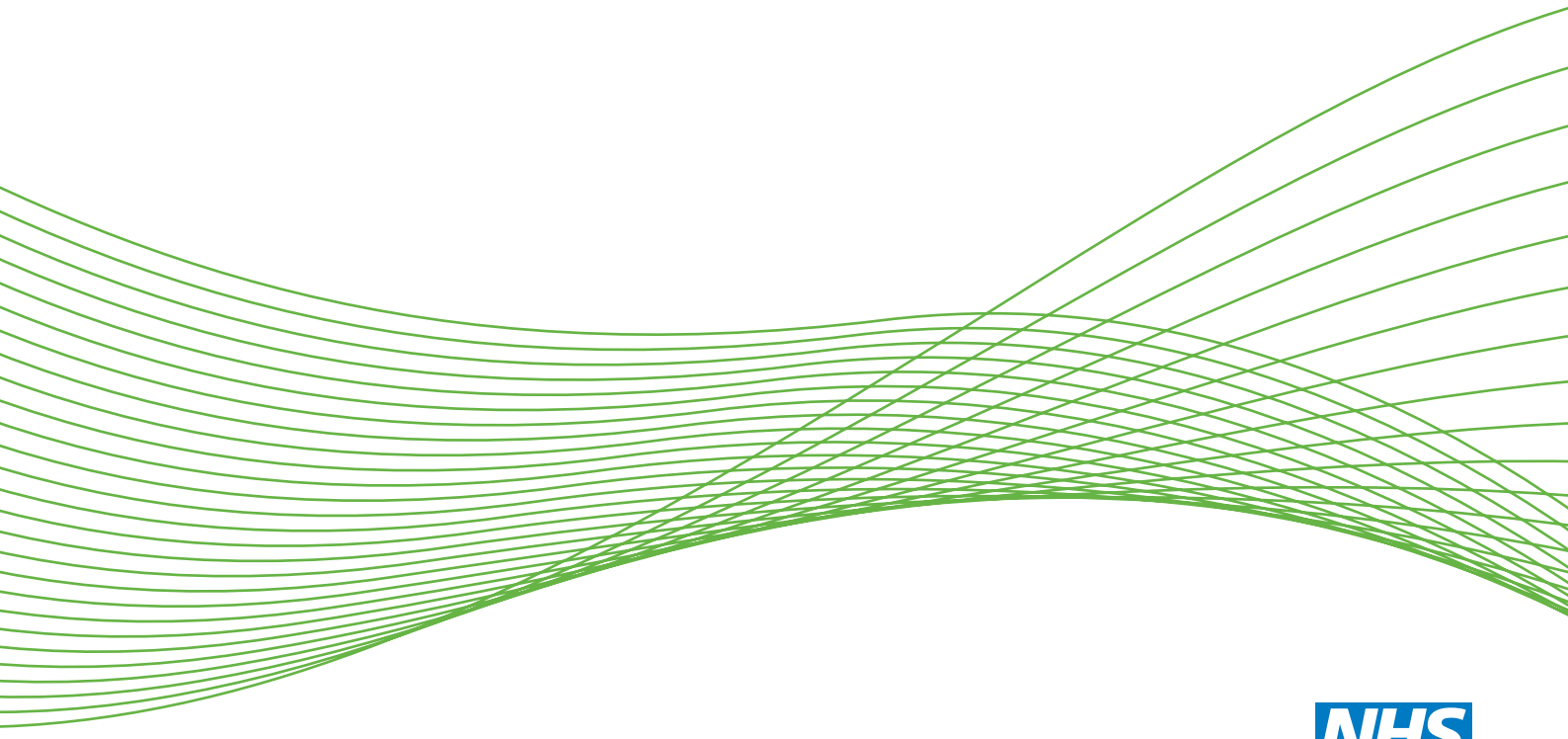
HEALTH SERVICES AND DELIVERY RESEARCH

VOLUME 5 ISSUE 28 OCTOBER 2017
ISSN 2050-4349

Quality and reporting standards, resources, training materials and information for realist evaluation: the RAMESES II project

Geoff Wong, Gill Westhorp, Joanne Greenhalgh, Ana Manzano, Justin Jagosh and Trisha Greenhalgh



NHS
National Institute for Health Research

DOI 10.3310/hsdr05280

Quality and reporting standards, resources, training materials and information for realist evaluation: the RAMESES II project

Geoff Wong,[1]* Gill Westhorp,[2] Joanne Greenhalgh,[3] Ana Manzano,[3] Justin Jagosh[4] and Trisha Greenhalgh[1]

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[2]Realist Research Evaluation and Learning Initiative, Charles Darwin University, Darwin, NT, Australia
[3]Sociology and Social Policy, University of Leeds, Leeds, UK
[4]Centre for Advancement in Realist Evaluation and Syntheses (CARES), University of Liverpool, Liverpool, UK

*Corresponding author

Declared competing interests of authors: Geoff Wong is a member of the National Institute for Health Research Health Technology Assessment programme Primary Care Panel, and is a panel member of the Health and Safety Executive External Peer Review Panel Evaluation Governance Group. During the course of the project Gill Westhorp worked as a consultant and consulting academic undertaking realist evaluations and reviews, and provided some capacity building and some PhD supervision on a commercial basis. These activities were not undertaken under the auspices of this project.

Published October 2017
DOI: 10.3310/hsdr05280

This report should be referenced as follows:

Wong G, Westhorp G, Greenhalgh J, Manzano A, Jagosh J, Greenhalgh T. Quality and reporting standards, resources, training materials and information for realist evaluation: the RAMESES II project. *Health Serv Deliv Res* 2017;**5**(28).

Health Services and Delivery Research

ISSN 2050-4349 (Print)

ISSN 2050-4357 (Online)

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HS&DR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hsdr. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Services and Delivery Research* journal

Reports are published in *Health Services and Delivery Research* (HS&DR) if (1) they have resulted from work for the HS&DR programme or programmes which preceded the HS&DR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

HS&DR programme

The Health Services and Delivery Research (HS&DR) programme, part of the National Institute for Health Research (NIHR), was established to fund a broad range of research. It combines the strengths and contributions of two previous NIHR research programmes: the Health Services Research (HSR) programme and the Service Delivery and Organisation (SDO) programme, which were merged in January 2012.

The HS&DR programme aims to produce rigorous and relevant evidence on the quality, access and organisation of health services including costs and outcomes, as well as research on implementation. The programme will enhance the strategic focus on research that matters to the NHS and is keen to support ambitious evaluative research to improve health services.

For more information about the HS&DR programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hsdr

This report

The research reported in this issue of the journal was funded by the HS&DR programme or one of its preceding programmes as project number 14/19/19. The contractual start date was in March 2015. The final report began editorial review in March 2017 and was accepted for publication in July 2017. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HS&DR editors and production house have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the final report document. However, they do not accept liability for damages or losses arising from material published in this report.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Abstract

Quality and reporting standards, resources, training materials and information for realist evaluation: the RAMESES II project

Geoff Wong,[1]* Gill Westhorp,[2] Joanne Greenhalgh,[3] Ana Manzano,[3] Justin Jagosh[4] and Trisha Greenhalgh[1]

[1]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[2]Realist Research Evaluation and Learning Initiative, Charles Darwin University, Darwin, NT, Australia
[3]Sociology and Social Policy, University of Leeds, Leeds, UK
[4]Centre for Advancement in Realist Evaluation and Syntheses (CARES), University of Liverpool, Liverpool, UK

*Corresponding author grckwong@gmail.com

Background: Many of the problems confronting policy- and decision-makers, evaluators and researchers today are complex, as are the interventions designed to tackle them. Their success depends both on individuals' responses and on the wider context of people's lives. Realist evaluation tries to make sense of these complex interventions. It is a form of theory-driven evaluation, based on realist philosophy, that aims to understand why these complex interventions work, how, for whom, in what context and to what extent.

Objectives: Our objectives were to develop (a) quality standards, (b) reporting standards, (c) resources and training materials, (d) information and resources for patients and other lay participants and (e) to build research capacity among those interested in realist evaluation.

Methods: To develop the quality and reporting standards, we undertook a thematic review of the literature, supplemented by our content expertise and feedback from presentations and workshops. We synthesised findings into briefing materials for realist evaluations for the Delphi panel (a structured method using experts to develop consensus). To develop our resources and training materials, we drew on our experience in developing and delivering education materials, feedback from the Delphi panel, the RAMESES JISCMail e-mail list, training workshops and feedback from training sessions. To develop information and resources for patients and other lay participants in realist evaluation, we convened a group consisting of patients and the public. We built research capacity by running workshops and training sessions.

Results: Our literature review identified 152 realist evaluations, and when 37 of these had been analysed we were able to develop our briefing materials for the Delphi panel. The Delphi panel comprised 35 members from 27 organisations across six countries and five disciplines. Within three rounds, the panels had reached a consensus on 20 key reporting standards. The quality standards consist of eight criteria for realist evaluations. We developed resources and training materials for 15 theoretical and methodological topics. All resources are available online (www.ramesesproject.org). We provided methodological support to 17 projects and presentations or workshops to help build research capacity in realist evaluations to 29 organisations. Finally, we produced a generic patient information leaflet for lay participants in realist evaluations.

Limitations: Our project had ambitious goals that created a substantial workload, leading to the need to prioritise objectives. For example, we truncated the literature review and focused on standards and training material development.

Conclusions: Although realist evaluation holds much promise, misunderstandings and misapplications of it are common. We hope that our project's outputs and activities will help to address these problems. Our resources are the start of an iterative journey of refinement and development of better resources for realist evaluations. The RAMESES II project seeks not to produce the last word on these issues, but to capture current expertise and establish an agreed state of the science. Much methodological development is needed in realist evaluation but this can take place only if there is a sufficient pool of highly skilled realist evaluators. Capacity building is the next key step in realist evaluation.

Funding: The National Institute for Health Research Health Services and Delivery Research programme.

Contents

List of tables

List of figures

List of boxes

List of abbreviations

CINAHL	Cumulative Index to Nursing and Allied Health Literature
CMOC	context–mechanism–outcome configuration
CPCI-S	Conference Proceedings, Citation Index – Science
ERIC	Education Resources Information Center
IQR	interquartile range
NIHR	National Institute for Health Research
RDS	Research Design Service
SCI	Science Citation Index
SSCI	Social Science Citation Index

Plain English summary

Realist evaluation is used to answer questions such as 'what works for whom, in what circumstances, how and why?' It is an approach to evaluating interventions or programmes in health and other fields. When we started this project, there were no standards setting out how to judge if realist evaluations were of high quality – something we have called quality standards. Nor did any standards exist to guide evaluators on how best to write up their evaluations – we have called these reporting standards. Although there were some resources and training materials for realist evaluation, more were needed that showed evaluators in detail how to rigorously undertake certain parts of an evaluation.

In this project, we developed quality and reporting standards and resources and training materials for realist evaluations. We used a range of methods (e.g. a review of the literature and a structured consensus-building process called a Delphi panel) to help us choose and agree on what should be in the standards and training materials. We used a pre-existing e-mail list for additional input. We asked researchers we worked with on realist evaluations for their comments, and we got feedback from researchers we trained in workshops or presented to at conferences. We analysed and wove together all this information to produce quality and reporting standards and resources and training materials. We needed to prioritise certain parts of the project as a result of its ambitious nature and the workload this created. We have made all of our project's outputs freely available online (www.ramesesproject.org).

Scientific summary

Background

Many of the problems confronting policy- and decision-makers, evaluators and researchers today are complex. For example, much health service need results from the effects of smoking, suboptimal diets (including obesity), excessive alcohol intake, inactivity or adverse family circumstances (e.g. partner violence), all of which, in turn, have multiple causes operating at both individual and societal level. Interventions or programmes designed to tackle such problems are themselves complex, with multiple, interconnected components delivered individually or targeted at communities or populations. Their success depends both on individuals' responses and on the wider context in which people strive (or not) to live meaningful and healthy lives. What works in one family, one organisation or one city may not work in another.

Designing and evaluating complex interventions is challenging. Randomised trials that compare 'intervention on' with 'intervention off', and their secondary research equivalent, meta-analyses of such trials, may produce statistically accurate statements (e.g. that the intervention works 'on average'), but these leave us none the wiser about where to target resources or how to maximise impact.

Realist evaluation seeks to address these problems. It is a form of theory-driven evaluation, based on realist philosophy, and it aims to advance understanding of why these complex interventions work, how, for whom, in what context and to what extent, as well as to explain the many situations in which a programme fails to achieve the anticipated benefit.

Realist evaluation assumes both that social systems and structures are 'real' (because they have real effects) and that human actors respond differently to interventions in different circumstances. To understand how an intervention might generate different outcomes in different circumstances, realism introduces the concept of mechanisms, which may be helpfully conceptualised as underlying changes in the reasoning and behaviour of participants who are triggered in particular contexts.

This project aims to develop quality and reporting standards, resources and training materials, to build research capacity and to develop materials for lay participants involved in realist evaluations.

Objectives

1. Recruit an interdisciplinary Delphi panel of, for example, researchers, support staff, policy-makers, patient advocates and practitioners with various types of experience relevant to realist evaluation.
2. Summarise the current literature and expert opinion on best practice in realist evaluation to serve as a baseline/briefing document for the panel.
3. Run three rounds (and more if needed) of the online Delphi panel to generate and refine items for a set of quality standards and reporting guidance.
4. In parallel with the Delphi panel:

    (a) provide ongoing advice and consultancy to up to 10 realist evaluations, including any funded by the National Institute for Health Research (NIHR), thereby capturing the 'real-world' problems and challenges of this methodology
    (b) host the RAMESES JISCMail list on realist research (www.jiscmail.ac.uk/RAMESES), capturing relevant discussions about theoretical, methodological and practical issues
    (c) feed problems and insights from 4a and 4b into the deliberations of the Delphi panel.

5. Write up the quality standards and guidance for reporting in an open access journal.
6. Collate examples of learning/training needs for researchers, postgraduate students and peer reviewers in relation to realist evaluation.
7. Develop, deliver and refine resources and training materials for realist evaluation. Deliver three 2-day 'realist evaluation' workshops and three 2-day 'training the trainers' workshops for a range of audiences [including interested NIHR Research Design Service (RDS) staff].
8. Develop, deliver and refine information and resources for patients and other lay participants in realist evaluation. In particular, draft template information sheets and consent forms that could be adapted for ethics and governance activity.
9. Disseminate training materials and other resources, for example via public-access websites.

Methods

In this project we used a range of methods to meet the objectives set out above. To fulfil objectives 1 and 2 we undertook a thematic review of the literature that was supplemented by our content expertise and by collating feedback from presentations and workshops. We synthesised our findings into briefing materials for realist evaluations. We recruited members to the Delphi panel, which had wide representation from researchers, students, policy-makers, theorists and research sponsors. We used the briefing materials to brief the Delphi panel so that they could help us in fulfilling objective 3. For the advice and consultancy in objective 4, we drew on not only our experience in developing and delivering education materials, but also relevant feedback from the Delphi panel, the RAMESES JISCMail e-mail list on realist research approaches, training workshops and the evaluations teams we had supported methodologically in the past. To help us refine our reporting standards (objective 5), we captured methodological and other challenges that arose within the realist evaluation projects we provided methodological support to. To produce the definitive reporting standards, quality standards and resources and training materials (objective 5), we synthesised expert input (from the Delphi panel), literature review and real-time problem analysis (e.g. feedback from the e-mail list, training sessions and workshops and presentations).

Throughout this project we did not set specific time points when we would refine the drafts of our project outputs. Instead, we iteratively and contemporaneously fed any data we captured into our draft reporting standards, quality standards and resources and training materials, making changes gradually. Only our Delphi panel ran within a specific time frame. The definitive guidance and standards were, therefore, the product of continuous refinements. To understand and develop information and resources for patients and other lay participants in realist evaluation (objective 8) we convened a group consisting of patients and the public. We addressed objective 9 through academic publications, online resources and delivery of presentations and workshops.

Results

Our literature review identified 152 realist evaluations, and when we had analysed 37 of these we had reached thematic saturation. Our analysis and discussion within the project team produced a summary of the published literature, and common questions and challenges in briefing materials for the Delphi panel. The Delphi panel comprised 35 members from 27 organisations across six countries and five disciplines. Within three rounds, the panels had reached a consensus on 20 key reporting standards, with an overall response rate of 76% and 80% for rounds 2 and 3, respectively. The RAMESES II reporting standards for realist evaluations have been published in an open-access journal and the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network (www.equator-network.org).

The quality standards and resources and training materials drew on the following sources of data: (1) personal expertise from researchers and trainers; (2) data from the Delphi panels; (3) feedback from participants at training sessions we ran; and (4) comments made on RAMESES JISCMail mailing list.

We developed eight quality criteria for realist evaluations with different versions for evaluators, researchers, peer reviewers and funders/commissioners of research. For our resources and training materials, we used the data we captured to identify the methodological topics that were highlighted by the majority of realist evaluators as most challenging. We developed training materials for 15 theoretical and methodological topics in realist evaluations. The quality standards and training materials are freely available online (www.ramesesproject.org).

We provided methodological support to 17 projects and presentations or workshops to help build research capacity in realist evaluations to 29 organisations, both nationally and internationally. This training included two 'training the trainers' workshops run in conjunction with the NIHR RDS East Midlands. Finally, we produced a generic patient information leaflet for lay participants in realist evaluations.

Conclusions

In conclusion, although realist evaluation holds much promise for developing theory and informing policy in some of the health and other sectors' most pressing questions, misunderstandings and misapplications of it is common. To try to address these problems, we have produced reporting and quality standards, and resources and training materials. In addition, we provided methodological support and advice to realist evaluation projects, ran training workshops for fellow realist evaluators and developed information and resources for patients and other lay participants in realist evaluation. However, for the quality of realist evaluations to improve, evaluators who wish to use realist evaluation will have to develop the necessary skills and use the materials we have developed.

We hope that our resources will be the start of an iterative journey of refinement and development of better resources for realist evaluations. Acknowledging that the science of evaluation should never be static, the RAMESES II project seeks not to produce the last word on these issues but to capture current expertise and establish an agreed state of the science on which future researchers will no doubt build. Much methodological development is needed in realist evaluation (e.g. work on appropriate quantitative methods, implications for research ethics, development of realist approaches in particular sectors and adaptation of existing evaluation tools for realist approaches). However, this can take place only if there is a sufficient pool of highly skilled realist evaluators. Capacity building through, for example, training and 'apprenticeships' of less experienced evaluators with more experienced ones is the next key step in realist evaluation.

Funding

Funding for this study was provided by the Health Services and Delivery Research programme of the National Institute for Health Research.

Chapter 1 Background

Many of the problems confronting policy- and decision-makers, evaluators and researchers today are complex. For example, much health service demand results from the effects of smoking, suboptimal diets (including obesity), excessive alcohol, inactivity or adverse family circumstances (e.g. partner violence), all of which, in turn, have multiple causes operating at both individual and societal level. Interventions or programmes designed to tackle such problems are themselves complex, often having multiple, interconnected components delivered individually or targeted at communities or populations. Their success depends both on individuals' responses and on the wider context in which people strive (or not) to live healthy lives. What works in one family, one organisation or one city may not work in another.

Similarly, the 'wicked problems' of contemporary health services research – how to improve quality and assure patient safety consistently across the service, how to meet rising need from a shrinking budget and how to realise the potential of information and communication technologies (which often promise more than they deliver) – require complex delivery programmes with multiple, interlocked components that engage with the particularities of context. What works in hospital A may not work in hospital B.

Designing and evaluating complex interventions is challenging. Randomised trials that compare 'intervention on' with 'intervention off', and their secondary research equivalent, meta-analyses of such trials, may produce statistically accurate statements (e.g. that the intervention works 'on average'), but may leave us none the wiser about where to target resources or how to maximise impact.

