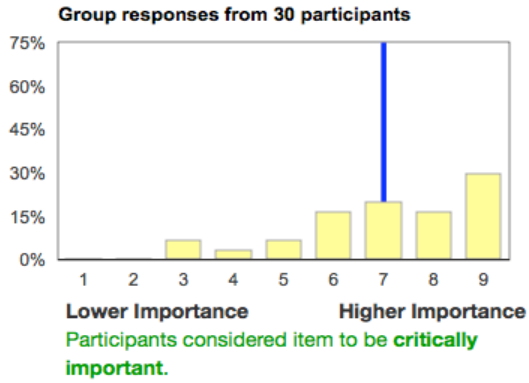


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

8. Selection of the reported result

Whether there is selected reporting of effect estimates based on multiple measurements and analyses

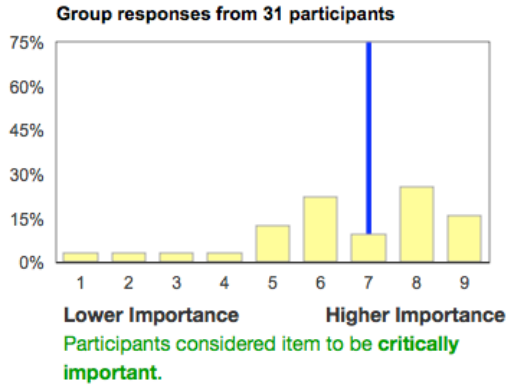


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

9. Follow-up

Whether the study had adequate follow-up of participants



Lower Importance	1	2	3	4	5	6	7	8	9	Higher Importance
	○	○	○	○	○	○	○	○	○	

Please, provide the rationale behind your answer here

Page 6

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Inconsistency of Effects in the Evidence Base

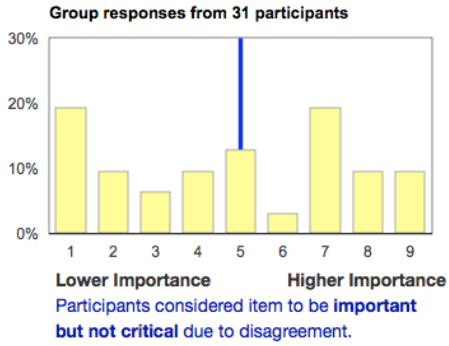
After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on differences in intervention effect estimates across studies included in the review.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Variability in point estimates

The degree to which point estimates vary across individual studies

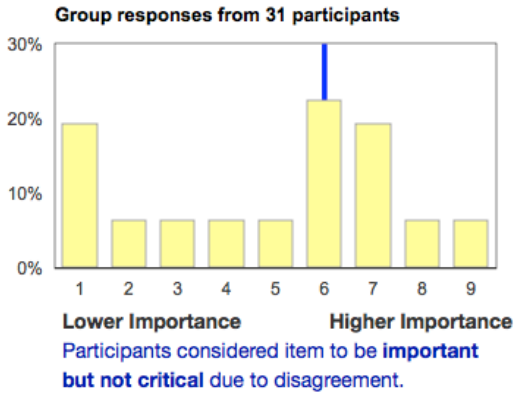


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence intervals

The degree to which confidence intervals overlap across individual studies

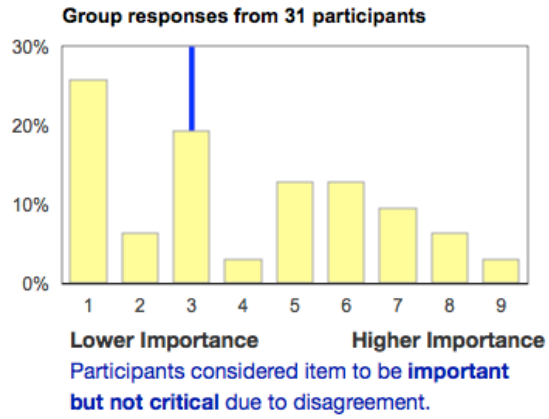


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. 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

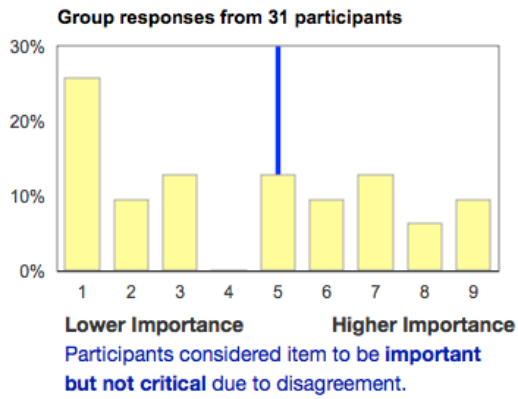


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. 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 (chance)

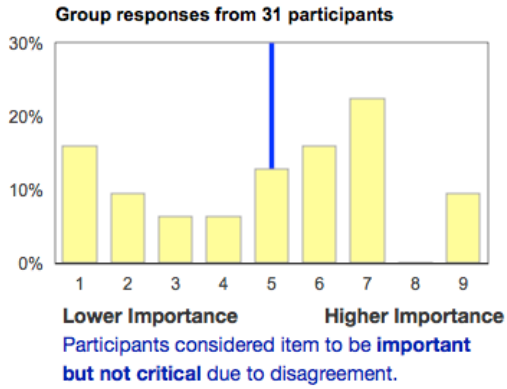


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. 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)

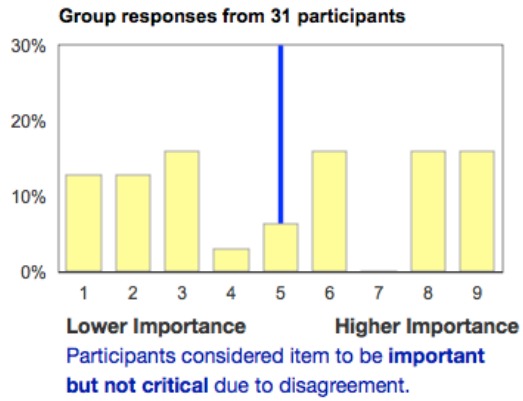


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Qualitative analyses exploring heterogeneity

Results of qualitative analyses of evidence exploring varying effects of interventions that help explain heterogeneity (e.g., qualitative comparative analysis)



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 7

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Inapplicability of the Evidence Base

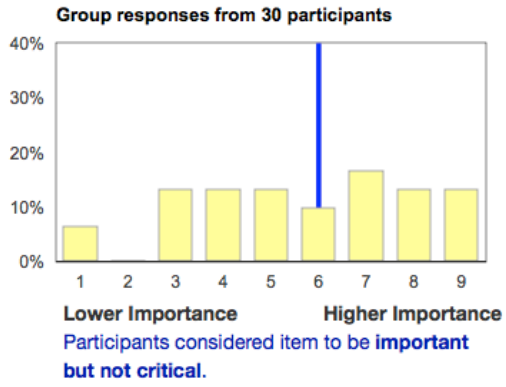
After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on the applicability of the body of evidence to the populations, interventions, outcomes, and settings of interest.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Inapplicability of study populations

Degree to which the participants in included studies compare to the population of interest

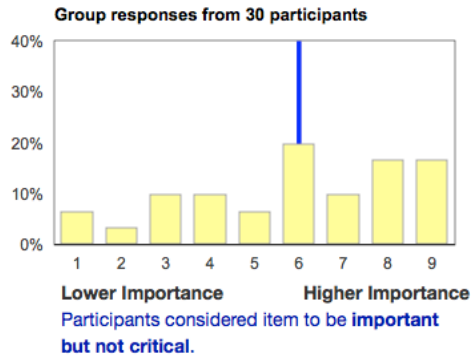


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Inapplicability of study interventions

Degree to which the interventions in included studies compare to the intervention of interest

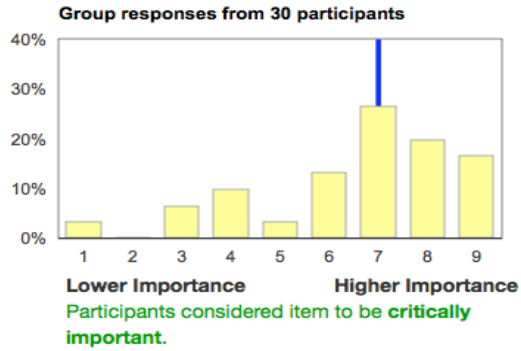


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Inapplicability of outcomes

Degree to which the outcomes considered in included studies compare to the outcomes of interest

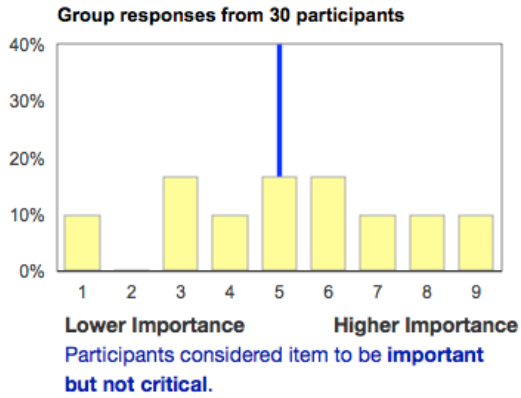


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Inapplicability of follow-up timing

Degree to which the timing of outcome assessments in included studies compares to the follow-up time-points of interest

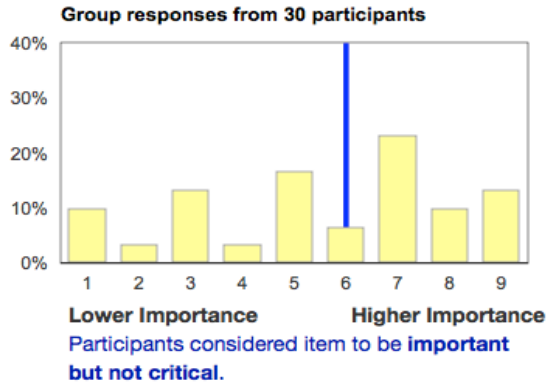


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Inapplicability of comparisons

Degree to which effect estimates are from comparison groups of interest



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 8

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Imprecision of the Effect Estimates

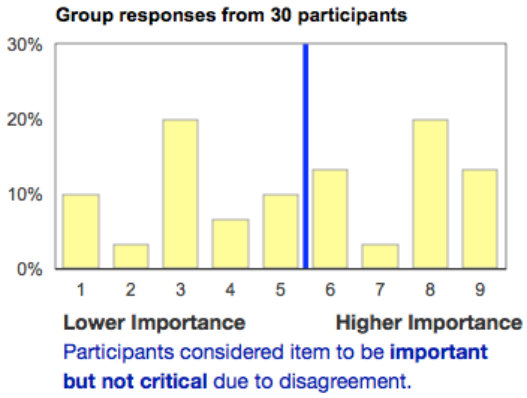
After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on the width of confidence intervals for intervention effect estimates.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Adequate sample size

Whether the total number of participants in the review meets a conventional sample size for a single adequately powered trial

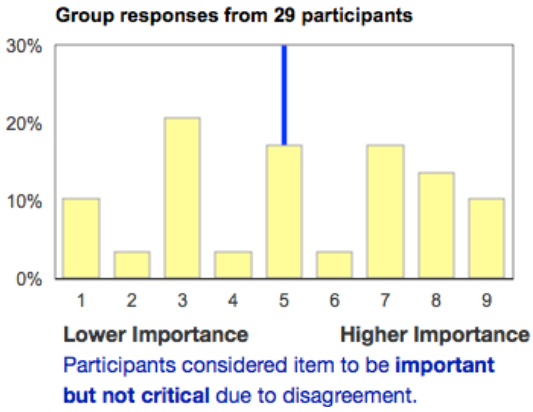


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Overlap of confidence interval with line of no effect

Whether the confidence interval for the overall estimate includes effects indicating both benefit and harm

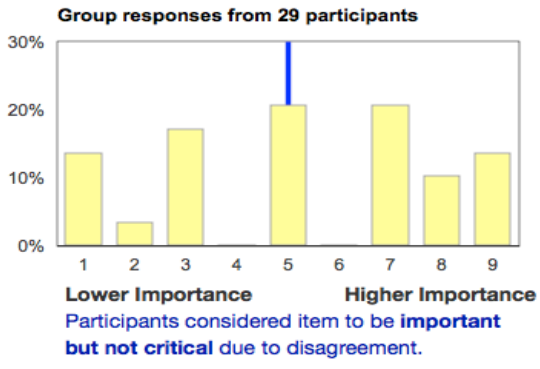


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Width of confidence interval

Whether the confidence interval includes estimates of important benefit and important harm



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 9

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Downgrading the Initial Quality Rating: Publication Bias

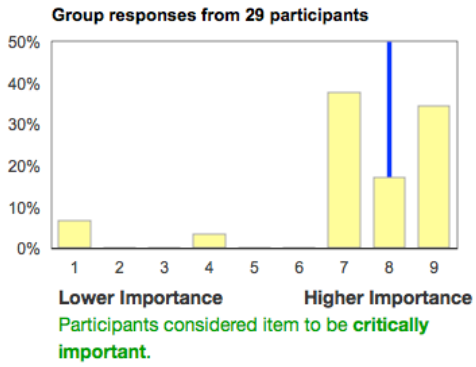
After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "downgrade" the initial quality rating (e.g., from "high" to "low"). This page lists criteria related to downgrading the initial quality rating based on systematic under-estimation or over-estimation of underlying effects due to the selective publication of studies.