Realist evaluation seeks to address these problems. It is a form of theory-driven evaluation, based on realist philosophy,[1] that aims to advance understanding of why these complex interventions work, how, for whom, in what context and to what extent, as well as to explain the many situations in which a programme fails to achieve the anticipated benefit.

Realist evaluation assumes both that social systems and structures are 'real' (because they have real effects) and that human actors respond differently to interventions in different circumstances. To understand how an intervention might generate different outcomes in different circumstances, realism introduces the concept of mechanisms – which may be helpfully conceptualised as underlying changes in the reasoning of participants who are triggered in particular contexts.[2] For example, a school-based feeding programme may work by relieving hunger in young children in a low-income rural setting where famine has produced overt nutritional deficiencies, but for teenagers in a troubled inner-city community where many young people are disaffected, it may work chiefly by making pupils feel valued and nurtured.[3] What constitutes 'working' is also likely to be somewhat different in the two settings.

Realist evaluations have addressed numerous topics of central relevance in health services research, including what works and for whom when 'modernising' health services,[4] introducing breastfeeding support groups,[5] using communities of practice to drive change,[6] involving patients and the public in research,[7] how robotic surgery impacts on team-working and decision-making within the operating theatre[8] and fines for delays in discharge from hospitals.[9] They have also been used in fields as diverse as international development, education, crime prevention and climate change.

What is realist evaluation?

Realist evaluation was developed by Pawson and Tilley in the 1990s,[10] originally in the field of criminology, to address the question, 'what works for whom in what circumstances and how?' in criminal justice interventions. This early work highlighted the following points:

- Social programmes (closely akin to what health service researchers call complex interventions) are an attempt to address an existing social problem (i.e. to create some level of social change).

- Programmes 'work' by enabling participants to make different choices (although choice-making is always constrained by such things as participants' previous experiences, beliefs and attitudes, opportunities and access to resources).
- Making and sustaining different choices may require a change in a participant's reasoning (e.g. in their values, beliefs, attitudes or the logic they apply to a particular situation) and/or the resources (e.g. information, skills, material resources, support) they have available to them. Programmes provide opportunities and resources. The interaction between what the programme provides and the participant's 'reasoning' is what enables the programme to 'work' and is known as a 'mechanism'.
- Programmes work in different ways for different people (that is, the contexts within programmes can trigger different change mechanisms for different participants).
- The contexts in which programmes operate make a difference to the outcomes they achieve. Programme contexts include features such as social, economic and political structures, organisational context, programme participants, programme staffing, geographical and historical context, and so on. In realist terms, context does not simply denote spatial, geographical or institutional locations. Context refers, among other things, to the sets of 'social rules, norms values and interrelationships' that operate within these locations.[10]
- Some aspects of the context enable particular mechanisms to be triggered. Other aspects of the context may prevent particular mechanisms from being triggered. That is, there is always an interaction between context and mechanism, and that interaction is what creates the programme's impacts or outcomes: context + mechanism = outcome.
- Because programmes work differently in different contexts and through different change mechanisms, they cannot simply be replicated from one context to another and automatically achieve the same outcomes. Theory-based understandings about 'what works for whom, in what contexts, and how' are, however, transferable.
- Therefore, one of the tasks of evaluation is to learn more about: 'what works', in what respects and to what extent, including intended and unintended outcomes; 'for whom', that is, for which subgroups of participants; 'in which contexts'; and 'what mechanisms are triggered by what programmes in what contexts'.

A realist evaluation approach assumes that programmes are 'theories incarnate'. That is, whenever a programme is implemented, it rests on a theory about what 'might cause change', even though that theory may not be explicit. One of the tasks of a realist evaluation is, therefore, to make the theories underpinning a programme explicit, by developing clear hypotheses about how, and for whom, programmes might 'work'. The implementation of the programme, and the evaluation of it, then tests those hypotheses. This means collecting data, not just about programme impacts or the processes of programme implementation, but about the specific aspects of context that might impact on programme intended and unintended outcomes, and about the specific mechanisms that might be creating change.

Pawson and Tilley[10] also argue that a realist approach has particular implications for the methods required to evaluate a programme. For example, rather than comparing changes for participants who have undertaken a programme with a group of people who have not (as is done in randomised controlled or quasi-experimental designs), a realist evaluation compares context–mechanism–outcome configurations (CMOCs) within programmes. It may ask, for example, whether a programme works more or less well, and/or through different mechanisms, in different localities (and if so, how and why) or for different subgroups of the population. Furthermore, they argue that different stakeholders will have different information and understandings about how programmes are supposed to work and whether or not they in fact do so. Data collection processes (interviews, focus groups, questionnaires and so on) should be constructed to identify and collect the particular information that those stakeholder groups will have, and thereby to confirm, refute or refine theories about how and for whom the programme 'works'.

Realist evaluation is underpinned by a realist philosophy of science ('realism').[11] Philosophically speaking, realism can be thought of as sitting between positivism ('there is a real external world which we can come to know directly through experiment and observation') and constructivism ('given that all we can know has

been interpreted through human senses and the human brain, we cannot know for sure what the nature of reality is'). However, it is worth pointing out that this is not to suggest that 'constructivism' and 'positivism' represent opposite poles on the same continuum. Realism holds that there is a real social world but that our knowledge of it is amassed and interpreted (partially and/or imperfectly) via our senses and brains, and filtered through our language, culture and past experience. In other words, realism sees the human agent as operating in a wider social reality, encountering experiences, opportunities and resources, and interpreting and responding to the world within particular personal, social, historical and cultural frames. For this reason, different people respond differently to the same experiences, opportunities and resources. Hence, a programme (or, in the language of health services research, a complex intervention) aimed at improving health outcomes is likely to have different levels of success with participants in different contexts, and even in the same context at different times.

The need for standards and training materials in realist evaluation

The RAMESES JISCMail listserv [www.jiscmail.ac.uk/RAMESES (an e-mail list for discussing realist approaches)] postings suggest that enthusiasm for realist evaluation and belief in its potential for application in many fields have outstripped the development and application of robust quality standards in the field. Two important prior publications have systematically shown that many so-called 'realist evaluations' were not applying the concepts appropriately and were, as a result, producing potentially misleading findings and recommendations.[12,13]

Pawson and Manzano-Santaella, in their paper, 'A realist diagnostic workshop', used case examples of flawed realist evaluations to highlight three common errors in such studies.[13] First, while it is possible to show associations and correlations in data from many types of evaluation, the focus of a realist evaluation should be to explore and explain why such associations occur. Second, they explain what may constitute valid data for use in realist evaluation. Producing a realist explanation is likely to require a mix of data types to provide explanations and support for the relationships within and between CMOCs. Third, realist explanations require CMOCs to be produced. Pawson and Manzano-Santaella note that some realist evaluations have presented finely detailed lists of contexts, mechanisms and outcomes, but have failed to produce a coherent explanation of how these contexts, mechanisms and outcomes were linked and related, or not related, to each other. Pawson and Manzano-Santaella called for greater emphasis on elucidating programme theory (the theory about what a programme or intervention is expected to do and, in some cases, how it is expected to work) expressed as CMO configurations.

Marchal *et al.*[12] undertook a review of the realist evaluation literature in health systems research to quantify and analyse the field. They identified 18 realist evaluations and noted a range of challenges that arose for researchers. Absence of prior theoretical and methodological guidance appeared to have led to recurring problems in the realist evaluations they appraised. Marchal *et al.*[12] noted that '[t]he philosophical principles that underlie realist evaluation are variably interpreted and applied to different degrees'. Different researchers had conceptualised concepts used in realist evaluation such as 'middle-range theory', 'mechanism' and 'context' differently. This, they concluded, was often related to fundamental misunderstandings, and the rigour of the evaluation suffered as a result.

These two papers[12,13] showed that, although realist evaluation had been embraced by parts of the health research community, it had also proven a challenging task for some who were unfamiliar with the practical application of realism. Both sets of authors called for methodological guidance to allay misunderstandings about the purpose, underlying philosophical assumptions, analytic concepts and methods of realist evaluation.

Chapter 2 Methods

Objectives

The project had both strategic and operational objectives, and, because it was funded through the health sector, the objectives were framed in relation to health. However, representatives from beyond the health sector were involved to ensure that the products were relevant to any realist evaluation.

Strategic objectives

(a) To develop quality standards, reporting guidance and resources and training materials for realist evaluation.
(b) To build capacity in health services research for supporting and assessing realist approaches to research.
(c) Acknowledging the unique potential of realist research to address the patient's agenda ('what will work for us in our circumstances?'), to produce resources and training materials for lay participants, and those seeking to involve them, in research.

Operational objectives

1. Recruit an interdisciplinary Delphi panel of, for example, researchers, support staff, policy-makers, patient advocates and practitioners with various types of experience relevant to realist evaluation.
2. Summarise the current literature and expert opinion on best practice in realist evaluation, to serve as a baseline/briefing document for the panel.
3. Run three rounds (and more if needed) of the online Delphi panel to generate and refine items for a set of quality standards and reporting guidance.
4. In parallel with the Delphi panel:

    (a) provide ongoing advice and consultancy to up to 10 realist evaluations, including any funded by the National Institute for Health Research (NIHR), thereby capturing the 'real-world' problems and challenges of this methodology
    (b) host the RAMESES JISCMail list on realist research, capturing relevant discussions about theoretical, methodological and practical issues
    (c) feed problems and insights from 4a and 4b into the deliberations of the Delphi panel.

5. Write up the quality standards and guidance for reporting in an open-access journal.
6. Collate examples of learning/training needs for researchers, postgraduate students and peer reviewers in relation to realist evaluation.
7. Develop, deliver and refine resources and training materials for realist evaluation. Deliver three 2-day 'realist evaluation' workshops and three 2-day 'training the trainers' workshops for a range of audiences [including interested NIHR Research Design Service (RDS) staff].
8. Develop, deliver and refine information and resources for patients and other lay participants in realist evaluation. In particular, draft template information sheets and consent forms that could be adapted for ethics and governance activity.
9. Disseminate training materials and other resources, for example via public-access websites.

Overview of methods

We first provide a brief overview of the range of methods we used to meet the objectives set out above and of how they related to each other. The methods we used in this project closely resemble those we

used in another project (the RAMESES project), which developed methodological guidance, reporting standards and training materials for realist and meta-narrative reviews.[14] We have previously published a protocol paper that outlined the methods we intended to use in this project.[15] The following methods sections outline, in more detail, specific aspects of the methods used.

To fulfil operational objectives 1 and 2, we undertook a thematic review of the literature. Findings were supplemented by our content expertise and with feedback collated from presentations and workshops for researchers using or intending to use realist evaluation. We synthesised our findings into briefing materials on realist evaluation for the Delphi panel. We recruited members to the Delphi panel, which had wide representation from researchers, students, policy-makers, evaluators, theorists and research sponsors. We used the briefing materials to inform the Delphi panel in preparation for the task, so they could contribute to developing standards (objective 3). For the advice and consultancy to realist evaluations (objective 4a), we drew on our experience in conducting realist evaluations and developing and delivering education materials, but also on relevant feedback from the Delphi panel, an e-mail list on realist research approaches (www.jiscmail.ac.uk/RAMESES) and the evaluations teams we had supported in the past. To help us refine our reporting standards (objective 5), we captured methodological and other challenges that arose within the realist evaluation projects to which we provided methodological support. All of these sources fed into the reporting standards, quality standards and resources and training materials (objective 7). We did not set specific time points when we would refine the drafts of our project outputs. Instead, we iteratively and contemporaneously fed the data we captured into our draft reporting standards, quality standards and resources and training materials, making changes gradually. Only our Delphi panel ran within a specific time frame. The final guidance and standards were, therefore, the product of continuous refinements. To understand and develop information and resources for patients and other lay participants in realist evaluation (objective 8), we convened a group consisting of patients and the public. We addressed objective 9 through academic publications, online resources and delivery of presentations and workshops. The project was overseen by a Project Advisory Group, which comprised three independent members (see *Acknowledgements*). This group met with the project team on three occasions (May 2015, November 2015 and May 2016) and provided advice to the project team. *Figure 1* provides a pictorial overview of how the different methods we used fed into each other.

Details of literature search methods

With input from an expert librarian, we identified reviews, scholarly commentaries, models of good practice and examples of (alleged) misapplication of realist evaluation. To identify the relevant documents we refined and developed the search used by Marchal *et al.*[12] for a previous review on a similar topic, and also applied contemporary search methods designed to identify 'richness' when exploring complex interventions.[16,17]

A search was conducted on 3 March 2015 across 10 databases. Free-text terms were selected to describe realist methods and thesaurus terms were used where available (see *Appendix 1*). The following databases were searched:

- Cumulative Index to Nursing and Allied Health Literature (CINAHL; via EBSCO*host*)
- The Cochrane Library (Wiley Online Library)
- Dissertations & Theses database (ProQuest)
- EMBASE (via OvidSP)
- Education Resources Information Center (ERIC; via EBSCO*host*)
- Global Health (via OvidSP)
- MEDLINE and MEDLINE In-Process & Other Non-Indexed Citations (via OvidSP)
- PsycINFO (via OvidSP)
- Scopus, Science Citation Index (SCI), Social Science Citation Index (SSCI) & Conference Proceedings, Citation Index – Science (CPCI-S)
- Web of Science Core Collection (Thomson Reuters Corporation, New York, NY, USA).

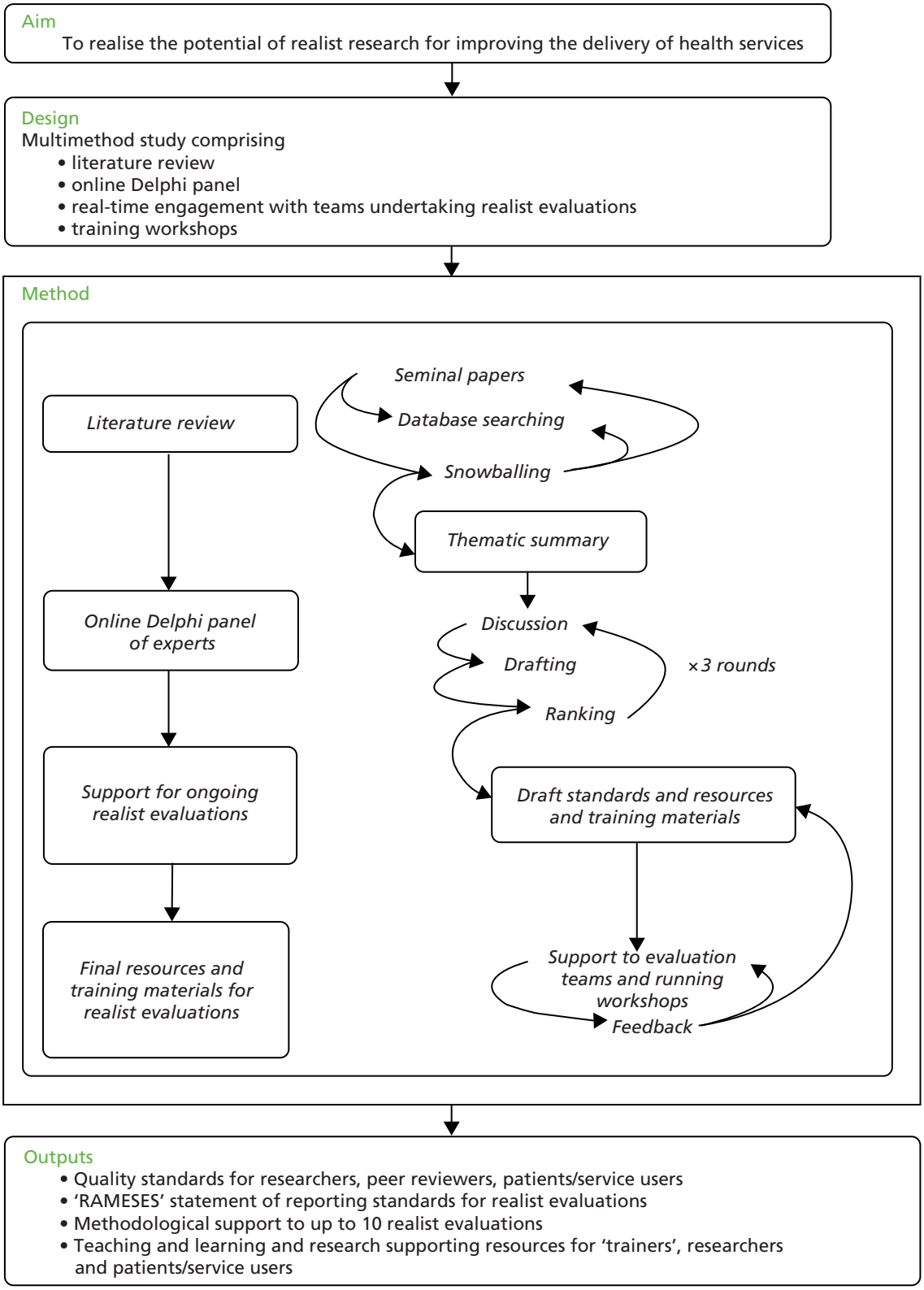


FIGURE 1 Overview of study processes.

A forward citation search was conducted via the Web of Science Core Collection for the following key text: Pawson R, Tilley N. *Realistic Evaluation*. London: Sage; 1997.[10]

No language or study design filters were applied. We included any document that referred to or claimed to be a realist evaluation that used the approach as set out by Pawson and Tilley in their key publication, *Realistic Evaluation*.[10] Documents were excluded if they were not realist evaluations, published prior to the year 2000, book reviews, letters and comment. We set the cut-off point at 2000, as we assumed that evaluations based on Pawson and Tilley's work would begin appearing in the literature from this point onwards. All citation screening was undertaken by Geoff Wong. The whole searching process, from start to the retrieval of all full-text documents, took approximately 1 month.

We decided that, because of the narrow purpose of our review and the number of relevant citations retrieved, we would stop analysing data when we had reached thematic saturation. As a strategy to manage the potential number of realist evaluations, we decided to start our analysis and synthesis from the most recent (i.e. from 2015) realist evaluations and work 'backwards'. The decision on when thematic saturation had been reached was made in discussion with the whole project team. For both practical reasons (e.g. resource constraints) and academic ones (no new data), we stopped including new papers when there was agreement that saturation of themes had been reached. Thematic saturation was reached once the group agreed by consensus that the new realist evaluations identified contained no new themes or only subthemes that related to the three questions listed below in bullet points.

The thematic analysis was led by Geoff Wong, who undertook all stages of the review and shared findings with the rest of the project team so that discussion, debate and refinement of interpretations of the data could take place. Findings were shared by e-mail and, when necessary, face-to-face meetings were conducted to discuss interpretations of the data.

In undertaking our thematic analysis, we familiarised ourselves with the included evaluations to identify patterns in the data. Aware that the purpose of the review was to produce briefing documents for the Delphi panel, we considered the following questions:

- What is considered by experts in realist evaluation to be current best practice (and what is the range and diversity of such practice)?
- What do experts in realist evaluation, and other researchers who have undertaken a realist evaluation, believe counts as high quality and necessary to report?
- What issues do researchers struggle with (based on thematic analysis of postings on the RAMESES JISCMail list archive as well as the published literature)?

In the panels, we wanted to achieve a consensus on quality and reporting standards, and thus what we needed from our review of the literature were data to inform us on what might constitute quality in executing and reporting realist evaluations. We accepted that we might need to refine, discard or add additional questions and topic areas in order to better capture our analysis and understanding of the literature as these emerged from our reading of the evaluations.

Data were extracted to a Microsoft Excel® (Microsoft Corporation, Redmond, WA, USA) spreadsheet that we iteratively refined to capture the data needed to produce our briefing materials. This review was undertaken in a short time frame. The time taken from obtaining full-text documents to producing the final draft for circulation of the briefing documents was approximately 12 weeks. The output of this phase was a provisional summary that addressed the questions above and highlighted, for each question, the key areas of knowledge, ignorance, ambiguity and uncertainty. This was distributed to the Delphi panel (as our briefing document) as the starting point for its work.