We want you to rate the importance of each criterion for considering whether to downgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Indexed literature search

The comprehensiveness of the review authors' search of indexed literature to identify eligible studies

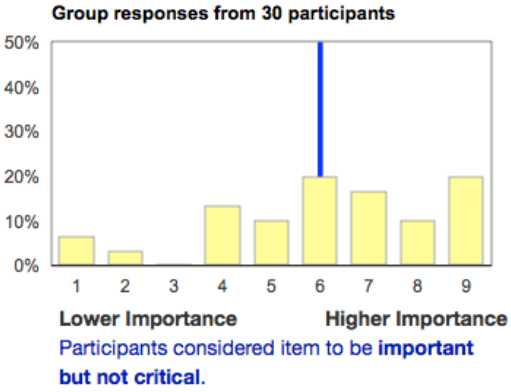


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Grey literature

The comprehensiveness of the review authors' search of grey literature to identify eligible studies

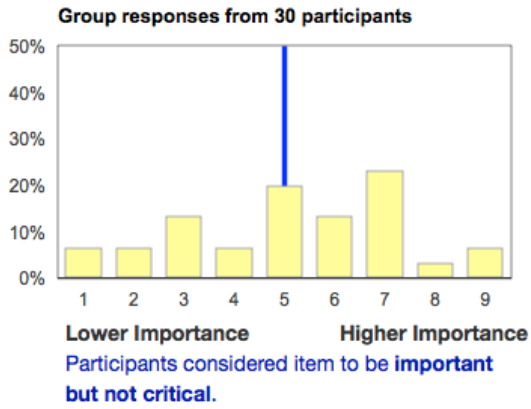


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Language of included manuscripts

Whether authors applied restrictions to study selection on the basis of language

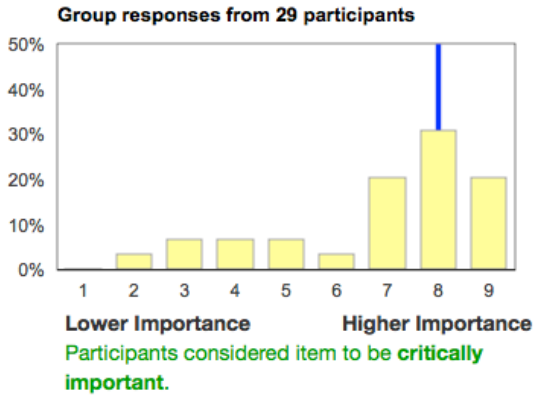


Lower Importance	1	2	3	4	5	6	7	8	9	Higher Importance
	○	○	○	○	○	○	○	○	○	

Please, provide the rationale behind your answer here

4. Study sponsorship

Whether developers and purveyors of the intervention had influence on studies included in the review

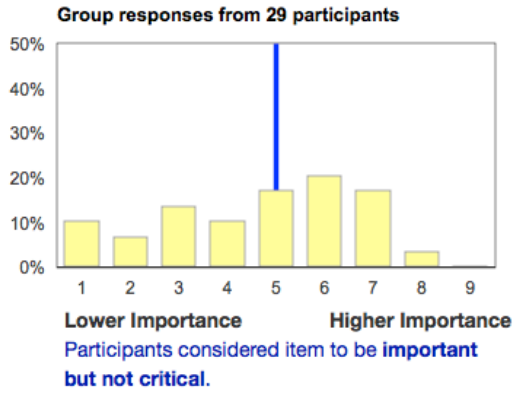


Lower Importance	1	2	3	4	5	6	7	8	9	Higher Importance
	○	○	○	○	○	○	○	○	○	

Please, provide the rationale behind your answer here

5. Number of small studies

Degree to which the body of evidence consists of studies with small sample sizes

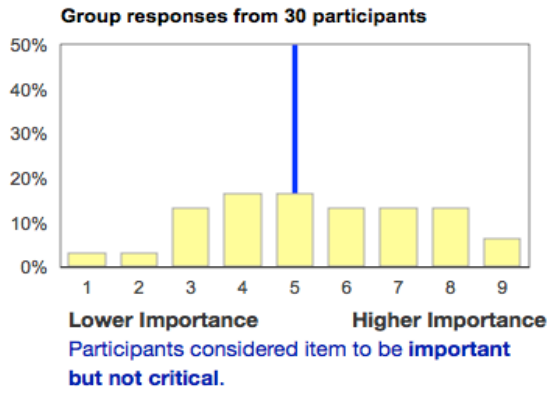


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Funnel plot asymmetry

Whether there was evidence of funnel plot asymmetry

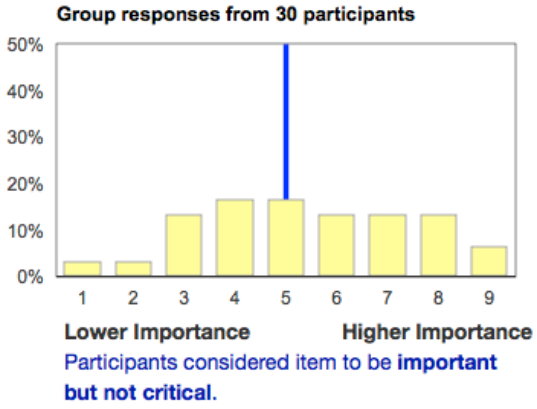


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

7. Discrepancies between published and unpublished studies

Results from any approaches to assess discrepancies in findings between published and unpublished studies



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Page 10

Rating the Quality of Evidence in Reviews of Complex Interventions (Panel B): Round Three

Round Three ends May 16, 2017 08:00 AM PT

In this study, you are Client X

Upgrading the Initial Quality Rating

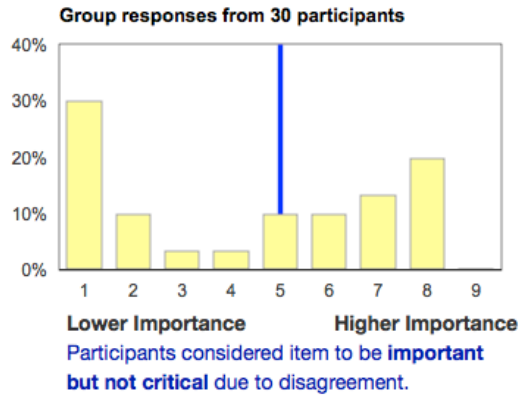
After assigning an initial rating for the quality of a body of evidence, many approaches assess criteria related to factors that could "upgrade" the initial quality rating (e.g., from "low" to "high") if that body of evidence has not been downgraded for any other reason.

We want you to rate the importance of each criterion for considering whether to upgrade the initial quality of the body of evidence for *all* reviews of complex interventions.

As a reminder, a rating of the quality of a body of evidence indicates the reviewers' confidence that an effect estimate for a specific outcome is correct, based on the body of evidence contributing to that effect estimate.

1. Large magnitude of an effect

Rating up the quality of a body of evidence from non-randomised studies that yield large or very large estimates of the magnitude of an intervention effect

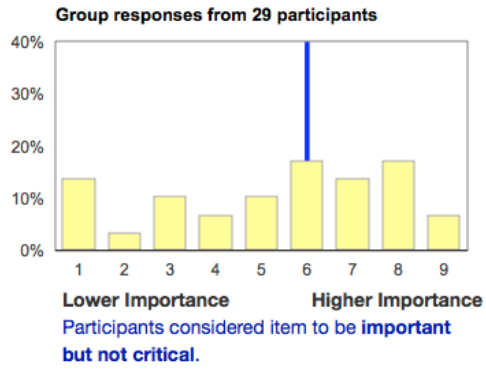


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

2. Dose-response gradient

Rating up the quality of a body of evidence from non-randomised studies when there is presence of a dose-response gradient

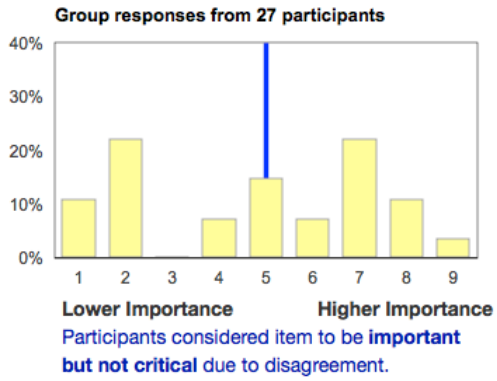


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

3. Effect of plausible residual confounding

Rating up the quality of a body of evidence from non-randomised studies when all plausible residual confounding from non-randomised studies are likely to reduce the demonstrated effect or increase the effect if no effect was observed

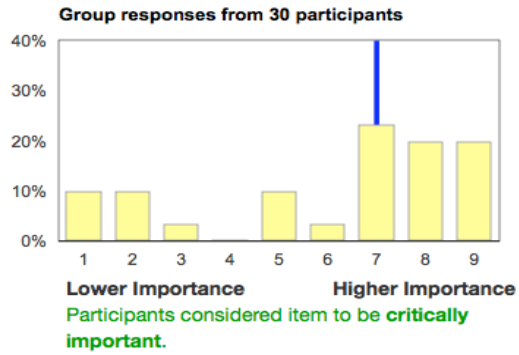


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

4. Consistency across diverse contexts

Rating up the quality of a body 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)

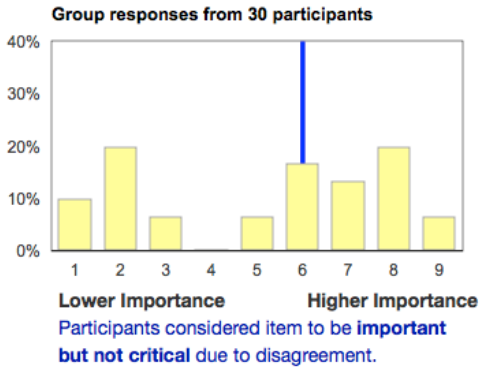


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

5. Analogous evidence

Rating up the quality of a body of evidence when there is supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanism(s)

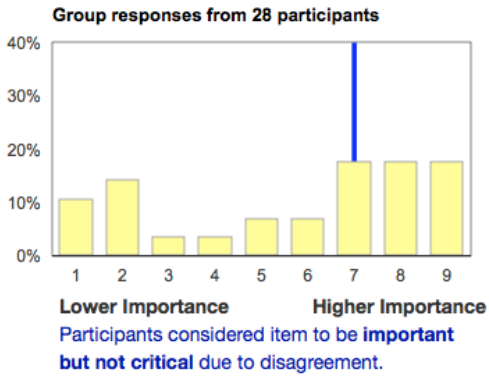


	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

6. Coherence of evidence for the causal pathway

Rating up the quality of a body of evidence when there is coherence of results in individual links in the causal pathway between intervention and distal outcomes



	1	2	3	4	5	6	7	8	9	
Lower Importance	○	○	○	○	○	○	○	○	○	**Higher Importance**

Please, provide the rationale behind your answer here

Appendix 11. List of codes (online expert panel)

Thematic analysis of the comments/discussions from the online expert panel

Descriptions of general categories and individual codes	Codes	Pre-defined
Features of complexity	**COMPLEXITY**	**Yes**
CI: variation in review elements	CI-VAR	No
CI: variation in intervention design	CI-VAR-INTERV	No
CI: variation on intervention implementation	CI-VAR-IMPLEMENT	No
CI: interactions with context	CI-CONTEXT	No
CI: use of diverse study designs	CI-DESIGN	No
CI: lack of intervention specification	CI-INTERV-REPORT-LACK	No
CI: lack of reporting on contextual factors	CI-CONTEXT-REPORT-LACK	No
CI: use of multiple interacting components	CI-MULT-COMPONENT	No
CI: use of a range of outcomes	CI-OUTCOMES-RANGE	No
CI: need to integrate qual and quant evidence	CI-QUAL+QUANT	No
CI: lack of meta-analysis	CI-META-LACK	No
CI: need for increased judgement	CI-JUDGE	No
CI: against algorithmic use of criteria	CI-ALGORITHM-NO	No
CI: systems perspective on complexity	CI-SYSTEMS	No
CI: thinking of complexity through dimensions	CI-DIMENSIONS	No
CI: thinking of complexity through questions	CI-QUESTIONS	No
Missing Criteria	**MISS**	**Yes**
MISS: guidance on narrative synthesis	MISS-NARRATIVE	No
MISS: lack of intervention specification	MISS-INTERV-REPORT-LACK	No
MISS: appreciating GRADE as it is, plus additional signalling questions	MISS-GRADE-PLUS	No
MISS: need to integrate qual and quant evidence	MISS-QUAL+QUANT	No
MISS: need to have something on theory and causal chain	MISS-THEORY	No
MISS: thinking of complexity through questions	MISS-QUESTIONS	No
Rating a mixed body of evidence	**MIX**	**Yes**
MIX: EBM-linear vs. Bayesian stepwise	MIX-LINEAR vs. STEP	No
MIX: challenges of GRADE hierarchy	MIX-GRADE-HIERARCHY	No

MIX: challenges of assessing RoB in different NRS designs	MIX-RoB-NRS	No
MIX: challenges with upgrading	MIX-UPGRADE	No
MIX: no challenges if appropriately weighed	MIX-WEIGH	No
MIX: challenges of interpreting inconsistent results	MIX-INCONSISTENT	No
MIX: consistency adds confidence	MIX-CONSISTENT	No
MIX: against mixing	MIX-AGAINST	No
MIX: use design for moderator analysis	MIX-MODERATOR	No
MIX: not fit for a meta-analytic paradigm	MIX-META-FIT-NO	No
MIX: requires more methodological expertise	MIX-EXPERTISE	No
MIX: challenges associated with aggregation/pooling	MIX-POOL	No
Qualification for using the guidance	**QUAL**	**Yes**
QUAL: requires postgraduate qualification	QUAL-POSTGRAD	No
QUAL: requires experience in research methods	QUAL-MEHTOD	No
QUAL: requires experience in reviewing	QUAL-REVIEW	No
QUAL: requires expertise in the subject	QUAL-SUBJECT	No
QUAL: requires methodological expertise	QUAL-METHODS	No
QUAL: requires training in GRADE	QUAL-GRADE-TRAIN	No
Dissemination of the guidance	**DISSEM**	**Yes**
DISSEM: through major organisations in evidence synthesis	DISSEM-ORG	No
DISSEM: recommends provision of training opportunities	DISSEM-TRAIN	No
DISSEM: recommends provision of training resources	DISSEM-RESOURCE	No
DISSEM: recommends provision of examples	DISSEM-EXAMPLE	No
DISSEM: through journals/editors	DISSEM-JOURNAL	No
DISSEM: email teaching course leaders	DISSEM-COURSE	No
DISSEM: the guidance would need to be revised in light of further understanding of complexity	DISSEM-REVISE	No
DISSEM: through a software for easy use	DISSEM-SOFT	No
Initial quality rating: study design	**DESIGN**	**Yes**
DESIGN: risk of bias as a quality indicator	DESIGN-RISK	No
DESIGN: randomisation is a gold standard	DESIGN-GOLD	No
DESIGN: RCTs are less prone to bias	DESIGN-RCT-BIAS	No
DESIGN: quick and dirty sorting of evidence	DESIGN-QUICK	No
DESIGN: weak versus strong designs	DESIGN-W/S	No
DESIGN: issues with upgrading	DESIGN-UP	No
DESIGN: potential for overrating	DESIGN-OVER	No
DESIGN: confusion in defining NRS	DESIGN-NRS-MIX	No
DESIGN: criterion 4 is difficult to operationalise	DESIGN-OPERATION	No
DESIGN: certain NRS are less prone to bias	DESIGN-NRS-BIAS	No
Limitations of the included studies	**STUDY**	**Yes**