Our purpose in identifying published reviews was not to complete a census of realist evaluations. We make no claims that the review we undertook was exhaustive; thus, we never intended that it should be

published as a stand-alone piece of research. In other words, the purpose of our review was not to produce definitive summaries in response to the themes above but to prepare a baseline set of briefing materials for the Delphi panel, and to deliberate on and add to them in the next step. As such, the review we undertook would be best considered as being a rapid, accelerated or truncated thematic review. Such an approach will predictably produce limitations, and these are discussed in *Chapter 4, Limitations*.

Details of online Delphi process

We recruited Delphi panel members purposefully, to ensure that we had representation from evaluators, researchers, funders, journal editors and experts in realist evaluation. Individuals were recruited through relevant organisations and targeted e-mails, and also through personal contacts and recommendations. Those interested in participating were provided with an outline of the study, and individuals who indicated the greatest commitment and potential to balance the sample were selected.

The Delphi panel was run online using SurveyMonkey (SurveyMonkey, Palo Alto, CA, USA). Participants in round 1 were provided with the briefing materials we developed from the literature review and were invited to suggest what might be included in the reporting standards. Responses were analysed and fed into the design of questionnaire items for round 2.

In round 2 of the Delphi Panel, participants were asked to rank each potential item twice on a 7-point Likert scale (1 = strongly disagree to 7 = strongly agree), once for relevance (i.e. 'Should an item on this theme/topic be included at all in the guidance?') and once for validity (i.e. 'To what extent do you agree with this item as currently worded?'). Those who agreed that an item was relevant, but disagreed on its wording, were invited to suggest changes to the wording via a free-text comments box. In this second round, participants were again invited to suggest additional topic areas and items. We did not prespecify stop-points for establishing when consensus has been achieved. This was because we wanted to have the flexibility to return to the Delphi panel items that we judged might need further input. Although we accept that this may have enabled us to preferentially return some items and not others, we guarded against this by sending all Delphi panel members an end-of-round report detailing all the findings, changes made to the text and items to be returned to the next round. Panel members were invited to contact us should they have any concerns with the items that were not returned for re-rating, such as believing that the item should be returned to the panel, or disagreeing with wording changes.

Participants' responses were collated and the numerical rankings were entered onto an Excel spreadsheet. The response rate, average, mode, median and interquartile range (IQR) for each item was calculated. Items that scored low on relevance were omitted from subsequent rounds. We invited further online discussion on items that scored high on relevance but low on validity (indicating that a rephrased version of the item was needed) and on those for which there was wide disagreement about relevance or validity. The panel members' free-text comments were also collated and analysed thematically.

Following analysis and discussion within the project team, we drew up a second list of statements that were circulated for ranking (round 3). Round 3 contained items for which consensus had not yet been reached. For items on which consensus had been reached, we did not return these to rounds 3, 4 or beyond for panel members to re-rate, even if we had made changes to the wording. This was because, when we undertook the RAMESES project, we had received informal feedback from the Delphi panel members indicating that round 2 of the online Delphi process had been very time-consuming. We were advised that to retain a high response rate for subsequent rounds, we should minimise the time commitment we asked of panel members. We planned that the process of collating responses, further e-mail discussion and re-ranking would be repeated until a maximum consensus was reached (rounds 4, 5, and so on). In practice, very few Delphi panels, online or face to face, go beyond three rounds because participants tend to 'agree to differ' rather than move towards further consensus. We used e-mail reminders to optimise our response rate from Delphi panel members. We considered consensus to be achieved when the median score was 6 or above.

We planned to report residual non-consensus as such and to report the nature of the dissent described (if any). Making such dissent explicit tends to expose inherent ambiguities, which may be philosophical or practical, and acknowledges that not everything can be resolved; such findings may be more use to those who use realist evaluation than a firm statement that implies that all tensions have been fixed. We used the findings from the Delphi panel to develop the reporting standards and methodological quality standards for realist evaluations.

Developing quality standards

The quality standards were designed to support professional development, assist evaluators to assess the quality of various aspects of the evaluation process and to assist reviewers with meta-evaluation (i.e. assessing the quality of evaluations).

To develop the quality standards, we drew on the following sources of data:

- free-text comments from participants and findings from the Delphi panels
- personal expertise as evaluators, researchers, peer reviewers and trainers in the field
- feedback from participants at workshops and training sessions run by members of the project team
- comments made on RAMESES JISCMail.

The data from the sources above were collated contemporaneously and discussed within the project team. Iterative cycles of discussion and revisions for content and clarity of the drafts were needed to develop the standards. *Box 1* provides an illustration of how we drew on the data sources to produce the quality standards.

Developing, delivering and refining resources and training materials for realist evaluation

An important part of our project was to produce publicly accessible resources to support training in realist evaluations. We anticipated that these resources will need to be adapted, and perhaps supplemented, for different groups of learners, and interactive learning activities added. We developed, and iteratively refined, draft learning objectives, example course materials and teaching and learning support methods. We drew on a range of sources to inform the content and format of our training materials as well as our experience as trainers and consultants on realist evaluations.

We sought out examples of the kinds of requests that are often made by evaluators for support on realist evaluation, for example using the rich archive of postings on the RAMESES JISCMail listserv from both novice and highly experienced practitioners, going back 3 years. We also proactively asked the list members for additional examples, and used our empirical data from the Delphi panel and our literature review to identify relevant examples. Finally, we sought input from UK RDS staff interested in realist evaluation to describe the kind of problems people bring to them, and where they feel that further guidance, support and resources are needed.

We used a thematic approach to classify examples into a list of problems and issues, each with a corresponding training need(s) and resources to address them. These were developed iteratively in regular discussions and meetings of the research team. Our goal was to develop a coherent and comprehensive curriculum for training realist researchers and for ‘training the trainers’.

BOX 1 Illustration of the type of data we drew on to identify the need for, and develop, quality standards

Quality standard: programme theories

Identification of need

As evaluators, researchers and trainers in realist evaluation, we had noted that there was some confusion among researchers about the nature, need and role of realist programme theory (or theories) in realist evaluations. To develop the briefing materials and initial drafts of the reporting standards for realist evaluations, we searched for and analysed a number of published evaluations and noted that our impressions were well founded.

When providing methodological support for a realist evaluation, the importance of programme theory emerged again. One of the project team commented, 'I felt the development of the initial "programme theory" pulled things together . . .' In our Delphi process, we encouraged participants to provide free-text comments. These closely reflected the comments we received about the importance of programme theory.

Development of the quality criteria

We drew on our content expertise of the topic area and published methodological literature to develop the quality criteria. In addition, we found that some of our Delphi panel participants provided us with clear indications that supported the criteria we set. For example, we suggested that realist evaluations should develop a programme theory and one that did not was 'inadequate'. Delphi panel participants' free-text comments echoed our suggestion:

> *Really important . . .*
>
> *Initial programme theories will be clearly stated . . .*
>
> *Many people's efforts at realist evaluation fall at the programme theory stage . . .*

We were also able to draw on the discussions that took place on JISCMail to support some of our criteria. For example, under 'adequate', we suggested that: 'initial tentative programme theory (or theories) were identified and (as far as possible) described in realist terms (that is, in terms of the causal relationship between contexts, mechanisms and outcomes). These were refined as the evaluation progressed'.

As illustration, a comment from JISCMail that we drew upon to support this criterion was:

> *It's good to read that you are planning to develop a programme theory. It may be that even before you start data collection that you may wish to develop an initial 'best guess' programme theory of the . . . intervention. Do not worry that it may be a best guess and has no CMOCs (i.e. is not particularly realist in nature) – it is a starting point. As the evaluation progresses your job is to gradually (iteratively) 'convert' it into a more detailed realist programme theory that has data to support any inferences you have made.*

Support and consultancy to realist evaluations

The support we offered to fellow evaluators and researchers using realist evaluations consisted of two overlapping and complementary levels:

1. Online discussion and support via JISCMail for evaluators and researchers, at any level, interested in or undertaking a realist evaluation. When questions or issues were raised, either one of the project team or another list member would reply. Where necessary, summaries were made of discussions and clarification was provided by members of the project team.
2. Direct requests for support and training. During the course of the study, members of the project team were frequently approached to provide methodological support to realist evaluation projects. The exact content, nature and duration of the support provided was discussed between the relevant team members to ensure that what was provided met the needs of those who requested the support.

Realist evaluation and 'training the trainers' workshops

Throughout the 24 months of the project, members of the project team offered training workshops to other evaluators, researchers and patient organisations on an as-requested basis. When asked to provide a workshop, the logistics and content of each workshop were discussed between the relevant project team member and the hosts.

For the 'training the trainers' workshops, we engaged with the NIHR's RDS. We did this by e-mailing each regional service and also asking for expressions of interest via e-mail lists and personal contacts.

Develop, deliver and refine information and resources for patients and other lay participants in realist evaluation

To develop these resources, we convened a panel of lay participants with the help of the Patient and Public Involvement Co-ordinator from the Nuffield Department of Primary Care Health Sciences at the University of Oxford. We sought to invite lay participants who had been involved in research studies and came from a range of backgrounds and ages. During the panel, we sought to understand what lay participants might wish to know if they were to participate in a realist evaluation and provided examples of the potential materials for their consideration.

Chapter 3 Results

We produced four outputs related to realist evaluations for this project, namely:

1. reporting standards
2. methodological quality standards
3. resources and training materials (for researchers, evaluators and lay participants)
4. capacity building.

This chapter provides details of the results we obtained from the methods and approaches we used, and how they contributed to the content of our outputs.

Literature search

We searched 10 electronic databases from inception (where applicable) to March 2015 and, along with citation tracking, retrieved 4426 documents after removal of duplications. A total of 1498 duplicates were removed, along with a further 737 papers that did not meet our inclusion criteria. A total of 2191 papers were screened by title and abstract for inclusion with 1503 excluded at this stage. *Figure 2* shows the disposition of the documents and *Table 1* the number of citations returned for each database searched.

One of the project team (GWo) screened the abstracts and titles and included documents that claimed to be realist evaluations. In total, 152 documents were judged to be realist evaluations. Because of the narrow focus of our review of the literature on realist evaluations, as discussed in *Chapter 2*, *Details of literature search methods*, we worked 'backward' from 2015 to earlier years and sought to stop analysis at the point of thematic saturation. We achieved thematic saturation after analysis of 37 out of the 152 realist evaluations. Out of these realist evaluations, 32 (from years 2015 and 2014 inclusively) evaluated health-related topics, and five (from years 2015 to 2012 inclusively) evaluated non-health-related topics. We made this distinction to ensure that we analysed realist evaluations that covered a range of topic areas, as the approach is used in a broad range of topic areas beyond health research. Hence, *Table 2* shows only the characteristics of the documents we analysed (evaluation title, type of document, year submitted for publication and topic area) and drew on to produce our briefing document for the Delphi panel.

Because many evaluation reports are not published and our search strategy focused on published materials, the great majority of documents we analysed were journal articles about evaluations rather than complete evaluation reports. We acknowledge that full evaluation reports may have provided greater detail. However, because journal articles usually require a description of both methods and findings, our focus was methodological and the literature review served only to identify issues to refer to the Delphi panel; therefore, we remain confident that the sample was adequate for the task.

We conducted a thematic analysis guided, initially, by the three questions set out above (see *Chapter 2, Details of literature search methods*) to produce the briefing documents for the realist evaluation Delphi panel (see *Appendix 2*). All the data we extracted were either entered into an Excel spreadsheet or written up directly into a draft of our briefing document. Of the three questions set out above, two refer to what experts in realist evaluation and researchers who have undertaken a realist evaluation consider to be best practice and high quality. Much of this information was contained in the documents listed in *Table 2*, but we also had to supplement our understanding by drawing on more methodological documents.[1,10,12,13]

Our first question [what is considered by experts to be current best practice (and what is the range and diversity of such practice)?] related to perceptions of methodological rigour in the execution of realist evaluations. Addressing this question required the most immersion and analysis. With this question,

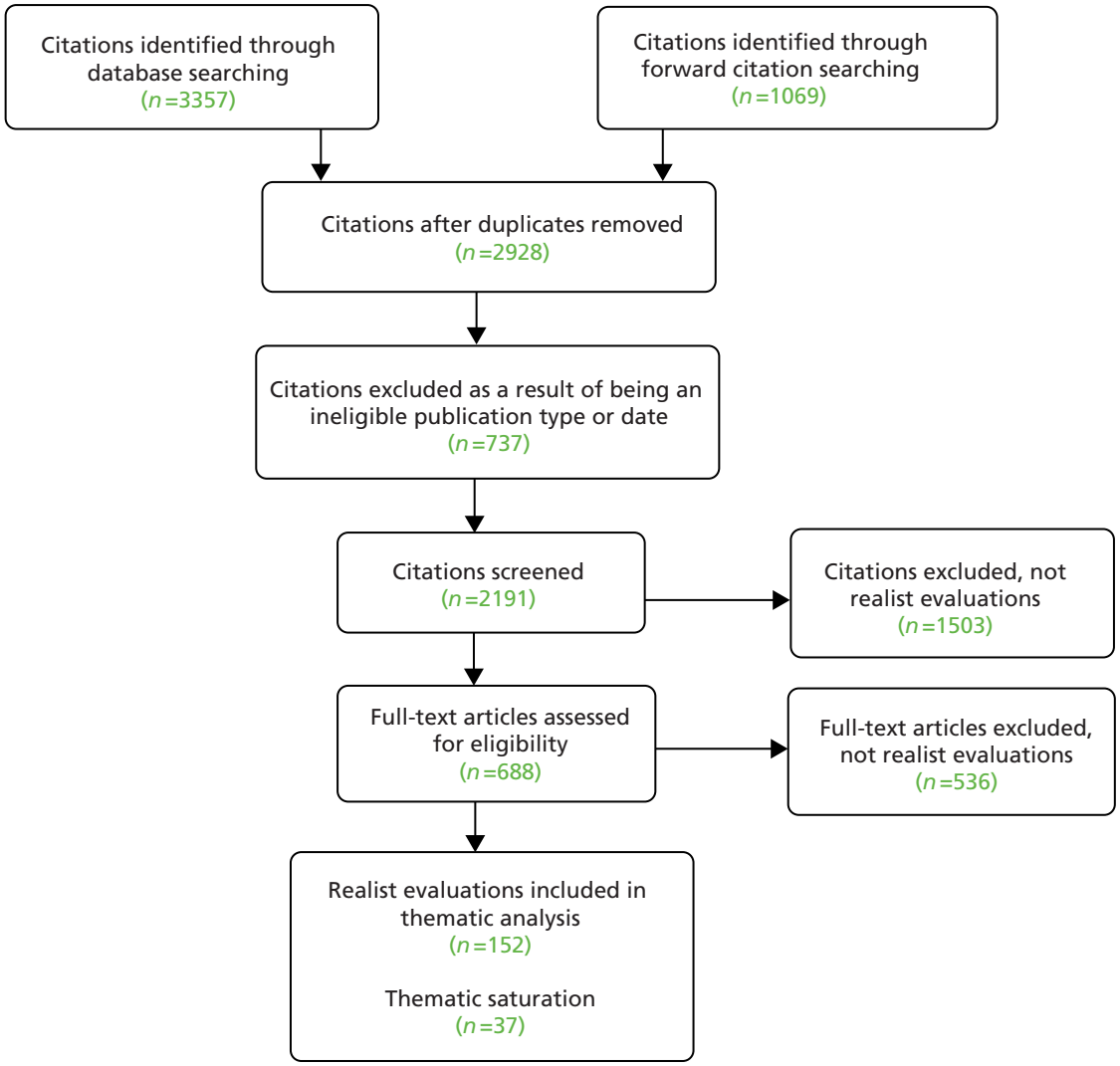


FIGURE 2 Flow diagram outlining the disposition of documents.

TABLE 1 Citations returned for databases searched

Database	Number of citations returned
CINAHL	215
The Cochrane Library	26
Dissertations & Theses	147
EMBASE	484
ERIC	209
Global Health	94
MEDLINE and MEDLINE In-Process & Other Non-Indexed Citations	455
PsycINFO	533
Scopus, SCI, SSCI and CPCI-S	854
Web of Science Core Collection	340
Citation tracking	
Pawson R, Tilley N. *Realistic Evaluation*. London: Sage; 1997[10]	1069

TABLE 2 Characteristics of realist evaluation documents used to inform Delphi process materials (listed by year submitted for publication)

Study title (reference and reference number)	Year submitted	Topic area
Health-related realist evaluations		
Grades in formative workplace-based assessment: a study of what works for whom and why (Lefroy *et al.*[18])	2015	Education – medical (work-based assessment)
What works in 'real life' to facilitate home deaths and fewer hospital admissions for those at end of life?: results from a realist evaluation of new palliative care services in two English counties (Wye *et al.*[19])	2015	Palliative care (home death and hospital admissions)
Faculty development for educators: a realist evaluation (Sorinola *et al.*[20])	2014	Education – medical (faculty development)
Reducing emergency bed-days for older people? Network governance lessons from the 'Improving the Future for Older People' programme (Sheaff *et al.*[21])	2014	Emergency bed-days for older people
Using interactive workshops to prompt knowledge exchange: a realist evaluation of a knowledge to action initiative (Rushmer *et al.*[22])	2014	Interactive workshops for knowledge exchange
Can complex health interventions be evaluated using routine clinical and administrative data? – a realist evaluation approach (Riippa *et al.*[23])	2014	Use of routinely collected data for evaluating complex interventions
Introducing Malaria Rapid Diagnostic Tests (MRDTs) at registered retail pharmacies in Ghana: practitioners' perspective (Rauf *et al.*[24])	2014	Implementation of malaria rapid diagnostic tests in retail pharmacies
Advancing the application of systems thinking in health: a realist evaluation of a capacity building programme for district managers in Tumkur, India (Prashanth *et al.*[25])	2014	Capacity building programme for district health managers
Stroke patients' utilisation of extrinsic feedback from computer-based technology in the home: a multiple case study realistic evaluation (Parker *et al.*[26])	2014	Stroke rehabilitation using computer-based technology
Educational system factors that engage resident physicians in an integrated quality improvement curriculum at a VA hospital: a realist evaluation (Ogrinc *et al.*[27])	2014	Quality improvement in resident physician training
Realistic nurse-led policy implementation, optimization and evaluation: novel methodological exemplar (Noyes *et al.*[28])	2014	Policy implementation
Putting context into organizational intervention design: using tailored questionnaires to measure initiatives for worker well-being (Nielsen *et al.*[29])	2014	Work well-being
Mechanisms that support the assessment of interpersonal skills: a realistic evaluation of the interpersonal skills profile in pre-registration nursing students (Meier *et al.*[30])	2014	Interpersonal skills assessment
Factors affecting the successful implementation and sustainability of the Liverpool Care Pathway for dying patients: a realist evaluation (McConnell *et al.*[31])	2014	Palliative care – Liverpool Care Pathway
Towards a programme theory for fidelity in the evaluation of complex interventions (Masterson-Algar *et al.*[32])	2014	Implementation fidelity – complex rehabilitation intervention for patients with stroke
Action learning sets in a nursing and midwifery practice learning context: a realistic evaluation (Machin and Pearson[33])	2014	Education – action learning sets in nursing
Advancing the application of systems thinking in health: realist evaluation of the Leadership Development Programme for district manager decision-making in Ghana (Kwamie *et al.*[34])	2014	Leadership development programme
Adolescents developing life skills for managing type 1 diabetes: a qualitative, realistic evaluation of a guided self-determination-youth intervention (Husted *et al.*[35])	2014	Chronic disease management – use of guided self-determination in diabetes

continued

TABLE 2 Characteristics of realist evaluation documents used to inform Delphi process materials (listed by year submitted for publication) (*continued*)