STUDY: supporting empirical evidence	STUDY-EVIDENCE	No
STUDY: Critical for judging risk of bias	STUDY-BIAS	No
STUDY: achievable for complex interventions	STUDY-ACHIEVE	No
STUDY: uncertainty about relevance	STUDY-UNCERTAIN	No
STUDY: impossible to assess in complex interventions	STUDY-IMPOSSIBLE	No
STUDY: use of multiple criteria	STUDY-MULTIPLE	No
STUDY: importance of the criterion is dependent on other considerations	STUDY-DEPEND	No
Inconsistency of effects in the evidence base	**IC**	**Yes**
IC: expected high levels of heterogeneity in systematic reviews of complex interventions	IC-HIGH	No
IC: assessment of heterogeneity is not an issues of quality	IC-QUALITY	No
IC: investigation of heterogeneity	IC-INVESTIGATE	No
IC: investigation using quantitative methods	IC-QUANT	No
IC: investigation using qualitative methods	IC-QUAL	No
IC: need for further guidance	IC-FURTHER	No
IC: using the criteria together	IC-TOGETHER	No
IC: automatic downgrading of evidence	IC-AUTOMATIC	No
IC: direction of the effect	IC-DIRECTION	No
IC: statistical test for heterogeneity	IC-STATS	No
Indirectness of the evidence base	**ID**	**Yes**
ID: the criterion is deemed more suitable for assessing evidence relevance vs. evidence quality	ID-REV/QUAL	No
ID: dependability of the importance of the criterion on the degree of differences	ID-DEPEND-DEGREE	No
ID: excluding evidence from reviews when the criterion is not met, rather than downgrading evidence	ID-EXCLUSION	No
ID: challenges associated with operationalising the criterion	ID-CHALLENGE-OPERATE	No
ID: importance for reasons of evidence generalisability	ID-GENERALISE	No
ID: the criterion is not specific for CIs	ID-NOT-SPECIFIC	No
ID: referring to empirical evidence for justify the importance of the criterion	ID-EVIDENCE	No
ID: importance of the criterion for reasons of adequately using the generated evidence	ID-USABILITY	No
ID: importance of the criterion for equity considerations	ID-EQUITY	No
ID: importance of Comparability vs. Representativeness aspects of indirectness assessment	ID-COMP vs. REPRESENT	No
ID: the criterion is deemed problematic to assess in reviews of CIs	ID-PROBLEMATIC	No

ID: dependability of the importance of the criterion on a specific review context	ID-DEPEND-REVIEW	No
ID: the criterion is deemed more suitable for assessing quality of systematic reviewing vs. quality of a body of evidence	ID-SR vs. BODY	No
ID: CIs are expected to have differences in implementation	ID-CI-IMPLEMENT	No
ID: important to understand the differences in reviews of CIs rather than to downgrade evidence	ID-UNDERSTAND	No
ID: extrapolation of evidence is deemed to be error-prone	ID-ERROR	No
ID: the criterion is challenging to assess for CIs	ID-CHALLENGE-CI	No
ID: need to better specify CIs in reviews	ID-SPEC-CI	No
ID: the criterion is deemed more suitable for assessing the scope of evidence vs. quality of a body of evidence	ID-SCOPE/BODY	No
ID: dependability of the importance of the criterion on the nature of the outcomes	ID-DEPEND-OUTCOME	No
ID: importance of considering intermediate outcome in reviews of CIs	ID-INTERMEDIATE	No
ID: importance of considering the causal chain in reviews of CIs	ID-CHAIN	No
ID: surrogate outcomes are deemed less informative	ID-SURROGATE	No
ID: the criterion is deemed more suitable for assessing the external validity vs. the internal validity	ID-EXTERNAL vs. INTERNAL	No
ID: the question/criterion is not well understood by the respondent	ID-UNDERSTAND-LACK	No
ID: comparators are not well defined in reviews	ID-COMP-DEFIN	No
ID: the criterion is deemed important, but not the most important	ID-NOT-MOST-IMPORT	No
ID: the criterion is deemed important in terms of assuring the longevity of the effects	ID-LONGEVITY	No
ID: the criterion is considered a more objective standard of indirectness	ID-OBJECTIVE	No
Imprecision of the effect estimates	**IP**	**Yes**
IP: uncertainty on whether the criterion warrants downgrading evidence	IP-DOWNGRADE	No
IP: the criterion is suggested to be combined with other IP criteria	IP-COMBINE	No
IP: highlights importance to assess variation in the direction of effect	IP-DIRECTION	No
IP: the criterion is deemed not specific for reviews of CIs	IP-NOT-SPECIFIC	No
IP: highlights the importance to also consider prediction intervals	IP-PI	No

IP: highlights the importance of clinically meaningful effects	IP-MEANINGFUL	No
IP: highlights the importance of expected variability over precision	IP-EXPECTED vs. PRECISE	No
IP: the importance of the criterion is deemed dependable on the power	IP-DEPEND-POWER	No
IP: the criterion is deemed irrelevant for assessing evidence quality	IP-QUALITY-IRRELEVANCE	No
IP: the criterion is deemed particularly important for reviews of CIs	IP-IMPORTANT-CIs	No
IP: the criterion is deemed to be an objective measure of precision	IP-OBJECTIVE	No
IP: highlights the importance of the criterion with regard to its impact on other considerations of precision	IP-CONSEQUENCE	No
IP: the importance of the criterion is deemed dependable on the study design	IP-DEPEND-DESIGN	No
IP: highlights the importance of power of individual studies over the review sample size	IP-POWER-INDIVIDUAL	No
IP: the importance of the criterion is questionable in reviews with multiple outcomes	IP-QUESIONABLE	No
IP: the criterion is perceived to be rarely used	IP-RARE	No
IP: the criterion is deemed unrealistic for reviews of CIs	IP-UNREALISTIC	No
IP: the criterion is perceived to ignore allocation at higher level units	IP-HIGHER-LEVELS	No
Publication bias	**PB**	**Yes**
PB: challenges associated with operationalising the criterion	PB-CHALLENGE-OPERATION	No
PB: challenges associated with assessing the criterion	PB-ASSESS	No
PB: the criterion is not perceived to inform the publication bias	PB-NOT	No
PB: the criterion is not perceived to be specific for complex interventions	PB-CI-NOT-SPECIFIC	No
PB: referring to empirical evidence for justify the rating of the criterion	PB-EVIDENCE	No
PB: referring to the lack of empirical evidence to justify the rating	PB-EVIDENCE-LACK	No
PB: referring to the mixed evidence to justify the rating	PB-EVIDENCE-MIXED	No
PB: highlights the importance to look at grey literature	PB-GREY	No
PB: the criterion is perceived to be less important with emergence of new methods	PB-EMER-METHODS	No
PB: the importance of the criterion is deemed dependable on the context	PB-DEPEND-CONTEXT	No

PB: highlights the need to use a judgement to determine the importance of the criterion	PB-JUDGEMENT	No
PB: lack of the understanding of the Q	PB-UNDERSTAND-LACK	No
PB: the criterion is not perceived to yield high quality research	PB-QUALITY-LACK	No
PB: the criterion is perceived to be the most important after searched in the indexed literature	PB-II-INDEXED	No
PB: the importance of the criterion is deemed dependable on the topic/scope of the review	PB-DEPEND-TOPIC	No
PB: the criterion is deemed feasible to assess	PB-FEASIBLE	No
PB: the criterion is not deemed feasible to assess	PB-FEASIBLE-LACK	No
PB: the criterion is not deemed to produce bias	PB-BIAS-FREE	No
PB: uncertainties are raised with regard to the importance of the criterion	PB-UNSURE	No
PB: CIs are deemed to have small sample sizes	PB-CI-SAMPLE-SMALL	No
PB: the criterion is perceived to be the most useful indicator of PB	PB-MOST-USEFUL	No
PB: the criterion is recommended to be an indicator of PB, but not a guarantee	PB-INDICATOR	No
PB: the criterion is deemed to be inapplicable to evidence with low number of included studies	PB-INAPPLICABLE	No
PB: the criterion is deemed to be problematic because of multiple possible causes of asymmetry	PB-MULTIPLE-CAUSE	No
Upgrading the initial quality rating	**UP**	**Yes**
UP: the criterion accords with those of Bradford Hill	UP-HILL	No
UP: need to broaden the definition of the criterion	UP-BROADEN	No
UP: need to define the dose for complex interventions	UP-DOSE-CI	No
UP: the criterion is perceived difficult to define/detect for CIs	UP-DIFFICULT-CI	No
UP: the criterion is deemed irrelevant for CIs	UP-IRRELEVANT-CI	No
UP: the criterion is perceived to be unconvincing	UP-UNCONVINCING	No
UP: the criterion is perceived to increase the confidence in CIs	UP: CONFIDENCE-CI	No
UP: the criterion is perceived to contribute to inclusion of the systems perspective in systematic reviews	UP: SYSTEMS	No
UP: the criterion is perceived to go beyond black-box assessment	UP-BLACK-BOX	No
UP: the criterion is perceived to help in identifying important gaps	UP-GAPS	No
UP: the criterion needs to be operationalised	UP-NEED-OPERATION	No
UP: the criterion is perceived to results in overuse	UP-OVERUSE	No

UP: the criterion is perceived to be involve subjective judgement	UP-SUBJECTIVE	No
UP: empirical evidence is brought against the importance of the criterion	UP-EVIDENCE-AGAINST	No
UP: the criterion is perceived to be rarely used in CIs	UP-RARE-CI	No
UP: the criterion is perceived to rely on many assumptions	UP-ASSUMPTIONS	No
UP: the criterion is perceived to belong to risk of bias assessment	UP-RoB	No
UP: the criterion is perceived difficult to understand	UP-DIFFICULT-UNDERSANTD	No
UP: the criterion is perceived to assess replicability of findings	UP-REPLICABILITY	No
UP: uncertainties are raised with regard to the importance of the criterion	UP-UNCERTAIN-IMPORTANCE	No
UP: the criterion is perceived to belong to imprecision assessment	UP-IMPRECISION	No
UP: the importance of the criterion is undermined by a likely consistently biased studies	UP-CONSISTENT-BIAS	No
UP: the criterion is perceived to be a matter of evidence interpretation	UP-INTERPRET	No
UP: the criterion is perceived to be conceptually similar to inconsistency assessment	UP-INCONSISTENCY	No
UP: the criterion is perceived to include too many dimensions	UP-HEAVY	No
UP: the criterion is perceived to be difficult to operationalise	UP-DIFFICULT-OPERATION	No
UP: the criterion needs to be appropriately interpreted	UP-INTERPRET-APPROPRIATE	No
UP: the criterion is not perceived to be a quality issue	UP-NOT-QUALITY	No
UP: the criterion is perceived to be due to biases	UP-BIASES	No
UP: the criterion is perceived to require a good theory and justification	UP-JUSTIFICATION	No
UP: the importance of the criterion is deemed to be dependent on the quality of evidence	UP-DEPEND-QUALITY	No
UP: the criterion is perceived to relate to a different body of evidence	UP-DIFFERENT-BODY	No
UP: the criterion is perceived to lack evidential support	UP-EVIDENCE-LACK	No
UP: the criterion is perceived to be subject to speculation	UP-SPECULATE	No
UP: the criterion is perceived to help explain the evidence, rather than rate confidence in the effects	UP-EXPLAIN-NOT-RATE	No

Appendix 12. IPR and IPRAS values

Round Three online expert panel

Criterion	IPR	IPRAS	DI
1: An initial quality rating of "high" when the body of evidence consists of randomized controlled trials (RCTs)	2	6.85	0.29
2: An initial quality rating of "moderate" when the body of evidence consists of non-randomized experimental study designs (e.g., natural experiments, quasi-experimental studies)	3	3.1	0.97
3: An initial quality rating of "low" when the body of evidence consists of non-experimental observational studies (e.g., cohort design, case-control study)	3	4.6	0.65
4: An initial rating of "high" for a body of evidence consisting of any type of study design	2	6.85	0.29
5: 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.)	2	6.85	0.29
6: Whether providers of the intervention are aware of the arm to which participants have been allocated	1.8	3.7	0.49
7: Whether recipients of the intervention are aware of the arm to which they have been allocated	1	3.1	0.32
8: Whether those assessing outcomes are aware of the arm to which participants have been allocated	2	6.85	0.29
9: Whether those analyzing data are aware of the arm to which participants have been allocated	2	5.35	0.37
10: 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	1	7.6	0.13
11: Whether there is incomplete or absent reporting of some outcomes and not others on the basis of the results	1	7.6	0.13
12: Whether studies have been ended early when beneficial results were found in interim analyses	2	3.85	0.52
13: Whether studies used outcome measures with low validity and/or reliability to assess intervention effects	1	6.1	0.16
14: Whether there were deviations in intervention implementation from what was intended	2	3.85	0.52
15: Whether the study used measures to adequately control for confounding	1	7.6	0.13
16: Whether the study developed and applied an appropriate comparison group	1	7.6	0.13

17: Whether the study used procedures to appropriately select participants into the study or into the analysis	2	6.85	0.29
18: Whether intervention groups were clearly defined	1.6	6.55	0.24
19: Whether there were deviations from the intended intervention beyond what would be expected in usual practice	2	3.85	0.52
20: Whether the study used appropriate analytic methods for dealing with participant missing data	1.2	6.25	0.19
21: Whether the study appropriately measured outcomes in all groups	1	7.6	0.13
22: Whether there is selected reporting of effect estimates based on multiple measurements and analyses	1.2	7.45	0.16
23: Whether the study had adequate follow-up of participants	2	5.35	0.37
24: The degree to which point estimates vary across individual studies	2	3.85	0.52
25: The degree to which confidence intervals overlap across individual studies	2.6	3.4	0.77
26: The magnitude of the P-value for a statistical test of the null hypothesis that all studies have the same underlying magnitude of effect	3	4.6	0.65
27: The magnitude of the I2 value, which indicates the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance)	3	3.1	0.97
28: Results of pre-specified quantitative analyses exploring moderators or methodological features that help explain heterogeneity (e.g., sub-group analyses, sensitivity analyses, meta-regressions)	1.6	4.15	0.39
29: Results of qualitative analyses of evidence exploring varying effects of interventions that help explain heterogeneity (e.g., qualitative comparative analysis)	2	3.85	0.52
30: Degree to which the participants in included studies compare to the population of interest	1	6.1	0.16
31: Degree to which the interventions in included studies compare to the intervention of interest	1	6.1	0.16
32: Degree to which the outcomes considered in included studies compare to the outcomes of interest	1	6.1	0.16
33: Degree to which the timing of outcome assessments in included studies compares to the follow-up time-points of interest	2	5.35	0.37
34: Degree to which effect estimates are from comparison groups of interest	1.9	5.43	0.35
35: Whether the total number of participants in the review meets a conventional sample size for a single adequately powered trial	2	3.85	0.52