Study title (reference and reference number)	Year submitted	Topic area
The management of long-term sickness absence in large public sector healthcare organisations: a realist evaluation using mixed methods (Higgins *et al.*[36])	2014	Sickness absence – long-term sickness absence in health-care workers
General practitioners' management of the long-term sick role (Higgins *et al.*[37])	2014	Sickness absence – GPs' management long-term sickness absence
More than a checklist: a realist evaluation of supervision of mid-level health workers in rural Guatemala (Hernández *et al.*[38])	2014	Supervision of mid-level health workers
Dialysis modality decision-making for older adults with chronic kidney disease (Harwood and Clark[39])	2014	Treatment decision-making – kidney dialysis
Housing, health and master planning: rules of engagement (Harris *et al.*[40])	2014	Housing regeneration
Public involvement in research: assessing impact through a realist evaluation (Evans *et al.*[41])	2014	Public involvement in research
Academic practice–policy partnerships for health promotion research: experiences from three research programs (Eriksson *et al.*[42])	2014	Health promotion – collaboration between academics, practitioners and policymakers
Schools' capacity to absorb a Healthy School approach into their operations: insights from a realist evaluation (Deschesnes *et al.*[43])	2014	Health in schools
A realist evaluation of a community-based addiction program for urban aboriginal people (Davey *et al.*[44])	2014	Substance use – First Nations, Inuit and Métis populations
Community resistance to a peer education programme in Zimbabwe (Campbell *et al.*[45])	2014	Health education – peer education of HIV
The transformative power of youth grants: sparks and ripples of change affecting marginalised youth and their communities (Blanchet-Cohen and Cook[46])	2014	Youth empowerment
The SMART personalised self-management system for congestive heart failure: results of a realist evaluation (Bartlett *et al.*[47])	2014	Chronic disease management – use of technology for self-management of health failure
Levels of reflective thinking and patient safety: an investigation of the mechanisms that impact on student learning in a single cohort over a 5 year curriculum (Ambrose and Ker[48])	2014	Education – teaching patient safety to medical students
People and teams matter in organizational change: professionals' and managers' experiences of changing governance and incentives in primary care (Allan *et al.*[49])	2014	Health services management – organisational change
Non-health-related realist evaluations		
Into the void: a realist evaluation of the eGovernment for You (EGOV4U) project (Horrocks and Budd[50])	2015	E-services designed to tackle social exclusion and disadvantage
Evaluating Criminal Justice Interventions in the Field of Domestic Violence – A Realist Approach (Taylor[51])	2014	Criminal justice – domestic violence interventions
How to use programme theory to evaluate the effectiveness of schemes designed to improve the work environment in small businesses (Olsen *et al.*[52])	2012	Work environment in small businesses
Improving outcomes for a juvenile justice model court: a realist evaluation (Kazi *et al.*[53])	2012	Criminal justice – juvenile justice model court
A model for design of tailored working environment intervention programmes for small enterprises (Hasle *et al.*[54])	2012	Work environment in small enterprises

GP, general practitioner.

we wanted to understand expert opinions about best practice to produce a high-quality realist evaluation. As a project team, we had our own ideas, but wanted to explore whether or not these were reflected in the included evaluations. We first had to decide whether or not we could agree among ourselves on which of the evaluations we analysed were of high, mixed or low quality. To do this, each evaluation was read in detail (GWo) and selected characteristics were extracted into an Excel spreadsheet. The headings on this spreadsheet were study name, type of document, year submitted, country, topic area, purpose of evaluation, understand realism?, methodological comments, lessons for methods, methods for reporting and challenges reported by reviewers' notes.

Once completed, the spreadsheet and the full-text documents were circulated to the rest of the project team. Through e-mail discussion and debate, a consensus was achieved on which studies were deemed high, mixed or low quality. The next step in the process was to re-read each of the included evaluations to determine which evaluation practices and processes were necessary to lead to a high-quality evaluation. Later on in the project, to develop reporting standards for realist evaluations, we used these findings to inform what needed to be reported to ensure that sufficient information was available to the reader, so that they were able to make judgments about methodological rigour. This addressed our second question (what do experts and other researchers believe count as high quality and necessary to report?). Again, this was led by Geoff Wong, and each issue that needed addressing was added to a draft of the briefing documents. To further strengthen the inferences we made on issues that needed to be addressed and, hence, included in our briefing materials, we looked back through the archives of the RAMESES JISCMail e-mail listserv to identify if the issues we had included had also been raised by other researchers. We also drew on the methodological issues raised in methods papers on realist evaluations in a similar way.[12,13]

The drafts of briefing materials were circulated to the project team and a consensus was achieved through discussion and debate. The briefing materials were the result of four rounds of revisions.

The contents of our briefing materials were as follows:

- terminology
- philosophical basis of realist evaluation
- classification
- title
- rationale for using realist evaluation
- methods
- data collection methods
- programme theory
- findings
- conclusion
- recommendations.

The complete briefing document circulated to the Delphi panels for realist reviews and meta-narrative reviews can be found in *Appendix 2*.

Delphi panel

We ran the Delphi panels between May 2015 and January 2016. We recruited 35 panel members from 27 organisations across six countries. The panel members comprised evaluators of health services (23), public policy (nine), nursing (six), criminal justice (six), international development (two), contract evaluators (three), policy- and decision-makers (two), funders of evaluations (two) and publishing (two) (note that some individuals had more than one role).

We started round 1 in June 2015 and circulated the briefing materials document to the panel. We sent two chasing e-mails to all panel members, and within 8 weeks all panel members who indicated that

they wanted to provide comments had done so. In round 1 of the Delphi panel, 33 members provided suggestions for items that should be included in the reporting standards and/or comments on the nature of the standards themselves. We used the suggestions from the panel members and the briefing document as the basis of the online survey for round 2.

Round 2 started at the end of September 2015 and ran until early November 2015. Panel members were invited to complete our online survey and asked to rate each potential item for relevance and validity. A copy of this survey can be found in *Appendix 3*. Where needed, up to three reminder e-mails were sent to the panel members. For round 2, the panel was presented with 22 items to rank. The overall response rate across all items for this round was 76%. Once the panel had completed its survey, we analysed their ratings for relevance and validity. Full details of the round 2 results can be found in *Table 3*. We also produced a post-round briefing document from round 2, which detailed for each item:

- the response rate
- mode
- median
- IQR
- the action we took for each item based on the panel's ratings
- an anonymised list of all the free-text comments made.

Based on the rankings and free-text comments, our analysis indicated that two items needed to be merged and one item removed. Minor revisions were made to the text of the other items based on the rankings and free-text comments. After discussion within the project team, we judged that only one item (the newly created merged item) needed to be returned to round 3 of the Delphi panel. Prior to the start of round 3, the post-round briefing document from round 2 was circulated to panel members. We did not receive any communication indicating that the panel members disagreed with the actions we undertook in response to their ratings and free-text comments from round 2.

For round 3, we asked the panel to consider again only the single item for which a consensus had not been reached. We produced an online survey for round 3 and, again, asked them to rate the item for relevance and validity. To keep the panel updated, we provided it with our post-round briefing document from round 2 (available on request from authors). Round 3 ran from late November 2015 to early January 2016. A copy of this survey can be found in *Appendix 4*. Two reminder e-mails were sent to the panel members. Once the panel had completed its survey, we analysed its ratings for relevance and validity (*Table 4*). The response rate for the single item included in round 3 was 80%. We produced a post-round briefing document from round 3 and circulated this to all our panel members for the sake of completeness (available on request from authors). We did not receive any communication indicating that the panel members disagreed with the actions we undertook in response to their ratings and free-text comments from round 3. Overall, consensus was reached within three rounds on both the content and wording of a 20-item reporting standard.

Using the data we gathered from the three rounds of the Delphi panel, we produced a final set of items to be included in the reporting for realist evaluations. These were published in June 2016 in *BMC Medicine*, an open-access journal.[55] Within this publication, we have provided an 'example' for each standard; that is, an example of good practice drawn from published evaluations. Our reporting standards have also been accepted and listed on the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network, a resource centre for good reporting of health research studies (www.equator-network.org).

Developing quality standards

We developed quality standards for two user groups, which are set out using rubrics:

1. evaluators and peer reviewers of realist evaluations
2. funders or commissioners of realist evaluations.

TABLE 3 Summary of results for round 2 of Delphi panel

Item	Relevance				Validity			
	Response rate (%)	Mode	Median	IQR	Response rate (%)	Mode	Median	IQR
Title	28/35 (80)	7	6.5	2.25	28/35 (80)	6	6	2
Summary or abstract	28/35 (80)	7	6	1	28/35 (80)	6	5.5	3
Rationale for evaluation	28/35 (80)	7	6	1	28/35 (80)	6	5	2.25
Programme theory	27/35 (77)	7	7	0	27/35 (77)	7	7	2
Evaluation questions, objectives and focus	27/35 (77)	7	7	1	27/35 (77)	7	6	3
Ethics	27/35 (77)	7	7	1	27/35 (77)	7	7	1
Rationale for using realist evaluation	27/35 (77)	7	7	1	27/35 (77)	7	6	1.5
Protocol or evaluation design	27/35 (77)	7	7	1	27/35 (77)	7	6	2.5
Setting(s) of the evaluation	27/35 (77)	7	7	1	27/35 (77)	6	6	2
Nature of the programme being evaluated	27/35 (77)	7	7	1	27/35 (77)	7	6	3
Recruitment process and sampling strategy	26/35 (74)	7	7	1	26/35 (74)	7	6	2
Data-gathering approaches[a]	26/35 (74)	7	7	0.75	26/35 (74)	7	6	1.75
Data documentation[a]	26/35 (74)	6	6	1.75	26/35 (74)	5	5.5	1
Data analysis	26/35 (74)	7	7	0.75	26/35 (74)	7	6	1.75
Processes used to ensure quality[b]	26/35 (74)	7	6	3	26/35 (74)	7	5	2.75
Characteristics of participants	26/35 (74)	7	6.5	1	26/35 (74)	7	6	2
Main findings	26/35 (74)	7	7	0.75	26/35 (74)	7	6	1
Summary of findings	26/35 (74)	7	7	1	26/35 (74)	6	6	1
Strengths, limitations and future research directions	26/35 (74)	7	6.5	1	26/35 (74)	6	6	1
Comparison with existing literature	26/35 (74)	7	7	1	26/35 (74)	7	6.5	1
Conclusion and recommendations	26/35 (74)	7	7	1	26/35 (74)	7	6	1.75
Funding	26/35 (74)	7	7	1	26/35 (74)	7	7	1

a These two items were combined, substantially reworded and included in round 3.
b This item was removed after discussion of its ratings with the project team.

TABLE 4 Summary of results for round 3 of Delphi panel

Item	Relevance				Validity			
	Response rate (%)	Mode	Median	IQR	Response rate (%)	Mode	Median	IQR
Data collection methods	28/35 (80)	7	7	1	28/35 (80)	7	6	2.25

Quality standards for evaluators and peer reviewers of realist evaluations

By peer reviewers, here, we specifically refer to individuals who have been asked to appraise the quality of completed evaluations. For each aspect of quality that requires a judgement about quality, we have provided a brief description of why the process is important, as well as descriptors of criteria against which a decision about quality might be arrived at. The quality standards for peer reviewers of realist evaluation reports are set out in *Table 5*.

As an illustrative example to explain how to use the layout of these quality standards, in the quality standard for '*4. Evaluation design*', this aspect of the evaluation could be judged as being adequate if, 'what was planned in the evaluation design, in what order and why was described and justified in detail'. For this aspect of an evaluation to be judged as 'good', we recommend that, as well as fulfilling the criteria for adequate (hence our use of the term 'adequate plus'), evaluations would need to ensure, among other things, that the 'adequate plus: the design "tested" multiple aspects of programme theory' criteria is fulfilled.

Quality standards for funders or commissioners of realist evaluations

As more and more realist evaluations are being undertaken, those commissioning the evaluations need to pass judgements on two broad areas: the proposed evaluation design and methodological expertise. We appreciate that many funding bodies and commissioners already have systems in place to guide their decision-making processes. However, a number of agencies have sought guidance about, or training in, how to assess the methodological aspects of realist tenders and proposals they have to deal with. As such, we see this guidance we have produced not as replacement for, but as a supplement to, existing organisational decision-making processes and guidance. We are also aware that funding bodies and commissioners have differences in the degree of involvement with the evaluations they have funded or commissioned. In response to these differences, these quality standards have been designed and worded in such a way that they may be used when an evaluation is still ongoing. The quality standards for realist evaluations for funders or commissioners of realist evaluations are set out in *Table 6*.

Developing, delivering and refining resources and training materials for realist evaluation

Two types of educational materials were developed: resource materials (made freely available online) and training materials.

The resource materials focus on topic areas that the literature review, Delphi panel and discussion list had identified as being most challenging and/or required further clarification. To make the materials accessible, we established a rough word limit of around 1000 words per topic, with very clearly defined topics, and written in as plain English as possible. This means that more introductory materials are accessible to those with very limited prior knowledge or experience of realist evaluation. It also means that more advanced readers can search for specific topics without having to wade through the more introductory resources, and that additional materials can easily be added in future.

Each of the resource materials provides references for those who wish to understand a topic area in greater depth. Many provide examples from previously completed evaluations to illustrate key points. Some provide direct links to more detailed articles on the same topic and/or to additional resources. For example, the 'realist interview' resource links to a longer journal article and to a list of questions that can be used in realist interviews or as a guide to start developing realist interview questions.

Most of the resource materials were written by one or two individuals within the project team and were then peer reviewed internally by a realist methodological expert. A couple were written by people outside the project team with interests in specific topics. These were each reviewed by at least two team members. The resource materials are open access and can be found on the RAMESES project website [http://ramesesproject.org (accessed 15 September 2017)].

TABLE 5 Quality standards for peer reviewers of realist evaluation reports

Quality standards for realist evaluation (for evaluators and peer reviewers)				
1. The evaluation purpose				
Realist evaluation is a theory-driven approach, rooted in a realist philosophy of science, which emphasises an understanding of causation and how causal mechanisms are shaped and constrained by context. This makes it particularly suitable for evaluations of certain topics and questions, for example complex interventions and programmes that involve human decisions and actions. A realist evaluation question contains some or all of the elements of 'what works, how, why, for whom, to what extent and in what circumstances, in what respect and over what duration?' and applies a realist logic to address the question(s). Above all, realist evaluation seeks to answer 'how' and 'why?' questions. Realist evaluation always seeks to explain. It assumes that programme effectiveness will always be conditional and is oriented towards improving understanding of the key contexts and mechanisms contributing to how and why programmes work				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
A realist approach is suitable for the purposes of the evaluation. That is, it seeks to improve understanding of the core questions for realist evaluation	• The evaluation does not seek to explain how and why the evaluand[a] works OR • There was no clear statement of the purpose(s) of the evaluation	• The evaluation seeks to explain how and why the evaluand works (or not) and to disaggregate outcomes for different subgroups and contexts • There is a statement of purpose(s) for the evaluation	Adequate plus: • The evaluation seeks to explain how and why the evaluand works differently in different contexts and for different subgroups: it seeks to explain how contexts affected mechanisms	Good plus: • Stated purpose clearly explains how the findings are intended to be used. There is a coherent argument as to why a realist approach is appropriate for those purposes
The evaluation question(s) are framed to be suitable for a realist evaluation	The evaluation question(s) are not structured to reflect the elements of realist explanation. For example, the question(s): • require only description; and/or • require only a numerical aggregation of outcomes; and/or • require only a summary of processes; and/or • rely exclusively on methods that are inadequate to generate realist understanding (e.g. 'a thematic analysis of . . . ')	The evaluation question(s) include a focus on how and why outcomes were generated in the evaluand, and contained at least some of the additional elements: *for whom, in what contexts, in what respects, to what extent and over what durations*	Adequate plus: • The questions address as many aspects of the realist question as are feasible within the constraints of the evaluation. The rationale for excluding any elements of 'the realist question' from the evaluation question(s) is explicit • (For example, the evaluation question may have sought only to explain how and why outcomes occur in certain contexts and not to what extent; the rationale for excluding 'extent' is described and reasonable in the circumstances) • The question(s) are sufficiently focused to be managed within the constraints of the evaluation	Good plus: • The evaluation question(s) are clear and as simple as possible. They can be understood by people without specialist methodological or content expertise

continued

TABLE 5 Quality standards for peer reviewers of realist evaluation reports (*continued*)

2. Understanding and applying a realist principle of generative causation in realist evaluations				
Realist evaluations are underpinned by a realist principle of generative causation – underlying mechanisms that operate (or not) in certain contexts to generate outcomes: Context + Mechanism = Outcome (CMO). Realist evaluations aim to understand how different mechanisms generate different outcomes in different contexts. This intent influences everything from the type of evaluation question(s) to an evaluation's design (e.g. the construction of a realist programme theory, recruitment process and sampling strategy, data collection methods, data analysis, to recommendations)				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
A realist principle of generative causation is applied	Significant misunderstandings of realist generative causation are evident. Common examples include the following: • Programme/intervention activities or strategies are mislabelled as mechanisms • Contexts are assumed to be directly causal, rather than affecting whether or not and how mechanisms operate • No attempts are made to uncover mechanisms • Outcomes are assumed to be caused by the programme/ intervention (rather than by underlying mechanisms) • Relationship(s) between an outcome, its causal mechanism(s) and context(s) are not explained or configured • If theory is provided, this is not explicitly linked to CMOCs	Some misunderstandings of realist generative causation are evident, but the overall approach is consistent enough that a recognisably realist analysis results from the process	Assumptions and methods used throughout the evaluation are consistent with a realist generative causation	Good plus: • The evaluation strategy demonstrates an exemplary understanding of a principle of realist generative causation, and application of methods consistent with that understanding throughout (e.g. in question(s), design and the evaluations outputs) • Emerging challenges arising as the evaluation unfolds are dealt with in ways that are consistent with realist generative causation

3. Constructing and refining a realist programme theory or theories

At an early stage in the evaluation, the main ideas that went into the making of an intervention, programme or policy (the programme theory or theories, which may or may not be realist in nature) are surfaced and made explicit. An initial tentative programme theory (or theories) is constructed, which sets out how and why an intervention, programme or policy is thought to 'work' to generate the outcome(s) of interest. Where possible, this initial tentative theory (or theories) will be progressively refined over the course of the evaluation

Over the course of the evaluation, if needed, programme theory (or theories) are 're-cast' in realist terms (describing the contexts in which, populations for which, and main mechanisms by which, particular outcomes are, or are expected to be, achieved). Ideally, the programme theory is articulated in realist terms prior to data collection in order to guide the selection of data sources about context, mechanism and outcome. However, in some cases, this will not be possible and the product of the evaluation will be an initial realist programme theory