36: Whether the confidence interval for the overall estimate includes effects indicating both benefit and harm	2	3.85	0.52
37: Whether the confidence interval includes estimates of important benefit and important harm	1.3	5.88	0.22
38: The comprehensiveness of the reviewer authors' search of indexed literature to identify eligible studies	1.7	7.08	0.24
39: The comprehensiveness of the reviewer authors' search of grey literature to identify eligible studies	1.6	4.15	0.39
40: Whether authors applied restrictions to study selection on the basis of language	1	3.1	0.32
41: Whether developers and purveyors of the intervention had influence on studies included in the review	2	6.85	0.29
42: Degree to which the body of evidence consists of studies with small sample sizes	1.3	2.88	0.45
43: Whether there was evidence of funnel plot asymmetry	2	2.35	0.85
44: Results from any approaches to assess discrepancies in findings between published and unpublished studies	1	6.1	0.16
45: Rating up the quality of a body of evidence from non-randomized studies that yield large or very large estimates of the magnitude of an intervention effect	3	4.6	0.65
46: Rating up the quality of a body of evidence from non-randomized studies when there is presence of a dose-response gradient	1	4.6	0.22
47: Rating up the quality of a body of evidence from non-randomized studies when all plausible residual confounding from non-randomized studies are likely to reduce the demonstrated effect or increase the effect if no effect was observed	2	3.85	0.52
48: Rating up the quality of a body 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)	2.9	4.68	0.62
49: Rating up the quality of a body of evidence when there is supporting evidence from similar or "analogous" interventions that are known to operate through the same or similar mechanism(s)	3	3.1	0.97
50: Rating up the quality of a body of evidence when there is coherence of results in individual links in the causal pathway between intervention and distal outcomes	1.9	3.93	0.48

Notes: IPR = inter-percentile range; IPRAS = inter-percentile range adjusted for symmetry; DI = disagreement index

Appendix 13. Expert meeting agenda

24 May 2017	
12:00 – 13:00	Arrivals
13:00 – 14:30	Lunch
Session 1: Introduction & background	
15:00 – 15:15	1.1. Welcome/introductions
15:15 – 15:30	1.2. Overview of objectives
15:30 – 15:45	1.3. Process for securing GRADE approval of guidance developed by project groups
15:45 – 16:00	1.4. Background: rating the certainty of evidence in the GRADE approach
16:00 – 16:30	1.5. Defining complex interventions (in complex systems)
16:30 – 16:45	Coffee/tea
16:45 – 17:30	1.6. Focus of the new guidance: results from systematic review, interviews and Delphi exercise
17:30 – 18:00	1.7. Discussion
19:30	Dinner (Magdalen Arms)

25 May 2017	
07:00 – 09:00	Breakfast
Session 2: GRADE conceptual framework for complex interventions	
09:00 – 09:15	2.1. Review of Day 1 and aims of Day 2
09:15 – 09:30	2.2. Clarifying the conceptual framework of GRADE ratings: update on the recent publication of the GRADE Working Group
09:30 – 09:45	2.3. Conceptualising “certainty of evidence” for complex interventions: response to the publication presented above
09:45 – 10:45	2.4. Discussion: GRADE guidance for complex interventions
10:45 – 11:00	Coffee/Tea
Session 3: GRADE domains of evidence for complex interventions	
11:00 – 11:45	3.1. Study design (initial categorisation of evidence)
11:45 – 12:30	3.2. Risk of bias

12:30 – 13:30	Lunch
Session 3: GRADE domains of evidence for complex interventions (cont.)	
13:30 – 14:15	3.4. Indirectness
14:15 – 15:00	3.5. Inconsistency
15:00 – 15:15	Coffee/Tea
Session 3: GRADE domains of evidence for complex interventions (cont.)	
15:15 – 16:00	3.5. Imprecision
16:00 – 16:45	3.6. Publication bias
16:45 – 17:30	3.7. Upgrading domains
17:30 – 18:00	3.8. Additional domains and missing issues
19:30	Dinner (Oxford Spires Hotel)

26 May 2017	
07:00 – 09:00	Breakfast
Session 4: Finalising GRADE guidance for complex interventions	
09:00 – 09:15	4.1. Review of Day 2 and aims of Day 3
09:15 – 09:45	4.2. Key issues from Day 2
09:45 – 10:45	4.3. Finalising the GRADE guidance for complex interventions
10:45 – 11:00	Coffee/Tea
Session 5: Next steps (write-up & implementation)	
11:00 – 11:15	5.1. Drafting & piloting of the guidance
11:30 – 12:00	5.2. Discussion on dissemination activities: GRADE Working Group, WHO, Cochrane/Campbell Collaborations, etc.
12:00 – 12:30	5.3. Wrap-up of the main meeting
12:30 – 13:30	Lunch

Appendix 14. List of codes (expert meeting)

Thematic analysis of the discussions from the face-to-face expert meeting

Descriptive labels of general categories and individual codes	Codes	Pre-defined
Conceptualising complexity for the GRADE guidance for CIs	**COMP-GRADE-CI**	**Yes**
COMP: Different perspectives on complexity	COMP-DIFFER	Yes
COMP: Complex interventions perspective	COMP-INTERV	Yes
COMP: Complex systems perspective	COMP-SYSTEM	Yes
COMP: Added value of complexity perspective	COMP-VALUE	No
COMP: Complexity as a framework	COMP-FRAME	No
COMP: Complexity as a continuum	COMP-CONT	No
COMP: Occam's Razor for complexity	COMP-RAZOR	No
COMP: Tools for assessing and reporting complexity	COMP-TOOLS	No
Specifying the construct of certainty of evidence for the GRADE guidance for CIs	**CERT-GRADE-CI**	**Yes**
CERT: Different approaches	CERT-DIFFER	Yes
CERT: The value of different approaches	CERT-DIFFER-VALUE	No
CERT: Fully contextualised	CERT-FULLY	Yes
CERT: Partly contextualised	CERT-PARTLY	Yes
CERT: Non-contextualised	CERT-NON	Yes
CERT: Non-contextualised rating of null effect	CERT-NON-NULL	Yes
CERT: Non-contextualised rating of confidence intervals	CERT-NON-CI	Yes
CERT: Non-contextualised rating of prediction intervals	CERT-NON-PI	No
CERT: Most useful definition for complexity perspective	CERT-MOST USEFUL	No
CERT: Implications of specifying certainty of evidence	CERT-IMPLICATIONS	No
CERT: Rating certainty of evidence in the point estimate	CERT-POINT	No
CERT: Providing flexibility of options	CERT-FLEX	No
CERT: Being explicit about the adopted approach	CERT-EXPLICIT	No
Initial categorisation of evidence in the GRADE guidance for CIs	**INCAT-GRADE-CI**	**Yes**
INCAT: discussion on ROBINS-I	INCAT-ROBINS	Yes
INCAT: starting everything as high	INCAT-HIGH	Yes
INCAT: introducing some designs as moderate	INCAT-MODERATE	Yes
INCAT: classifying study designs	INCAT-CLASS	No