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*

An initial tentative programme theory (or theories) is identified and developed. Programme theory is 're-cast' and refined as realist programme theory	Programme theory (or theories) are:
	• not developed; or
	• not articulated; or
	• described but not used in the evaluation; or
	• offered but not 're-cast' and refined as realist programme theory at any stage of the evaluation. In other words, the programme theory is not expressed in terms of the causal relationship between contexts, mechanisms and outcomes
		• Initial tentative programme theory (or theories) are identified and, as far as possible, described in realist terms (that is, in terms of the causal relationship between contexts, mechanisms and outcomes). These are refined as the evaluation progresses
		• Appropriate data are used to 'test' (confirm, refute or refine) selected aspects of programme theory (or theories)
		• Aspects of theory to be 'tested' (or not) are:
		  ○ specified and justified in the evaluation design
		  ○ appropriate to the purpose of the evaluation
		• The refined theory (or theories) are consistent with the evidence provided
		• Basic implications of the final programme theory (or theories) for practice in contexts examined in the evaluation are described
			Adequate plus:
			• Programme theory (or theories) are initially described in realist terms and used to inform all aspects of the evaluation (e.g. focusing an evaluation, identifying questions, determining what types of data are needed, from whom and where)
			• A range of appropriate types of data is used to test selected aspects of the theory, including triangulating evidence
			• Implications of the final programme theory (or theories) for practice in a range of contexts are described. A clear rationale is provided for the contexts in which the findings are applicable or not applicable
			• Where relevant, the programme theory or theories take into account the physical/material (e.g. environmental) and social aspects of systems necessary to answer evaluation questions and purposes
				Good plus:
				• The relationships between the programme theory (or theories) and relevant substantive theory (or theories) is articulated
				• A wide range of primary and secondary data are used to consolidate programme theory
				• Refinements to substantive theory are described, where appropriate
				• The final realist programme theory (or theories) comprises one or more CMOCs, describing how and why different mechanisms are triggered (or not) in different contexts to generate different outcomes
				• Implications of the final programme theory for a diverse range of contexts are comprehensively described. Relevant contexts which are not included in the evaluation were expressly addressed

continued

TABLE 5 Quality standards for peer reviewers of realist evaluation reports (*continued*)

4. Evaluation design				
Descriptions and justifications of what is planned in the evaluation design, in what order and why should be clearly articulated. Realist evaluations are ideally adaptive – that is, the evaluation question(s), scope and/or design may be adapted over the course of the evaluation to 'test' (confirm, refute or refine) aspects of the programme theory as it evolves. If changes are made to the evaluation design, these should be clearly described and justified. At the start of an evaluation, where possible, any changes that might be needed should be anticipated and contingencies planned				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
The evaluation design is described and justified	• The evaluation design is not clearly described or is not coherent • There is a lack of clarity as to what was planned in the evaluation design, in what order and why • The evaluation design does not clearly relate to or test the programme theory. For example, data collection methods used were unlikely to collect the relevant data needed to 'test' aspects of programme theory (see *Data collection methods* for more details) • Planned analyses are inconsistent with the assumptions underpinning realist evaluation (see *Data analysis* for more details)	• What was planned in the evaluation design, in what order and why is described and justified in detail • The evaluation design is informed by initial programme theory or theories, and 'tests' important or priority aspects of these OR • The evaluation is appropriately designed to develop realist programme theory • The design is coherent, with a logical flow from purpose through focus, questions, data collection and analysis methods	Adequate plus: • The design 'tests' multiple aspects of programme theory • The design enables alternative explanations to be investigated	Good plus: • The design is efficient, adding value by, for example, maximising use of existing data or increasing portability of findings • The design enables consideration of the extent to which the intervention contributes to overall outcomes, and/or identification of other aspects of the context (e.g. other policies or programmes) that are likely to contribute to outcomes
Ethical clearance is obtained if required	• No consideration is given to whether or not the evaluation required ethics approval • Ethics approval should have been sought, but was not (or was sought and declined)	• Protocols for ethics approval are considered and approval sought if required • Where ethics approval is sought, actions throughout the evaluation are consistent with the requirements of the ethics clearance obtained	• Proposals for ethics approval clearly distinguish the implications of the evaluation for different groups and different contexts • The proposal for ethics approval identifies the strategies for iteration in the design and steps to manage ethics in relation to such iteration	Specific implications of realist methodology are explained in the proposal for ethics approval [e.g. the need to link data across context, mechanism and outcome; the role of the evaluator(s) in relation to other stakeholders and the programme] and specific strategies to address those implications are included

5. Data collection methods

In a realist evaluation, a broad range of data increases the robustness of the theory 'testing' process and a range of methods used to collect them. Data will be required for all of context, mechanism and outcome, *and* to inform the relationships between them. Data collection methods should be adequate to capture not only intended, but also (as far as possible) unintended, outcomes (both positive and negative) and the context–mechanism interactions that generated them. Realist evaluation is usually multimethod (i.e. it uses more than one method to gather data). Where possible, data about outcomes should be triangulated (at least using different sources, if not different types, of information)

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*
Data collection methods are suitable for capturing the data needed in a realist evaluation	Within the realist evaluation project: • data collection methods are unclear; and/or • data collection methods are not theory driven (i.e. informed by the need to find data to confirm, refute or refine the programme theory); and/or • methods used are unlikely to capture necessary data (i.e. all of context, mechanism and outcome and the relationships between them)	• Methods for collecting and documenting data are driven by the programme theory (or theories) and: ○ capture the necessary data, including sampling necessary to test the programme theory; and ○ capture intended and unintended outcomes; and ○ address the evaluation questions • The rationale for the methods used is explained	Adequate plus: • Data collection methods are explicitly consistent with realist methodology (e.g. realist interviewing) • Quality control processes are adopted to ensure that data collection methods are applied rigorously and consistently • Allowance is made to collect additional data for further refinement of programme theory (or theories) and/or CMOCs as the evaluation unfolds • Data management processes (e.g. data bases, use of participant identifiers) are constructed to enable intended analyses (e.g. subgroup analyses, tracking participants over time)	• New data collection methods, tools and processes are adapted and/or developed where required and are consistent with realist principles • The specific techniques used or adaptations made to instruments or sampling processes are justified

continued

TABLE 5 Quality standards for peer reviewers of realist evaluation reports (*continued*)

6. Sample recruitment strategy				
In a realist evaluation, data are required for contexts, mechanisms and outcomes. One key source is respondents or key informants. Data are used to develop and refine theory about how, for whom and in what circumstances programmes generate their outcomes. This implies that any processes used to invite or recruit individuals need to identify an adequate sample of individuals who are able to provide information about contexts, mechanisms, outcomes and/or programme theory				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
The respondents or key informants recruited are able to provide sufficient data needed for a realist evaluation	• Recruitment is not designed to find respondents who could provide information about contexts, mechanisms and/or outcomes [e.g. recruitment was ad hoc and/or not informed by the programme theory (or theories)] • Random samples are used to generalise to whole populations (as distinct from sampling within theory-specified subgroups) • Convenience samples are used to 'test' (as distinct from develop) programme theories	Recruitment is: • designed to find an appropriate sample of respondents who can provide information about contexts, mechanisms and/or outcomes and the programme theory • purposive, with samples selected to test specific aspects of programme theory	Adequate plus: • Where needed, further recruitment is undertaken to collect the data needed for refinement of programme theory and/or CMOCs	• Sampling follows a rigorous and sequenced process of theory testing • A sufficiently large and diverse sample of relevant respondents is recruited to provide evidence across contexts and subgroups • When needed, respondents are re-interviewed as new evidence emerges, to explore context and mechanism extensively • Where applicable, sampling involves sensitive strategies to successfully recruit respondents from disenfranchised communities or other 'hard to reach' groups

7. Data analysis

Data analysis in realist evaluation is not a specific method but a way of interrogating programme theory (or theories) with data, *and* a way of using theory to understand patterns in data. In other words, data analysis is a way of teasing out what works, for whom, in what contexts, in what respects, over what duration and so on

In a realist evaluation, where possible, the analysis process should occur iteratively. The overall approach to data analysis is retroductive[b] (i.e. it moves between inductive and deductive processes, includes and tests researcher 'hunches' and aims to provide the best possible explanation of acknowledged-to-be-incomplete data). The processes used to analyse the data and integrate them into one or more realist programme theories should be consistent with a central principle of realism – namely generative causation. How these data are then used to further develop, confirm, refute or refine one or more programme theories should be clearly described and justified

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*
The overall approach to analysis is retroductive[b]	• The approach to analysis is not retroductive OR • The overall approach to analysis is not clear	• The approach to analysis moves between theory and data, data and theory appropriate to the stage of theory development	Adequate plus: • Theory is developed and refined through the use of retroductive reasoning. Evaluators' 'hunches' are clearly described • Theories that remain untested are specified and described	Good plus: • Analysis clearly links data, programme theory and formal theory
Data analyses processes applied to gathered data are consistent with a realist principle of generative causation	• Analytic processes are not described • Analysis is not disaggregated by subgroups (i.e. 'for whom') and/or contexts • Subgroup analyses are undertaken without reference to programme theory (e.g. disaggregating by gender, age or other demographic subgroups without specifying how they are relevant to theory, rather than on theory-relevant groupings)	• Qualitative analysis moves beyond thematic categorisation to identify and explain the relationships between contexts, mechanisms and outcomes • Quantitative analysis is hypothesis-driven to 'test' differences between subgroups or contexts, in relation to programme theory • Findings from analysis are organised to demonstrate relationships between context, mechanism and outcome (i.e. evidence is aligned against programme theory)	Adequate plus: • Specific analyses are conducted to 'test' the relationships within, and between, CMOCs (e.g. correlations analysis for quantitative data; analysis of narrative, argument or speech/ text to identify causal relationships in qualitative data) • That is, evidence is not just aligned against programme theory: the linkages within the programme theory are 'tested' • Weaknesses in analytic methods for realist purposes were acknowledged and choices justified	Good plus: • When iterations in evaluation design and/or programme theory required additional analytic methods, these methods were consistent with realist principles

continued

TABLE 5 Quality standards for peer reviewers of realist evaluation reports (*continued*)

7. Data analysis				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***

A realist logic of analysis is applied to develop and refine theory	The analysis does not:
	• identify contexts, mechanisms or outcomes
	• identify the relationships between contexts, mechanisms and outcomes
	• describe how the programme theory (or theories) was further developed, confirmed, refuted and refined
		• Data are analysed to develop and refine initial programme theory (or theories) into realist programme theory (or theories)
		• The realist analysis:
		  ○ Assigns conceptual labels of C, M or O to each data element or finding within a context–mechanism–outcome configuration (CMOC) – (e.g. 'in this aspect of the analysis, this item of data are functioning as context')
		  ○ Identifies the relationship between contexts, mechanisms and outcomes within particular CMOCs
		  ○ Identifies relationships across CMOCs [i.e. the location and interactions between CMOCs within a programme theory (or theories)]
			Adequate plus:
			• Analysis integrates a range of data sources (e.g. qualitative and quantitative, primary and secondary data) and describes how multiple data types were integrated to support inferences
				Data analysis is iterative over the course of the evaluation, with earlier stages of analysis being used to refine programme theory and/or refine evaluation design for subsequent stages

8. Reporting

Realist evaluations may be reported in multiple formats – detailed reports, summary reports, articles, websites and so on. Reports should be consistent with the RAMESES II reporting standards for realist evaluations (see https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0643-1)

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*
The evaluation is reported using the items listed in the RAMESES II reporting standard for realist evaluations	Key items are missing. For example: • No defined evaluation question(s) • Limited or no reporting of the evaluation's methods • Limited or no explanations and justifications provided for any adaptations made to the realist evaluation approach • Insufficient detail to enable readers to judge the trustworthiness and plausibility of findings	Most items in the RAMESES II reporting standards for realist evaluations are reported. In particular: • Item 3: rationale for evaluation • Item 4: programme theory • Item 5: evaluation questions(s), objectives and focus • Item 6: ethics • Method section items 8 (environment for the evaluation), 9 (description of the evaluand), 11 (data collection methods), 12 (recruitment and sampling), 13 (data analysis) and 15 (main findings)	• All items are clearly reported and in sufficient detail for an external reader to understand and judge the methods used and the trustworthiness and plausibility of findings • Where an item is not reported, a justification is provided	Good plus: • Additional materials are made available for external readers to investigate aspects of the evaluation in more detail
Findings and implications are clear and reported in formats that are consistent with realist assumptions	• Findings are unclear or difficult to follow • Findings are not reported in realist format (e.g. average results are reported but do not address issues such as 'for whom' or 'in what circumstances') • Lists of contexts, mechanisms and outcomes are provided without reporting causal relationships between them • Evidence is not clearly linked to context, mechanism or outcome	• Findings are clearly reported • All conclusions follow logically from the analyses • Findings explain how and why different patterns of outcomes are generated in different contexts or for different groups	Adequate plus: • Implications for policy, programmes and/or practice are clearly explained and follow logically from the analysis • Implications and/or recommendations take into account issues or strategies for different contexts or groups • Summaries of findings maintain patterns of outcomes (e.g. findings are not summarised by resorting to average effects)	Good plus: • The report is well written, transparent and easy to understand • Various reporting formats are used to present relevant findings to different audiences

C, context; M, mechanism; O, outcome.

a Evaluand is defined as 'that which is being evaluated'. For example, an intervention, programme, policy, product or initiative, or, in some cases, sets of programmes, policies or initiatives.

b For more details on retroduction see 'Retroduction in realist evaluation', which may be found in the standards and training materials section of the RAMESES projects website (www.ramesesproject.org).

TABLE 6 Quality standards for funders or commissioners of realist evaluations

Quality standards for realist evaluation (for funders or commissioners of realist evaluations)				
1. The evaluation purpose				
Realist evaluation is a theory-driven approach, rooted in a realist philosophy of science, which emphasises an understanding of causation and how causal mechanisms are shaped and constrained by context. This makes it particularly suitable for evaluations of certain topics and questions, for example complex interventions and programmes that involve human decisions and actions. A realist evaluation question contains some or all of the elements of 'what works, how, why, for whom, to what extent and in what circumstances, in what respect and over what duration?' and applies a realist logic to address the question(s). Above all, realist evaluation seeks to answer 'how?' and 'why?' questions. Realist evaluation always seeks to explain. It assumes that programme effectiveness will always be conditional and is oriented towards improving understanding of the key contexts and mechanisms contributing to how and why programmes work				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
A realist approach is suitable for the purposes of the evaluation	• There is no statement of the purpose of the evaluation AND/OR • The evaluation does not seek to explain how and why the evaluand[a] works	• There is a clear statement of purpose for the evaluation AND/OR • The evaluation seeks to explain how and why the evaluand works	Adequate plus: • The evaluation seeks to explain how and why the evaluand works differently in different contexts and for different subgroups	Good plus: • Stated purpose clearly explains how the findings are intended to be used • There is a coherent argument as to why a realist approach is appropriate
The evaluation question(s) are framed in such a way as to be suitable for a realist evaluation	The evaluation question(s) are not structured to reflect the elements of realist explanation. For example, answering the questions: • requires only description; and/or • requires only a numerical aggregation of outcomes; and/or • requires only summary of processes; and/or • relies exclusively on methods that are inadequate to generate realist understanding (e.g. 'a thematic analysis of . . .')	The evaluation question(s) include a focus on how and why outcomes are likely to be generated, and contain at least some of the additional elements, 'for whom, in what contexts, in what respects, to what extent and over what durations'	Adequate plus: • The rationale for excluding any elements of 'the realist question' from the evaluation question(s) is explicit • The question(s) are sufficiently focused to be managed within a realist evaluation	Good plus: • The evaluation question(s) are clear and as simple as possible. They can be understood by people without specialist methodological or content expertise

2. Understanding and applying a realist principle of generative causation in realist evaluations

Realist evaluations are underpinned by a realist principle of generative causation. That is, underlying causal processes (called 'mechanisms') operate (or not) in certain contexts to generate outcomes. The explanatory framework is Context + Mechanism = Outcome (CMO). Realist evaluations aim to understand how different mechanisms generate different outcomes in different contexts. This intent influences everything from the type of evaluation question(s) to an evaluation's design (e.g. the construction of a realist programme theory, recruitment process and sampling strategy, data collection methods, data analysis, to recommendations)

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*
A realist principle of generative causation is applied	Significant misunderstandings of realist generative causation are evident. Common misunderstandings include the following: • Programme activities or strategies are mislabelled as mechanisms • Contexts are assumed to cause outcomes directly, rather than affecting whether or not and how mechanisms operate • Outcomes are assumed to be caused directly by the programme/intervention (rather than by underlying mechanisms) • No attempts are made to understand underlying mechanisms • Relationships between an outcome, its causal mechanism(s) and context(s) are not explained • If theory is provided, this is not explicitly linked to CMOCs	Some misunderstandings of realist generative causation exist, but the overall approach is consistent enough that a recognisably realist analysis results from the process	Assumptions and methods used throughout the evaluation are consistent with a realist generative causation	Good plus: • The evaluation strategy demonstrates exemplary understanding of a principle of realist generative causation, and application of methods consistent with that understanding throughout [e.g. in question(s), design and the evaluations outputs] • Emerging challenges arising as the evaluation unfolds are dealt with in ways that are consistent with realist generative causation

continued

TABLE 6 Quality standards for funders or commissioners of realist evaluations (*continued*)

3. Constructing and refining a realist programme theory or theories				
At an early stage in the evaluation, the main ideas that went into the making of an intervention, programme or policy (the programme theory or theories, which may or may not be realist in nature) are identified and described. An initial tentative programme theory (or theories) is constructed, which sets out how and why an intervention, programme or policy is thought to 'work' to generate the outcome(s) of interest. Where possible, this initial tentative theory (or theories) is progressively refined over the course of the evaluation Over the course of the evaluation, if needed, programme theory (or theories) is 're-cast' in realist terms (describing the contexts in which, populations for which and main mechanisms by which particular outcomes are expected to be achieved). Ideally, the programme theory is articulated in realist terms prior to data collection in order to guide the selection of data sources about context, mechanism and outcome. However, in some cases, this will not be possible and the product of the evaluation will be an initial realist programme theory				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***

An initial tentative programme theory (or theories) is, or will be, identified and developed. Programme theory is or will be 're-cast' and refined as realist programme theory	Programme theory (or theories):
	
	• are not or will not be developed; or
	• are described but it is not clear how they were or will be used in the evaluation; or
	• are offered but it is not clear how they were or will be refined as realist programme theory during the evaluation
		• Initial tentative programme theory (or theories) are or will be identified and, as far as possible, described in realist terms (that is, in terms of the causal relationship between contexts, mechanisms and outcomes). These are or will be refined as the evaluation progresses
		• Where possible, aspects of theory to be 'tested' are:
		  ○ Specified and justified in the evaluation design
		  ○ Appropriate to the purpose of the evaluation
		
		• Aspects that will not be tested are identified and explanation is provided as to why
			Adequate plus:
			
			• Programme theory (or theories) are described in realist terms and used to inform all aspects of the evaluation (e.g. focus an evaluation, identify questions, determine what types of data need to be collected and from whom and where)
			• Where relevant, the programme theory or theories take into account the physical/material (e.g. environmental) and social aspects of systems necessary to answer evaluation questions
				Good plus:
				
				• The relationships between the programme theory (or theories) and relevant formal theory (or theories) will be sought
				• Where relevant, contexts which are not included in the evaluation are expressly addressed
				• The final realist programme theory (or theories) comprise one or more CMOCs, describing how and why different mechanisms are triggered (or not) in different contexts to generate different outcomes

4. Evaluation design

Descriptions and justifications of what is planned in the evaluation design, in what order and why should be clearly articulated. Realist evaluations are ideally adaptive; that is, the evaluation question(s), scope and/or design may be adapted over the course of the evaluation to 'test' (confirm, refute or refine) aspects of the programme theory as it evolves. If changes are made to the evaluation design, these should be clearly described and justified. At the start of an evaluation, where possible, any changes that might be needed should be anticipated and contingencies planned

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*
The evaluation design is described and justified	• The evaluation design is not clearly described or is not coherent • There is a lack of clarity as to what is planned in the evaluation design, in what order and why • The evaluation design does not clearly relate to or test the programme theory • The analyses are inconsistent with the assumptions underpinning realist evaluation	• What is planned in the evaluation design, in what order and why is described and justified in detail • The evaluation design is informed by an initial programme theory or theories, and sets out 'tests' important or priority aspects of these • The design is coherent, with a logical flow from purpose through focus, questions, data collection and analysis methods	Adequate plus: • The design tests multiple aspects of programme theory • The design enables alternative explanations to be investigated	Good plus: • The design is efficient, adding value by, for example, maximising use of existing data or increasing portability of findings • The design identifies or will identify the extent to which the interventions contribute to overall outcomes, and/or identifies other aspects of the context (e.g. other policies or programmes) that are likely to contribute to outcomes
Ethical clearance is or will be obtained if required	No consideration is given to whether or not the evaluation requires ethics approval	Protocols for ethics approval are considered and approval sought if required	Proposals for ethics approval clearly distinguish the implications of the evaluation for different groups and different contexts	Where relevant, specific implications of realist methodology are explained in the proposal for ethics approval and specific strategies to address those implications are provided

continued

TABLE 6 Quality standards for funders or commissioners of realist evaluations (*continued*)

5. Data collection methods				
In a realist evaluation, a broad range of data increases the robustness of the theory 'testing' process and a range of methods used to collect data. Data will be required for all of context, mechanism and outcome, *and* to inform the relationships between them. Data collection methods should be adequate to capture not only intended, but also, as far as possible, unintended, outcomes (both positive and negative), and the context–mechanism interactions that generated them. Realist evaluation is usually multimethod (i.e. uses more than one method to gather data). Where possible, data about outcomes should be triangulated (at least using different sources, if not different types, of information)				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***

Data collection methods are suitable for capturing the data needed in a realist evaluation	Within the realist evaluation project:
	• It is unclear which data collection methods are used; and/or
	• Data collection methods are not informed by the need to find data to confirm, refute or refine the programme theory; and/or
	• Methods used are unlikely to capture necessary data to test the programme theory
		• Methods for collecting and documenting data are driven by the programme theory (or theories) and
		  ○ Will capture the necessary data; and
		  ○ Will capture intended and unintended outcomes
		They will also consider:
		• The sampling needed to 'test' programme theory; and
		• The evaluation questions; and
		• The rationale for the methods and its implications for data analysis is explained
			Adequate plus:
			• Data collection methods are explicitly consistent with realist methodology (e.g. realist interviewing)
			• Quality control processes ensure that data collection methods are applied rigorously and consistently
			• Allowance is made to collect additional data for further refinement of programme theory (or theories) and/or CMOCs as the evaluation unfolds
			• Data management processes (e.g. data bases, use of participant identifiers) are or will be constructed to enable intended analyses (e.g. subgroup analyses, tracking participants over time)
				• New data collection methods, tools and processes are adapted and/or developed where required and are consistent with realist principles
				• Any specific techniques used, or adaptations made, to instruments or sampling processes are justified

6. Sample recruitment strategy

In a realist evaluation, data are required for all of the context, mechanisms and outcomes. One key source is respondents or key informants. Data are used to develop and refine theory about how, for whom and in what circumstances programmes generate their outcomes. This implies that any processes used to invite or recruit individuals need to identify an adequate sample of individuals who are able to provide information about contexts, mechanisms, outcomes and/or programme theory

Criterion	*Inadequate*	*Adequate*	*Good*	*Excellent*
The respondents or key informants recruited are likely to be able to provide sufficient data needed for a realist evaluation	• Recruitment is or was ad hoc, opportunistic and/or not informed by the programme theory • Random samples are or will be used to generalise to whole populations (as distinct from sampling within theory-specified subgroups) • Convenience samples not related to programme theory are or will be used to test programme theories	Recruitment is: • Designed to find an appropriate sample of respondents who can provide information about contexts, mechanisms and/or outcomes for the programme theory • Purposive, with samples selected to test specific aspects of programme theory	Adequate plus: • Where needed, further recruitment is or will be undertaken to collect the data needed for further refinement of programme theory	• Sampling follows a rigorous and sequenced process of theory testing • A sufficiently large and diverse sample of relevant respondents is or will be recruited to provide evidence across contexts • When needed, respondents will be approached again as new evidence emerges, to explore context and mechanism more extensively • Where applicable, sampling will involve sensitive strategies to successfully recruit respondents from disenfranchised communities or other 'hard to reach' groups

continued

TABLE 6 Quality standards for funders or commissioners of realist evaluations (*continued*)

7. Data analysis				
Data analysis in realist evaluation is not a specific method but a way of interrogating programme theory (or theories) with data, *and* a way of using theory to understand patterns in data. In other words, data analysis is a way of teasing out what works, for whom, in what contexts, in what respects, over what duration and so on In a realist evaluation, where possible, the analysis process should occur iteratively. The overall approach to data analysis is retroductive[b] (i.e. it moves between inductive and deductive processes, includes and tests researcher 'hunches' and aims to provide the best possible explanation of acknowledged-to-be-incomplete data). The processes used to analyse the data and integrate them into one or more realist programme theories should be consistent with a central principle of realism – namely generative causation. How these data are then used to further develop, confirm, refute or refine one or more programme theories should be clearly described and justified				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
The overall approach to analysis is or will be retroductive[b]	• The approach to analysis is not retroductive OR • The overall approach to analysis is not clear	• The approach to analysis moves or will move between theory and data, data and theory, appropriate to the stage of theory development	Adequate plus: • Any theory (or theories) are developed and refined through the use of retroductive reasoning. Evaluators' 'hunches' are clearly described • Theories that remain untested at the end of the evaluation are identified	Good plus: • The analysis clearly links data, programme theory and formal theory
Data analyses processes are consistent with a realist principle of generative causation	• Analytic processes are not described • Analysis is not or will not be disaggregated by subgroups (i.e. 'for whom') or contexts • Subgroup analyses are planned without reference to programme theory (e.g. disaggregating by demographic subgroups rather than theory-relevant groupings)	• Qualitative analysis identifies and explains the relationships between contexts, mechanisms and outcomes • Quantitative analysis 'tests' differences between subgroups or contexts, in relation to programme theory • Findings from analysis are aligned against programme theory	Adequate plus: • Specific analyses are or will be conducted to 'test' the relationships within and between CMOCs. That is, evidence is not just aligned against programme theory: the linkages within the programme theory are 'tested'	Good plus: • When iterations in evaluation design and/or programme theory require additional analytic methods to be employed, those used are consistent with realist principles

7. Data analysis				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***

A realist logic of analysis is used to develop and refine theory	The analyses used or planned do not:
	• Identify contexts, mechanisms or outcomes
	• Identify the relationships between contexts, mechanisms and outcomes; and/or
	• Explain how the programme theory (or theories) are or will be further developed, confirmed, refuted and refined
		• Data are or will be analysed to develop and refine initial programme theory (or theories) into realist programme theory (or theories)
		• The realist analysis has or will:
		  ○ Assign conceptual labels of C, M or O to each data element or finding within a CMOC – (e.g. 'in this aspect of the analysis, this item of data are functioning as context within this CMOC')
		  ○ Identify the relationship of contexts, mechanisms and outcomes within particular CMOCs
		  ○ Identify relationships across CMOCs; that is, the location and interactions between CMOCs within a programme theory (or theories)
			Adequate plus:
			• Analysis integrates a range of data sources (e.g. qualitative and quantitative, primary and secondary data) and describes how the multiple data types were or will be integrated to support inferences
				• Data analysis is or will be iterative over the course of the evaluation, with earlier stages of analysis being used to refine programme theory and/or refine evaluation design for subsequent stages

8. Reporting				
Realist evaluations may be reported in multiple formats – detailed reports, summary reports, articles, websites and so on. Reports should be consistent with the RAMESES II reporting standards for realist evaluations (see https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0643-1)				
Criterion	***Inadequate***	***Adequate***	***Good***	***Excellent***
The realist evaluation is or will be reported using the items listed in the RAMESES II reporting standards for realist evaluations	• No information is provided on whether or not the RAMESES II reporting standard for realist evaluations will be used	• The RAMESES II reporting standard for realist evaluations is or will be used	A firm commitment is made to: • Use the RAMESES II reporting standard for realist evaluations • Provide justifications where items will not been reported	Good plus: • The report is well written, transparent and easy to understand • Various reporting formats are used to present relevant findings to different audiences

C, context; M, mechanism; O, outcome.

a Evaluand is defined as 'that which is being evaluated'; for example, an intervention, programme, policy, product or initiative, or in some cases, sets of programmes, policies or initiatives.

b For more details on retroduction see 'Retroduction in realist evaluation', which may be found in the Standards and Training materials section of The RAMESES Projects website (www.ramesesproject.org).

An overview of the topic areas currently covered may be found in *Table 7*. Additional topics are also planned by members of the project team, to be added at a later date.

Support and consultancy to realist evaluations

We were approached by a wide range of evaluators who asked us for help with their realist evaluation projects. Selection was done on a 'first-come, first-served' basis. An overview of the 17 realist evaluation projects we provided methodological support and consultancy to may be found in *Table 8*.

Realist evaluation and 'training the trainers' workshops

We provided training workshops to organisations interested in learning more about realist evaluation on a first-come, first-served basis. When we were contacted, we entered into discussion with the individuals who contacted us and arranged bespoke training to meet their needs. These ranged from short 15-minute presentations to whole-day workshops. *Table 9* lists the 29 realist evaluation presentations or workshops we ran nationally and internationally.

TABLE 7 Summary of the topics covered in the training materials for realist evaluations

Topic area	Brief summary of contents
Realist evaluation, realist synthesis, realist research – what's in a name?	Definition and explanations of the differences between realist evaluation, review and research
What is a mechanism? What is a programme mechanism?	Explanation of the concept of a mechanism
What do realists mean by context, or, why nothing works for everywhere for everyone	Explanation of the concept of context
Protocols and realist evaluation	Explains what a realist evaluation protocol consists of and why
Philosophies and evaluation design	A short description of factors to be taken into account in designing a realist evaluation and how these may differ from some other designs
Realist evaluation and ethical considerations	Issues in writing research ethics applications and strategies to address them
Developing realist programme theories	Processes for developing (or 'surfacing') initial programme theories for realist evaluations
The realist interview	Explanation of how realist interviews differ from other interviews, and their role in realist evaluation
Realist evaluation interviewing – a 'starter set' of questions	Provides evaluators with a series of example questions and the rationale for their use
A realist understanding of programme fidelity	Discussion of the idea of fidelity within realist evaluation
'Theory' in realist evaluation	Explains the different types of theory used in realist evaluation
Working with a librarian on a realist review	Some realist evaluations involve an initial realist review. This document provides hints about how librarians may be able to assist, and how to enable them to support researchers, in realist research and evaluation
Realist evaluation: an introduction for commissioners	A short introduction for commissioners of evaluations, including when to commission a realist evaluation, what to include in the request for tender and how to assess tenders
Retroduction in realist evaluation	Explains what retroduction is and how it is used in realist research
Frequently asked questions about realist evaluation	Covers the frequently asked questions about realist evaluations and signposts reader to further resources

TABLE 8 Overview of the realist evaluations for which the project team provided methodological support or consultancy

Evaluation title	Evaluation aim(s)/questions(s)/focus	Funder/commissioner	Type of support provided
When cure is not likely: What do young adults with cancer and their families need and how can it best be delivered? A BRIGHTLIGHT companion study	• The most important parts of care in the last year of life for people with cancer aged 16–40 years • Whether or not differences exist between the experiences of people with cancer who are aged 16–24 years and those aged 25–40 years • How young adults and their families can be supported in the last year of life to achieve their preferences for care • The challenges that exist for health and social care professionals providing care	Marie Curie, UK	• Bespoke realist evaluation training • Attending project team meetings • Assistance with data analysis • Assistance with project publications
Is bigger better? Lessons for large-scale general practice	• How is the landscape of general practice changing? How quickly, and in what form, are new large-scale general practice organisations emerging? What are the factors driving the formation of these new organisations? • For a small sample of mature large-scale general practice organisations, how have they emerged and evolved over time? • How have organisational, local, national and other contextual factors affected the abilities of mature large-scale general practice organisations to achieve their goals over time? • What impacts are the organisations having on their patients, staff and the local health economy? • What impacts on quality of care can we measure?	Nuffield Trust, UK	• Bespoke realist evaluation training • Attending project team meetings • Assistance with data analysis Note that for logistical reasons the evaluation team decided not to undertake a realist evaluation
Determinants of effectiveness of a novel community health workers programme in improving maternal and child health in Nigeria	To better understand to what extent, and under what conditions, a community health workers programme (with or without conditional cash transfers) contributes to achieving equitable access to quality services and maternal and child health outcomes in Nigeria	MRC Joint DFID/ESRC/MRC/ Wellcome Trust Health Systems Research Initiative	• Assistance with theory development and refinement • Attendance at project meetings • Assistance with project publications • Workshops and webinars on realist evaluation methods
Investigating the communications component of dental complaints: towards a needs-based communications resource	To explore the characteristics of dental communication between dentists who are vulnerable to receiving complaints and their patients, so as to design a needs-based communications resource	NIHR doctoral fellowship	Assistance with study design and initial programme theory development. Proposal to be submitted in 2017

continued

TABLE 8 Overview of the realist evaluations for which the project team provided methodological support or consultancy (*continued*)

Evaluation title	Evaluation aim(s)/questions(s)/focus	Funder/commissioner	Type of support provided
Building Capacity to Use Research Evidence (BCURE)	To build the capacity of policy-makers in several low- and middle-income countries to use research evidence more effectively in decision-making	UK Department for International Development	Guidance on qualitative data collection techniques (topic guides for interview and focus groups)
Sea swimming and mental wellbeing	• To evaluate the benefits of sea swimming for mental health and wellbeing in coastal areas • To understand the potential and limitations of water cures – sea bathing in particular – to be used to manage health and improve wellbeing. To consider the extent to which these ideas and practices were/are modified by the age, gender, class, power and ethnicity of patients	Arts and Humanities Research Council	• Attending project team meetings • Workshop on realist evaluation • Assistance with project design Project was not funded
Developing and evaluating a collaborative care intervention for offenders with common mental health problems, near to and after release	To develop a way of organising care for men with common mental health problems as they approach being released from prison	NIHR's Programme Grant for Applied Research programme	Guidance on best-practice examples of collecting primary quantitative data in realist evaluations
Involving radiographers in mammography image interpretation and reporting in symptomatic breast clinics: a realist evaluation	In what circumstances, how and why can radiographers substitute the work of radiographers in mammography image interpretation and reporting in symptomatic breast clinics	NIHR doctoral training fellowship held by Anne-Marie Culpan	Acted as a doctoral supervisor to Anne Marie Culpan and provided support in: • study design • data collection methods • analysis • thesis structure and write up
A realist process evaluation of robotic surgery: integration into routine practice and impacts on communication, collaboration, and decision making	• What are the components on which successful integration of robotic surgery depends? • What contextual factors impact integration of robotic surgery? • How does communication and teamwork differ between laparoscopic and robotic surgery? • What are the consequences of differences in communication and teamwork for outcomes?	NIHR's HSDR programme	• Assistance with theory development and refinement • Attendance at project and steering group meetings • Assistance with writing a chapter in the final report • Assistance with project publications
Realist evaluation of adapted sex offender treatment interventions for people with learning disabilities	What works on Adapted Sex Offender Treatment Programs (ASOTPs) for whom, in what contexts, why and how?	ESRC new investigator award held by Andrea Hollomotz	Mentor to Andrea Hollomotz • Assistance with study design and analysis • Assistance with project publications

TABLE 8 Overview of the realist evaluations for which the project team provided methodological support or consultancy (*continued*)

Evaluation title	Evaluation aim(s)/questions(s)/focus	Funder/commissioner	Type of support provided
Values based recruitment: what works, for whom, why, and in what circumstances?	How have education and service providers implemented values-based recruitment approaches and what are the impacts on service delivery and care?	DH's Policy Research programme	• Member of the advisory group • Advice on study design • Attendance at project steering group meetings plus ad hoc meetings
The use of Pressure Ulcer Risk Assessment Instruments in clinical practice: A Realist Evaluation	To understand how hospital ward teams use PURPOSE-T and another commonly used risk assessment form and how their use impacts on: • the care that patients receive • communication between health professionals, patients and carers • patient outcomes (e.g. pressure ulcer development and management)	NIHR postdoctoral fellowship (started October 2016) held by Susanne Coleman	Supervisor on Susanne Coleman's successful NIHR postdoctoral fellowship proposal • Assistance with study design and submission of proposal • Attendance at supervision team meetings • Assistance with some analysis
Assessing the feasibility of implementing and evaluating a new problem-solving model for patients at risk of self-harm and suicidal behaviour in prison	Assessment of the feasibility and acceptability of the problem solving intervention, using qualitative methods	NIHR research for patient benefit	• Attendance at project team meetings • Analysis of data • Continuing to provide support to the study Note that for logistical reasons the team adopted a theories-of-change approach to the study, rather than realist evaluation
An Evaluation of the Leeds Curriculum	An evaluation of the impact of the Leeds Curriculum on the delivery of student education and student experience	Internal – University of Leeds, Leeds, UK	• Assistance with study design and discussion with stakeholders to elicit programme theories • Attendance at project team meetings Note that the team decided to adopt a development evaluation approach to the study, and support discontinued in December 2016
Medical Technologies Innovation – Closing the Early Stage Translation Gap in the Leeds City Region	How does sector-specific support in research translation, innovation training and development, and access to wider networks of project partners, support and embed research translation capability in Medical Technologies across five partner universities within the Leeds City region?	Higher Education Funding Council	• Meetings with the project managers and project stakeholders • Workshop with project stakeholders to identify programme theories underlying the programme Note that as of October 2016 the team recruited their own evaluation manager and felt support was no longer needed

ESRC, Economic and Social Research Council; DFID, Department for International Development; DH, Department of Health; HSDR; Health Services and Delivery Research; MRC, Medical Research Council.

TABLE 9 List of realist evaluation presentations and workshops

Date	Venue
April 2015	University of Oxford, Oxford, UK
May 2015	Nuffield Trust, London, UK
June 2015	University of Leeds, Leeds, UK
July 2015	University of Waterloo, Waterloo, ON, Canada
July 2015	White Rose Doctoral Training Centre, University of Leeds, Leeds, UK
August 2015	London School of Hygiene and Tropical Medicine, London, UK
September 2015	Diakonhjemmet University College/Gjøvik University College, Oslo, Norway
October 2015	Oxford Policy Management, Oxford, UK
October 2015	21st Qualitative Health Research Conference, Toronto, ON, Canada
November 2015	Centre for Evidence Based Intervention, Oxford, UK
November 2015	Researching Medical Education Conference, London, UK
November 2015	Realism Leeds Conference, Leeds, UK
February 2016	University College Cork, Cork, Ireland
March 2016	HM Treasury, London, UK
April 2016	University of Oxford, Oxford, UK
May 2016	Keele University, Keele, UK
May 2016	Health and Wellbeing Research Institute – Sheffield Hallam University, Sheffield, UK
June 2016	RDS East Midlands, Nottingham, UK
June 2016	Health Services Management Centre, University of Birmingham, Birmingham, UK
July 2016	ESRC Research Methods Conference, Bath, UK
July 2016	University of Plymouth, Plymouth, UK
July 2016	White Rose Doctoral Training Centre, University of Leeds, Leeds, UK
September 2016	DFID Joint Evaluation and Statistics Professional Development Conference, Oxford, UK
September 2016	European Evaluation Society Conference, Maastricht, the Netherlands
October 2016	RDS East Midlands, Nottingham, UK
October 2016	Cochrane Colloquium, Seoul, South Korea
October 2016	International Conference on Realist Evaluation and Synthesis, London, UK
December 2016	Division of Rehabilitation and Ageing, University of Nottingham, Nottingham, UK
February 2017	University of Leeds, Leeds, UK

ESRC, Economic and Social Research Council; DFID, Department for International Development; HM Treasury, Her Majesty's Treasury.

In terms of 'training the trainers' workshops, we wanted to build capacity within the NIHR RDS. We discussed what the training needs might be initially with colleagues at the RDS London's East London Team. Their feedback was supplemented with comments we received from our project's Advisory Group. After the publication of our project's protocol paper,[15] we were contacted by colleagues from the RDS East Midlands and, with their assistance, organised two workshops for regional staff.