Rating Risk of bias in the GRADE guidance for CIs	**ROB-GRADE-CI**	**Yes**
ROB: blinding of participants and providers	ROB-BLINDING-PP	Yes
ROB: blinding of outcome assessors	ROB-BLINDING-OA	Yes
ROB: fidelity	ROB-FIDELITY	Yes
ROB: moving from individual studies to a body of evidence	ROB-BODY	No
ROB: GRADE and Cochrane tools	ROB-COCHRANE TOOLS	No
ROB: Meta-level bias	ROB-META	No
Rating inconsistency in the GRADE guidance for CIs	**IC-GRADE-CI**	**Yes**
IC: 5Es	IC-5Es	No
IC: relation with certainty definition	IC-CERT	No
IC: need to provide further examples	IC-EXAMPLES	No
IC: flexibility of options	IC-FLEX	No
IC: importance of logic models	IC-LOGIC	No
IC: importance of pre-specification	IC-PRESPEC	No
Rating indirectness in the GRADE guidance for CIs	**ID-GRADE-CI**	**Yes**
ID: chain of evidence discussions	ID-CHAIN	Yes
ID: rating of intermediate outcomes	ID-INTERMED	No
ID: correlation of outcome	ID-CORRELATION	No
ID: rating different bodies of evidence	ID-DIFFER	No
ID: lumping vs. splitting	ID-LUMP vs. SPLIT	No
ID: Relevance vs. representativeness	ID-RELEV vs. REP	No
Rating imprecision in the GRADE guidance for CIs	**IP-GRADE-CI**	**Yes**
IP: assessment of OIS criterion	IP-OIS	Yes
IP: size of the body of evidence	IP-SIZE	Yes
IP: relation with certainty definition	IP-CERT	No
IP: differentiation from inconsistency assessment	IP-IC	No
Rating publication bias in the GRADE guidance for CIs	**PB-GRADE-CI**	**Yes**
PB: industry sponsorship	PB-INDUSTRY	No
PB: developer bias	PB-DEVELOPER	No
PB: allegiance bias	PB-ALLEGIANCE	No
PB: comprehensiveness of search	PB-COMPREHENSIVE	No
PB: multilingual searches	PB-LANGUAGE	No
PB: searching protocol databases	PB-PROTOCOL	No
PB: obtaining information about research funders	PB-FUNDERS	No
Upgrading evidence in the GRADE guidance for CIs	**UP-GRADE-CI**	**Yes**
UP: upgrading when starting as high	UP-HIGH	No

UP: upgrading for counteracting plausible confounding	UP-PLAUSE	Yes
UP: upgrading for dose-response	UP-DOSE	Yes
UP: upgrading for large magnitude of effect	UP-LARGE	Yes
Additional considerations for the GRADE guidance for CIs	**ADD-GRADE-CI**	**Yes**
ADD: assessing casual pathway of an intervention	ADD-CAUSAL	Yes
Drafting the GRADE guidance for CIs	**DRAFT-GRADE-CI**	**No**
DRAFT: need to accommodate different disciplines	DRAFT-DISCIPLINES	No
DRAFT: need to specify the remit	DRAFT-REMIT	No
DRAFT: need to specify the timeline	DRAFT-TIMELINE	No
DRAFT: need to accommodate different organisations	DRAFT-ORGAN	No
DRAFT: need to provide examples outside of clinical medicine	DRAFT-EXAMPLES	No
DRAFT: need to provide many options to users	DRAFT-FLEX	No
DRAFT: need for feasibility	DRAFT-FEASIBLE	No
DRAFT: opportunity for bringing new thinking	DRAFT-NOVELTY	No
DRAFT: need to collaboration with relevant groups in GWG	DRAFT-COLLABORATE	No
DRAFT: need to collaborate with the WHO project	DRAFT-WHO PROJECT	No
DRAFT: following the format of the GRADE Equity guidelines	DRAFT-GRADE EQUITY	No
Dissemination of the GRADE guidance for CIs	**DISS-GRADE-CIs**	**No**
DISS: need to have more visibility outside of clinical medicine	DISS-VISIBLE	No
DISS: need to do more PR of GRADE	DISS-PR	No
DISS: need to provide training on how to use GRADE	DISS-TRAINING	No
DISS: for non-academic audiences	DISS-NON ACADEMIC	No
DISS: approval from GWG	DISS-GWG	No
Activities out outputs of the GRADE Working Group	**GWG**	**No**
GWG: description of the GRADE approach	GWG-GRADE	No
GWG: papers published by GWG	GWG-PAPERS	No
GWG: weighting different domains in GRADE	GWG-WEIGHT	No
GWG: clarifying the upgrading domains	GWG-UP	No
GWG: need to exercise judgement	GWG-JUDGE	No
GWG: approval by GWG	GWG-APPROVE	No
GWG: project groups	GWG-GROUPS	No
Areas for future research	**FUTURE**	**No**

FUTURE: decision-making processes	FUTURE-DECISION	No
FUTURE: systems thinking	FUTURE-SYSTEMS	No
FUTURE: reviewing mixed bodies of evidence	FUTURE-MIXED	No
FUTURE: upgrading evidence when starting high	FUTURE-UP	No
FUTURE: assessing non-randomised studies	FUTURE-NON RANDOM	No
FUTURE: extending the GRADE categories of evidence	FUTURE-GRADE SCALE	No
FUTURE: communicating low certainty ratings	FUTURE-COMMUN LOW	No
FUTURE: extending ROBINS-I tools	FUTURE-ROBINS-I	No
Challenges of reviews and GRADE use	**CHALLENGE**	**No**
CHALLENGE: lack of specification	CHALLENGE-SPECIFICATION	No
CHALLENGE: lack of permeation of GRADE outside of clinical practice	CHALLENGE-PERMEATE	No
CHALLENGE: lack of discriminant validity	CHALLENGE-DISCRIMINANT	No
CHALLENGE: lack of granularity	CHALLENGE-GRANULARITY	No
CHALLENGE: inflexibility of GRADE	CHALLENGE-INFLEXIBLE	No
CHALLENGE: lack of understanding of GRADE by users	CHALLENGE-UNDERSTAND	No
Codes not categorsised		
Specifying the review purpose	REVIEW PURPOSE	No
Lumping vs. splitting debate in systematic reviews	LUMP vs. SPLIT	No
Thinking outside of the GRADE box	GRADE BOX	No
Fine-tuning of the GRADE approach	GRADE FINE-TUNE	No
Defining logic models	DEFINE-LOGIC MODELS	No
Added value of the GRADE guidance for CIs	GUIDACNCE-ADDED VALUE	No
The audience of the GRADE guidance for CIs	GUIDANCE-AUDIENCE	No
The aims of the GRADE guidance for CIs	GUIDANCE-AIMS	No