Develop, deliver and refine information and resources for patients and other lay participants in realist evaluation

To develop resources for patients and other lay participants in realist evaluation, we first discussed, within the project team, what might be required. We also sought input from our project's Advisory Group. We then drafted a specimen document outlining what a realist evaluation is and when it might be used, and this also explained what might be expected of a participant when taking part in a realist evaluation. We did not develop any materials for seeking ethics approvals as we established that organisations or institutions had a diverse range of processes, and so a one-size-fits-all set of documents was not likely to be useful. To gain feedback on the documents, we convened a 90-minute face-to-face meeting with six members of the public from diverse backgrounds in September 2016 in Oxford (only five out of the six invited participants attended on the day). This meeting was facilitated by Geoff Wong, who made contemporaneous notes. At this meeting, we introduced ourselves and then proceeded to explain the purpose of the session. The participants then spent time refining and providing feedback on the documents we provided. We also discussed their ideas about best how to present this information. The session finished with a summary of what they had suggested, and also a way of taking their proposals forward. Based on their suggestions and feedback, Geoff Wong drafted new materials and these were sent round to the participants for comments and feedback.

In brief, after some clarification, the participants felt that it probably does not matter to the person who is being recruited into a realist evaluation what exactly a realist evaluation is. In other words, the detail of what a realist evaluation is or is not was unlikely to matter to the potential participant, so much of the detail in the text of the documents we initially provided was not needed. We were advised the text should be short and kept to half of a side of A4- or one side A5-sized paper. The agenda and notes from the session may be found in *Appendix 5*. The only new material that the participants felt was needed was a 'generic' text that could be used in a patient information leaflet when recruiting to a realist evaluation, and this can be found in *Box 2*.

BOX 2 Generic text for patient information leaflets

[INSERT PROJECT TITLE]

Example text: Evaluation of the NHS Health Checks programme

[INSERT BRIEF DESCRIPTION OF THE PURPOSE OF THE PROJECT]

Example text: The NHS Health Checks programme is a national programme that offers a free 'MOT' or health check to anyone over the age of 40.

We are researchers/evaluators [DELETE AS APPROPRIATE] from [INSERT ORGANISATION]. We are trying to find out why this programme does, or does not, work for different people. For this, we need your help.

We are interested to know your reasons for taking part in this programme or, if you are not taking part, what your reasons are.

To do this we will . . . [INSERT DATA COLLECTION METHODS]

Example text . . . ask you some questions/watch what happens when you take part in the programme/ask you to join a group where we discuss the programme/ask you to write a diary about the programme, etc.

BOX 2 Generic text for patient information leaflets *(continued)*

We will be using a research method called 'realist evaluation'. If you want to find out more about this method, please . . .. [INSERT PROCESS]

Example text: ask a member of our project team/visit the website, etc.

We hope you agree to take part, and thank you in advance for your time.

To take part please . . .. [INSERT RECRUITMENT PROCESS]

Example text: speak to a member of our project team/e-mail . . . /call . . . /visit our website at . . .

Chapter 4 Discussion

For this project, we developed reporting standards, quality standards and teaching and learning resources for realist evaluation. In addition, we provided methodological support and advice to realist evaluation projects, gave presentations to, and ran training workshops for, fellow realist evaluators and developed information and resources for patients and other lay participants in realist evaluation. Realist evaluation has now been used for close to 20 years in health services research and other disciplines, but there are still many evaluators, researchers and commissioners who were not trained in the approach and to whom it remains 'new'. It offers great promise in unpacking the black box of the many complex interventions or programmes that are increasingly being developed and used. We see this project as a start to the long journey of advancing the rigour of how realist evaluations are carried out and reported.

As relatively experienced users of realist evaluation, we had noted a number of common and recurrent challenges that face grant-awarding bodies, peer reviewers, evaluators and knowledge users. These centred on two closely related questions:

1. How can we judge if a realist evaluation, or a proposal for such a evaluation, is of high quality (including, for completed evaluations, how credible and robust findings are)?
2. How can we undertake such evaluations?

Our experience suggested that we could go a long way towards answering these questions by developing resources that help fellow evaluators to give due consideration to the theoretical and conceptual underpinnings of realist evaluations, outlined briefly below.

Realist evaluation is based on a realist philosophy of science as set out by Pawson and Tilley,[10] which permeates and informs its underlying epistemological assumptions, methodology and quality considerations. One of the most common misapplications we have noted is that evaluators have not always appreciated the underlying philosophical basis of realist evaluation or the implications of this for how the evaluations should be conducted. Instead, they have based their evaluations explicitly or implicitly on fundamentally different philosophical assumptions, commonly taking either the positivist notion that generalisable truths are best generated from controlled experiments, especially randomised trials, or a constructivist position that perceptions are all important. Another common misunderstanding is that realist evaluation is no more than a set of research or evaluation methods. For example, in our review of realist evaluations, we came across many instances where the evaluators appear to assume that realist evaluation is a form of qualitative research, whereas in practice it more commonly uses multiple methods. The appreciation that realist evaluation is an approach, or 'lens', through which to understand phenomena was often missing. In other words, many evaluators did not appreciate that realist evaluation uses a realist understanding of generative causation (as captured in the heuristic: context + mechanism = outcome) to:

- develop realist explanatory theories about phenomena through the use of data
- confirm, refute or refine ('test') realist explanatory theories using data.

A wide range of data-gathering methods may be used. No specific set of data-gathering methods must be used in a realist evaluation. Those chosen should, however, enable the collection of enough relevant data for realist theory development or 'testing'.

Even when a realist philosophy of science has been understood and adhered to in a realist evaluation, many evaluators – ourselves included – struggled with recurring conceptual and methodological issues. Mechanisms present a particular challenge in realist evaluations – how to define them, where to locate them, how to identify them and how to confirm, refute and refine them.[2,56] Realist evaluation trades on the use of realist theoretical explanations to make sense of the observed data. Realist evaluators commonly grapple with how to define a theory (e.g. what is the difference between a programme theory and a

middle-range theory?) and what level of abstraction is appropriate in different circumstances. On a more pragmatic level, those who seek to produce theory-driven evaluations of heterogeneous topic areas wrestle with a broad range of 'how to' issues: how to define the scope of the evaluation; how, and to what extent, to refine this scope as the evaluation unfolds; what should the evaluation design be; what data are needed; which data-gathering methods should be used; who to recruit and sample; how to collate, analyse and synthesise findings; and how to make recommendations that are academically defensible and useful to policy-makers and so on. We believe that the resources we have produced from this project will go some way to addressing the challenges we have highlighted above.

In undertaking this project, we were faced with one main dilemma that related to how best to allocate time and resources to the multiple work packages. For example, we could easily have spent more time on our literature review, but this may potentially have been at the expense of neglecting our Delphi panels, provision of support to review teams or development of resources and training materials. In retrospect, our project was very ambitious in its aims and, as such, we had to prioritise some aspects of the project above others. For example, we felt that it was more important to devote more time to (a) getting our Delphi process right so that we had a solid consensus on which to develop our quality and publication standards (and, to a lesser extent, our resources and training materials) and (b) the resources and training materials themselves. This meant that our literature review had to be rapid/truncated/abbreviated (see *Chapter 2, Details of literature search methods* and *Chapter 3, Literature search* for more details). Another example of prioritisation was in the breadth and depth of our resources and training materials. Entire textbooks could be written for these, but instead we chose to focus on common challenges. Our hope is that we have started the journey towards addressing some of the issues around the realist evaluation approach as set out by Pawson and Tilley – namely, how do you judge quality, how do you report it and how do you do X, Y or Z? We do, however, fully accept that more work is needed and, therefore, we have provided recommendations in *Chapter 4, Research recommendations and implications for practice*.

Changes to the protocol

Near the start of this project we published our project protocol.[15] During the course of the project we varied the following aspects of our protocol. One of the objectives of our project was to produce resources and training materials for lay participants, and those seeking to involve them, in realist evaluations. We have partially addressed this objective, in that some of the resources and training materials we have produced about aspects of realist evaluations are such that they are accessible to those with no to limited prior knowledge or experience of realist evaluations (see *Chapter 3, Developing, delivering and refining training materials for realist evaluation*). From our discussions within the project team, with other realist evaluators (e.g. in training workshops) and our project's Advisory Group, we came to the judgement that these materials would be accessible and helpful to lay participants who are more involved in realist evaluations, for example in their capacity as co-applicants or co-investigators in a project, and would help them understand more about realist evaluations.

However, for individuals who will be recruited into a realist evaluation, we had initially intended to develop draft template information sheets and consent forms that could be adapted for ethics and governance activity. On the issue of consent forms, again from discussion within the project team, other realist evaluators and our project's Advisory Group, we came to the judgement that there was too much diversity between organisations that grant ethics approval for us to be able to produce a generic template. Different organisations had such diverse processes and requirements for seeking ethics approval that we judged it best for those seeking such approvals to consult and adhere to their organisation's requirements. As such, we did not produce draft consent forms for realist evaluations. We were, however, able to develop a resource and training material entitled 'Realist evaluation and ethical considerations' (see *Table 7*) that will help to guide realist evaluators when developing information sheets and consent forms for recruiting participants into realist evaluations.

We had planned to deliver three 2-day 'realist evaluation' workshops and three 2-day 'training the trainers' workshops for a range of audiences. When we approached, or were approached by, those interested, we negotiated with them the logistics and content of each workshop. The preference from those interested was overwhelmingly for shorter workshops, so we ended up providing more workshops, but of a shorter duration, than we had planned. We were unable to find a mutually convenient time before the end of the project to organise any further 'training the trainer' workshops beyond the two 1-day workshops we provided to RDS East Midlands in June and October 2016.

Limitations

To develop the briefing materials for our Delphi panels, we undertook a literature review. This review has limitations that are likely to have introduced a number of biases and so, potentially, at least, they limit the inferences that can be made from the included evaluations and methodological pieces. For example, the search process for the review, despite being developed by an expert librarian, was not exhaustive. All the screening for inclusion and exclusion was undertaken by one screener and no quality checks were undertaken. Both processes may mean that we are likely to have missed some evaluations. However, given that the intent was to reach theoretical saturation, and that we retrieved many more evaluations than were necessary to achieve it, this is unlikely to have caused a significant problem to the other stages of the project.

An additional challenge we faced during the literature review was that, at the time of the project, there were no quality standards against which to judge the quality of realist evaluations; it was a function of this project to develop them. This was identified as a need in a range of methodological pieces we analysed as part of the review.[12,13] Therefore, we had to use the project team's collective judgement, informed by our experience in conducting and teaching realist evaluations, and the literature, to judge the quality of the realist evaluations included in the review. This is an important limitation of our review processes.

Once evaluations had been included, data extraction was undertaken by one researcher, and omissions in data extraction are likely to have occurred. However, all the included evaluations and the data extraction spreadsheet were circulated to all project team members, and so a degree of informal quality checking did occur.

Decision-making on what should be included in the Delphi panel's briefing materials was undertaken by the entire project team. We are aware that any item or topic included in the briefing materials was included as a result of our subjective interpretations, raising questions about reproducibility. However, the briefing documents we produced were not an end product in themselves, but the starting point for the Delphi panel to build a consensus. In addition, we deliberately asked Delphi panel members to enter into a discussion and suggest items for inclusion in the quality and reporting standards. We also provided the panel members with an end-of-round report and invited them to contact us should they have any concerns about the actions we had taken after we analysed their ratings. As such, we expected that changes would occur as we ran each round of the Delphi process, and thus we are confident that any omissions as a result of the review's limitations processes are unlikely to have a significant impact on the final reporting and quality standards. We accept that the review of the literature could have been more thorough (e.g. all evaluations analysed and more than one reviewer involved), but we made the judgement that the findings of the review contributed only part (albeit an important part) of the data to inform the Delphi Panel's briefing document. Other sources of data were the project team's expertise, that of the Delphi panel itself and data from the RAMESES JISCMail list. We felt that, in order for us to ensure that we delivered as much as possible on all the objectives of this project, the review needed to be truncated and our energies spent elsewhere. To provide transparency on what we have done, we have reported, in detail, all stages of the review itself and the rest of the project.

We recognise that there is much more to cover in terms of the breadth and depth of the training materials we have produced. Because realist evaluation is developing as an approach to evaluation, the 'wish list' we were able to elicit from our fellow evaluators who have used this approach was quite long. Given the time and resources allocated for this project, we elected to focus on providing sufficient depth in an accessible manner, rather than breadth on the issues that were the most challenging. With time, we hope to use the community of practice we have developed to address more, and more methodological, challenges.

As experience grows with the use of realist evaluation, it is very likely that many of the resources we have produced will need to be updated. We welcome and invite methodological development in realist evaluations. We expect that what we have produced should be gradually refined and updated as methodological developments take place with increasing use of realist evaluation. Thus, we view the reporting and quality standards and resources and training materials more as a starting point than as definitive resources that must not be altered in any way.

We are aware that realist evaluation is used to evaluate a wide range of topics and by evaluators from a broad range of disciplines and affiliations. The level of expertise of the users of our resources will also vary considerably, from novice to seasoned evaluators. These two aspects mean that some latitude is needed in the use of the resources we have produced. For example, not all the items in the reporting standards will be applicable for all evaluations. Or, when assessing the quality of an evaluation, there may be justifiable reasons for an evaluation to not meet some quality criteria. We have tried to anticipate the varied uses that realist evaluation might be put to by providing a degree of flexibility in our standards. For example, in our reporting standards, if adaptations are made to the evaluation design (as originally described), then evaluators are invited to provide an explanation for any such adaptations.

Finally, we were not able to produce detailed generic templates of draft information and consent forms for participant recruitment into realist evaluations. We have explained why this was the case above (see *Chapter 3*, *Develop, deliver and refine information and resources for patients and other lay participants in realist evaluation*, and *Chapter 4*, *Changes to the protocol*).

Research recommendations and implications for practice

Realist evaluation, despite having been first introduced in 1997, still has a great deal to do in terms of capacity building and methodological development. This is because it is only in the last few years that its popularity has grown as a form of theory-driven evaluation approach to make sense of complex interventions or programmes. This has created a situation in which some evaluators are using the realist evaluation approach for the first time on projects and some struggle with it.

Thus, capacity building is the priority for realist evaluation as an approach. Dedicated training courses, run by experienced realist evaluators, are needed. We anticipate that developing and running such courses will be easier with the key topic areas and consensus standards identified in this study, although even with such resources, some learners may still struggle to engage with the philosophical basis of the realist evaluation approach. Practical 'how to' resources and training materials were limited before, but this project has developed 15 of these to help fill this need. Course developers now have a reference point from which to build their training courses, and learners, a yard-stick against which to judge the quality of their work. The resources and training materials we have developed for this project are designed to be accessible to the novice but also to signpost more advanced learners to further resources. As such, they may be used as part of the basic building blocks of a 'curriculum' for realist evaluation courses.

As experience with realist evaluation grows and more evaluations are undertaken, new methodological insights are likely to occur. These need to be captured and analysed to determine if the quality and reporting standards we have produced continue to be fit for purpose or need to be updated. At present, no formal process exists to advance this agenda. Ideally, further funding might enable a project similar to

this one – that is, RAMESES III – to address the updating of the standards, although, because much groundwork has already been done, a more truncated project may suffice.

At present, those interested in realist evaluation (and review) might initially need to make small and gradual methodological gains, perhaps by embedding an element of methodological development with in their projects. Disseminating what they have learnt from undertaking their realist evaluations (or reviews) will be a key activity. At present, only ad hoc, informal help and support from more experienced realist researchers given to novices, the sharing of tips and templates used, and debate and discussion of contentious issues takes place. Some of this activity is happening on the RAMESES JISCMail list that we set up as part of the RAMESES Project. There is the potential that the RAMESES JISCMail list could be further developed and supported to serve as an avenue for advancing and disseminating methodological lessons in realist evaluation (and review). For example, it may be one way for realist evaluators to address the issue of generic templates of draft information and consent forms for participant recruitment into realist evaluations, an area that we did not fully address in our project. This might be through the sharing of particularly useful examples of these resources between evaluators and researchers.

Realist evaluators might want to consider learning from the example of organisations like the Cochrane Collaboration, in which motivated researchers have collaborated in a more organised way to systematically and gradually undertake methodological development. At present, many who contribute to, and support, the RAMESES JISCMail do so voluntarily, and with the end of this project all inputs to this list will be on a voluntary basis. Building some sort of future structure that is more sustainable is important. A potential benefit of being more organised is that priorities can be established on which methodological issues in realist evaluation (and review) need more attention, and duplication can be avoided. For example, the resources and training materials we developed are focused on what we were able to identify as the main issues that fellow evaluators found the most challenging to understand and/or execute. There are additional issues that we have not focused on or have only been able to address in passing. Further work is needed to develop resources for these, and other issues, as they arise. The resources and training materials we have designed are also intentionally brief. Developing new resources and building on the ones we developed, by drawing on the methodological lessons learnt from undertaking realist evaluations (or reviews), could potentially be a focus of a better-organised body of realist evaluators and researchers.

Finally, there is a dearth of research to demonstrate that quality and reporting standards necessarily change practice and improve the quality of research.[57,58] This will also be true for the standards we have produced and, therefore, research to demonstrate a change in practice and improvement in the quality of realist evaluations is needed at some point. There is also a counter-theory that such standards may constrain innovation in the development and application of realist methods, and testing this theory could form part of any evaluation of the standards.

Chapter 5 Conclusion

Although realist evaluation holds much promise for developing theory and informing policy in some of the health and other sectors' most pressing questions, misunderstandings and misapplications of it are common. To try to address these problems, we used a range of methods to gather the data needed to produce reporting and quality standards and resources and training materials. These included a literature review, Delphi panel, feedback from fellow realist evaluators, participants from training workshops and an e-mail list dedicated to realist research. In addition, we provided methodological support and advice to realist evaluation projects, gave presentations and ran training workshops for fellow realist evaluators and developed some resources for patients and other lay participants in realist evaluation. Undertaking this project was not without its challenges; our ambitious objectives meant that we had to shorten some aspects of the project (e.g. the literature review) and adapt others (such as workshop formats) to meet the needs of those we were training. We also found that we had over-anticipated the informational requirements of patients and other lay participants who might be involved in realist evaluation, thus narrowing our range of outputs for this group. We hope that what we have developed will be the start of an iterative journey of refinement and development of better resources for realist evaluations. An important priority for the realist evaluation approach is to build capacity. Acknowledging that the science of evaluation should never be static, the RAMESES II project seeks not to produce the last word on these issues but to capture current expertise and establish an agreed state of the science that future researchers will use and, no doubt, build on.

Acknowledgements

This project was funded by the National Institute for Health Research Health Services and Delivery Research programme. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Health Services and Delivery Research programme, NIHR, NHS or Department of Health. Trisha Greenhalgh's salary is part-funded by the Oxford Biomedical Research Centre, NIHR grant number BRC-1215-20008.

We would like to thank Nia Roberts from the Bodleian Library, University of Oxford, for her help with developing and running our literature search.

We are most grateful for the time, invaluable feedback and advice we received from our Project Advisory Group, who are all from the University of Leeds – Nick Emmel (chairperson), Jane Nixon and Rebecca Randall.

The following contributed to the patient and public panel: Maria Clark, Roger Ede, Matthew Le Croissette, Jo Lewis-Wood and Jeanne Nicholls. We wish to thank them for their help with this project.

We also wish to thank Ray Pawson and Nick Tilley for their advice, comments and suggestions when we were developing these reporting standards.

Finally, we are indebted to the Delphi Panel members, who freely and generously gave us their time and shared their wisdom:

Brad Astbury, University of Melbourne, Melbourne, VIC, Australia.

Paul Batalden, Dartmouth College, Hanover, NH, USA.

Annette Boaz, Kingston and St George's University, London, UK.

Rick Brown, Australian Institute of Criminology, Canberra, ACT, Australia.

Richard Byng, Plymouth University, Plymouth, UK.

Margaret Cargo, University of South Australia, Adelaide, SA, Australia.

Simon Carroll, University of Victoria, Victoria, BC, Canada.

Sonia Dalkin, Northumbria University, Newcastle, UK.

Helen Dickinson, University of Melbourne, Melbourne, VIC, Australia.

Dawn Dowding, Columbia University, New York, NY, USA.

Nick Emmel, University of Leeds, Leeds, UK.

Andrew Hawkins, ARTD Consultants, Sydney, NSW, Australia.

Gloria Laycock, University College London, London, UK.

Frans Leeuw, Maastricht University, Maastricht, the Netherlands.

Mhairi Mackenzie, University of Glasgow, Glasgow, UK.

Bruno Marchal, Institute of Tropical Medicine, Antwerp, Belgium.

Roshanak Mehdipanah, University of Michigan, Ann Arbor, MI, USA.

David Naylor, King's Fund, London, UK.

Jane Nixon, University of Leeds, Leeds, UK.

Peter O'Halloran, Queen's University Belfast, Belfast, UK.

Ray Pawson, University of Leeds, Leeds, UK.

Mark Pearson, Exeter University, Exeter, UK.

Rebecca Randell, University of Leeds, Leeds, UK.

Jo Rycroft-Malone, Bangor University, Bangor, UK.

Robert Street, Youth Justice Board, London, UK.

Nick Tilley, University College London, London, UK.

Robin Vincent, freelance consultant, Sheffield, UK.

Kieran Walshe, University of Manchester, Manchester, UK.

Emma Williams, Charles Darwin University, Darwin, NT, Australia.

All of the authors were also members of the Delphi panel.

Contributions of authors

Geoff Wong (Clinical Research Fellow, Realist Research Methodologist) carried out the literature review, analysed the findings from the review, produced the materials for the Delphi panel, analysed the results of the Delphi panel and developed the patient and lay materials.

Gill Westhorp (Professorial Research Fellow, Evaluator and Realist Research Methodologist), **Joanne Greenhalgh** (Associate Professor and Realist Research Methodologist), **Ana Manzano** (Lecturer in Health and Social Policy, Social Research Methodologist), **Justin Jagosh** (Senior Research Fellow and Realist Research Methodologist) and **Trisha Greenhalgh** (Professor of Primary Care and Social Scientist) analysed the findings from the review, produced the materials for the Delphi panel and analysed the results of the Delphi panel.

Joanne Greenhalgh assisted in the development of the patient and lay materials.

Gill Westhorp, **Joanne Greenhalgh**, **Ana Manzano** and **Justin Jagosh** developed and internally peer reviewed the resources and training materials.

Trisha Greenhalgh conceived the study and all the authors participated in its design.

All the authors provided realist evaluation support and training to various organisations during this study. All authors read and contributed critically to the contents of this report and approved the final manuscript.

Publication

Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. *BMC Med* 2016;**14**:96.

Data sharing statement

All non-personal data from this project can be obtained from the corresponding author.

References

1. Pawson R. *The Science of Evaluation: A Realist Manifesto*. London: Sage; 2013. https://doi.org/10.4135/9781473913820
2. Dalkin SM, Greenhalgh J, Jones D, Cunningham B, Lhussier M. What's in a mechanism? Development of a key concept in realist evaluation. *Implement Sci* 2015;**10**:49. https://doi.org/10.1186/s13012-015-0237-x
3. Greenhalgh T, Kristjansson E, Robinson V. Realist review to understand the efficacy of school feeding programmes. *BMJ* 2007;**335**:858–61. https://doi.org/10.1136/bmj.39359.525174.AD
4. Greenhalgh T, Humphrey C, Hughes J, Macfarlane F, Butler C, Pawson R. How do you modernize a health service? A realist evaluation of whole-scale transformation in London. *Milbank Q* 2009;**87**:391–416. https://doi.org/10.1111/j.1468-0009.2009.00562.x
5. Hoddinott P, Britten J, Pill R. Why do interventions work in some places and not others: a breastfeeding support group trial. *Soc Sci Med* 2010;**70**:769–78. https://doi.org/10.1016/j.socscimed.2009.10.067
6. Ranmuthugala G, Cunningham FC, Plumb JJ, Long J, Georgiou A, Westbrook JI, Braithwaite J. A realist evaluation of the role of communities of practice in changing healthcare practice. *Implement Sci* 2011;**6**:49. https://doi.org/10.1186/1748-5908-6-49
7. Cowe A, Cowe M, Goodman C, Kendal S, Mathie E, McNeilly E, *et al. RAPPORT: ReseArch with Patient and Public invOlvement: A Realist evaluaTion*. NIHR INVOLVE Conference, Leeds, 2012.
8. Randell R, Greenhalgh J, Hindmarch J, Dowding D, Jayne D, Pearman A, *et al.* Integration of robotic surgery into routine practice and impacts on communication, collaboration, and decision making: a realist process evaluation protocol. *Implement Sci* 2014;**9**:52. https://doi.org/10.1186/1748-5908-9-52
9. Manzano-Santaella A. A realistic evaluation of fines for hospital discharges: incorporating the history of programme evaluations in the analysis. *Evaluation* 2011;**17**:21–36. https://doi.org/10.1177/1356389010389913
10. Pawson R, Tilley N. *Realistic Evaluation*. London: Sage; 1997.
11. Pawson R. *Evidence-Based Policy: A Realist Perspective*. London: Sage; 2006. https://doi.org/10.4135/9781849209120
12. Marchal B, van Belle S, van Olmen J, Hoerée T, Kegels G. Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research. *Evaluation* 2012;**18**:192–212. https://doi.org/10.1177/1356389012442444
13. Pawson R, Manzano-Santaella A. A realist diagnostic workshop. *Evaluation* 2012;**18**:176–91. https://doi.org/10.1177/1356389012440912
14. Wong G, Greenhalgh T, Westhorp G, Pawson R. Development of methodological guidance, publication standards and training materials for realist and meta-narrative reviews: the RAMESES (Realist And Meta-narrative Evidence Syntheses – Evolving Standards) project. *Health Serv Deliv Res* 2014;**2**:30. https://doi.org/10.3310/hsdr02300
15. Greenhalgh T, Wong G, Jagosh J, Greenhalgh J, Manzano A, Westhorp G, Pawson R. Protocol – the RAMESES II study: developing guidance and reporting standards for realist evaluation. *BMJ Open* 2015;**5**:e008567. https://doi.org/10.1136/bmjopen-2015-008567

16. Booth A, Harris J, Crott E, Springett J, Campbell F, Wilkins E. Towards a methodology for cluster searching to provide conceptual and contextual 'richness' for systematic reviews of complex interventions: case study (CLUSTER). *BMC Med Res Methodol* 2013;**13**:118. https://doi.org/10.1186/1471-2288-13-118

17. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ* 2005;**331**:1064–5. https://doi.org/10.1136/bmj.38636.593461.68

18. Lefroy J, Hawarden A, Gay SP, McKinley RK, Cleland J. Grades in formative workplace-based assessment: a study of what works for whom and why. *Med Educ* 2015;**49**:307–20. https://doi.org/10.1111/medu.12659

19. Wye L, Lasseter G, Percival J, Duncan L, Simmonds B, Purdy S. What works in 'real life' to facilitate home deaths and fewer hospital admissions for those at end of life?: results from a realist evaluation of new palliative care services in two English counties. *BMC Palliat Care* 2014;**13**:37. https://doi.org/10.1186/1472-684X-13-37

20. Sorinola OO, Thistlethwaite J, Davies D, Peile E. Faculty development for educators: a realist evaluation. *Adv Health Sci Educ Theory Pract* 2015;**20**:385–401. https://doi.org/10.1007/s10459-014-9534-4

21. Sheaff R, Windle K, Wistow G, Ashby S, Beech R, Dickinson A, *et al.* Reducing emergency bed-days for older people? Network governance lessons from the 'Improving the Future for Older People' programme. *Soc Sci Med* 2014;**106**:59–66. https://doi.org/10.1016/j.socscimed.2014.01.033

22. Rushmer RK, Hunter DJ, Steven A. Using interactive workshops to prompt knowledge exchange: a realist evaluation of a knowledge to action initiative. *Public Health* 2014;**128**:552–60. https://doi.org/10.1016/j.puhe.2014.03.012

23. Riippa I, Kahilakoski O, Linna M, Hietala M. Can complex health interventions be evaluated using routine clinical and administrative data? A realist evaluation approach. *J Eval Clin Pract* 2014;**20**:1129–36. https://doi.org/10.1111/jep.12175

24. Rauf A, Anto B, Koffuor G, Buabeng K, Abdul-Kabir M. Introducing malaria rapid diagnostic tests (MRDTs) at registered retail pharmacies in Ghana: practitioners' perspective. *Br J Pharm Res* 2014;**4**:943–53. https://doi.org/10.9734/BJPR/2014/8910

25. Prashanth NS, Marchal B, Devadasan N, Kegels G, Criel B. Advancing the application of systems thinking in health: a realist evaluation of a capacity building programme for district managers in Tumkur, India. *Health Res Policy Syst* 2014;**12**:42. https://doi.org/10.1186/1478-4505-12-42

26. Parker J, Mawson S, Mountain G, Nasr N, Zheng H. Stroke patients' utilisation of extrinsic feedback from computer-based technology in the home: a multiple case study realistic evaluation. *BMC Med Inform Decis Mak* 2014;**14**:46. https://doi.org/10.1186/1472-6947-14-46

27. Ogrinc G, Ercolano E, Cohen ES, Harwood B, Baum K, van Aalst R, *et al.* Educational system factors that engage resident physicians in an integrated quality improvement curriculum at a VA hospital: a realist evaluation. *Acad Med* 2014;**89**:1380–5. https://doi.org/10.1097/ACM.0000000000000389

28. Noyes J, Lewis M, Bennett V, Widdas D, Brombley K. Realistic nurse-led policy implementation, optimization and evaluation: novel methodological exemplar. *J Adv Nurs* 2014;**70**:220–37. https://doi.org/10.1111/jan.12169

29. Nielsen K, Abildgaard J, Daniels K. Putting context into organizational intervention design: using tailored questionnaires to measure initiatives for worker well-being. *Human Relations* 2014;**67**:1537–60. https://doi.org/10.1177/0018726714525974

30. Meier K, Parker P, Freeth D. Mechanisms that support the assessment of interpersonal skills: a realistic evaluation of the interpersonal skills profile in pre-registration nursing students. *J Pract Teach Learn* 2014;**12**:6–24. https://doi.org/10.1921/7701240205

31. McConnell T, O'Halloran P, Donnelly M, Porter S. Factors affecting the successful implementation and sustainability of the Liverpool Care Pathway for dying patients: a realist evaluation. *BMJ Support Palliat Care* 2015;**5**:70–7. https://doi.org/10.1136/bmjspcare-2014-000723

32. Masterson-Algar P, Burton CR, Rycroft-Malone J, Sackley CM, Walker MF. Towards a programme theory for fidelity in the evaluation of complex interventions. *J Eval Clin Pract* 2014;**20**:445–52. https://doi.org/10.1111/jep.12174

33. Machin AI, Pearson P. Action learning sets in a nursing and midwifery practice learning context: a realistic evaluation. *Nurse Educ Pract* 2014;**14**:410–16. https://doi.org/10.1016/j.nepr.2014.01.007

34. Kwamie A, van Dijk H, Agyepong IA. Advancing the application of systems thinking in health: realist evaluation of the Leadership Development Programme for district manager decision-making in Ghana. *Health Res Policy Syst* 2014;**12**:29. https://doi.org/10.1186/1478-4505-12-29

35. Husted GR, Esbensen BA, Hommel E, Thorsteinsson B, Zoffmann V. Adolescents developing life skills for managing type 1 diabetes: a qualitative, realistic evaluation of a guided self-determination-youth intervention. *J Adv Nurs* 2014;**70**:2634–50. https://doi.org/10.1111/jan.12413

36. Higgins A, O'Halloran P, Porter S. The management of long-term sickness absence in large public sector healthcare organisations: a realist evaluation using mixed methods. *J Occup Rehabil* 2015;**25**:451–70. http://dx.doi.org/10.1007/s10926-014-9553-2.

37. Higgins A, Porter S, O'Halloran P. General practitioners' management of the long-term sick role. *Soc Sci Med* 2014;**107**:52–60. https://doi.org/10.1016/j.socscimed.2014.01.044

38. Hernández AR, Hurtig AK, Dahlblom K, San Sebastián M. More than a checklist: a realist evaluation of supervision of mid-level health workers in rural Guatemala. *BMC Health Serv Res* 2014;**14**:112. https://doi.org/10.1186/1472-6963-14-112

39. Harwood L, Clark AM. Dialysis modality decision-making for older adults with chronic kidney disease. *J Clin Nurs* 2014;**23**:3378–90. https://doi.org/10.1111/jocn.12582

40. Harris P, Haigh F, Thornell M, Molloy L, Sainsbury P. Housing, health and master planning: rules of engagement. *Public Health* 2014;**128**:354–9. https://doi.org/10.1016/j.puhe.2014.01.006

41. Evans D, Coad J, Cottrell K, Dalrymple J, Davies R, Donald C, *et al.* Public involvement in research: assessing impact through a realist evaluation. *Health Serv Deliv Res* 2014;**2**(36).

42. Eriksson C, Fredriksson I, Froding K, Geidne S, Pettersson C. Academic practice-policy partnerships for health promotion research: experiences from three research programs. *Scand J Pub Health* 2014;**42**:88–95. https://doi.org/10.1177/1403494814556926

43. Deschesnes M, Drouin N, Tessier C, Couturier Y. Schools' capacity to absorb a Healthy School approach into their operations: insights from a realist evaluation. *Health Educ* 2014;**114**:208–24. https://doi.org/10.1108/HE-10-2013-0054

44. Davey C, McShane K, Pulver A, McPherson C, Firestone M. A realist evaluation of a community-based addiction program for urban aboriginal people. *Alcohol Treat Q* 2014;**32**:33–57. https://doi.org/10.1080/07347324.2013.831641

45. Campbell C, Scott K, Mupambireyi Z, Nhamo M, Nyamukapa C, Skovdal M, Gregson S. Community resistance to a peer education programme in Zimbabwe. *BMC Health Serv Res* 2014;**14**:574. https://doi.org/10.1186/s12913-014-0574-5

46. Blanchet-Cohen N, Cook P. The transformative power of youth grants: sparks and ripples of change affecting marginalised youth and their communities. *Child Soc* 2014;**28**:392–403. https://doi.org/10.1111/j.1099-0860.2012.00473.x

47. Bartlett YK, Haywood A, Bentley CL, Parker J, Hawley MS, Mountain GA, Mawson S. The SMART personalised self-management system for congestive heart failure: results of a realist evaluation. *BMC Med Inform Decis Mak* 2014;**14**:109. https://doi.org/10.1186/s12911-014-0109-3

48. Ambrose LJ, Ker JS. Levels of reflective thinking and patient safety: an investigation of the mechanisms that impact on student learning in a single cohort over a 5 year curriculum. *Adv Health Sci Educ Theory Pract* 2014;**19**:297–310. https://doi.org/10.1007/s10459-013-9470-8

49. Allan H, Brearley S, Byng R, Christian S, Clayton J, Mackintosh M, *et al.* People and teams matter in organizational change: professionals' and managers' experiences of changing governance and incentives in primary care. *Health Serv Res* 2014;**49**:93–112. https://doi.org/10.1111/1475-6773.12084

50. Horrocks I, Budd L. Into the void: a realist evaluation of the eGovernment for You (EGOV4U) project. *Evaluation* 2015;**21**:47–64.

51. Taylor H. *Evaluating Criminal Justice Interventions in the Field of Domestic Violence: A Realist Approach*. PhD thesis. Birmingham: University of Birmingham; 2014.

52. Olsen K, Legg S, Hasle P. How to use programme theory to evaluate the effectiveness of schemes designed to improve the work environment in small businesses. *Work* 2012;**41**(Suppl. 1):5999–6006. https://doi.org/10.3233/WOR-2012-0036-5999

53. Kazi M, Frounfelker S, Bartone A, Buchanan P. Improving outcomes for a juvenile justice model court: a realist evaluation. *Juven Fam Court J* 2012;**63**:37–54. https://doi.org/10.1111/j.1755-6988.2012.01079.x

54. Hasle P, Kvorning L, Rasmussen C, Smith L, Flyvholm M. A model for design of tailored working environment intervention programmes for small enterprises. *Saf Health Work* 2012;**3**:181–91. https://doi.org/10.5491/SHAW.2012.3.3.181

55. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. *BMC Med* 2016;**14**:96. https://doi.org/10.1186/s12916-016-0643-1

56. Astbury B, Leeuw F. Unpacking black boxes: mechanisms and theory building in evaluation. *Am J Eval* 2010;**31**:363–81. https://doi.org/10.1177/1098214010371972

57. Cobo E, Cortés J, Ribera JM, Cardellach F, Selva-O'Callaghan A, Kostov B, *et al.* Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial. *BMJ* 2011;**343**:d6783. https://doi.org/10.1136/bmj.d6783

58. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;**339**:b2700. https://doi.org/10.1136/bmj.b2700

Appendix 1 Example of search terms use for MEDLINE (via OvidSP)

Search number	Search terms	References found
1	(realist adj5 (evaluat* or analys* or asses* or intervention? or stud*)).ti,ab.	121
2	(realist adj5 (approach* or understand* or theor* or methodolog* or framework*)).ti,ab.	188
3	(realistic adj (evaluat* or analys* or asses* or intervention? or stud*)).ti.	52
4	(realistic adj (approach* or understand* or theor* or methodolog* or framework*)).ti.	103
5	Program Evaluation/ and realist.mp.	33
6	realist.ti.	175
7	1 or 2 or 3 or 4 or 5 or 6	455

Appendix 2 RAMESES II Delphi Panel Briefing Document: developing reporting standards for realist evaluations

RAMESES II

Delphi Panel Briefing Document: developing reporting standards for realist evaluations

Trish Greenhalgh, Geoff Wong, Justin Jagosh, Joanne Greenhalgh, Ana Manzano, Gill Westhorp, Ray Pawson, Nick Tilley

What we would like you to do, how and when

The task is to produce consensus reporting standards for realist evaluation. You have agreed to be a member of our Delphi panel. A Delphi panel is a way of working towards consensus on a topic or question. It consists of a number of rounds. In a preliminary round, you will be asked to suggest topics which you would like to see covered (or statements you would like to see included). In each subsequent round (usually two more), you will be asked to do a task which involves *scoring* a draft set of statements. There will be a deadline for this, because we can't analyse the responses until everyone has replied.

After each *scoring* round, you will be sent your own scores *and* the average score for everyone in the group. If you find you are an 'outlier', you have two choices: amend your score (after reflecting on the statement and why you scored it as you did) – or stand your ground and argue your case to the group (they won't know how you scored the statement). Even if you scored a statement similarly to the group average, you may be swayed to change your score by arguments put subsequently.

Each statement is scored on two dimensions: [a] relevance (should we include this topic / theme at all?) and [b] content (should we word it like this?). High scores for relevance *and* content mean the statement will be included 'as is'. High scores for relevance but low scores for content means we need to word the statement differently (we'll ask for suggestions). Low scores for relevance mean the statement gets dropped. But when some panel members score a statement high and others score it low, we need a discussion. For references on the validity and methodology of the Delphi process, please ask us.

Here's what we'd like you to do:

- Pull out now if you've changed your mind (so you don't count as a 'withdrawal')
- For ROUND 1, please read this background paper (and, if you've got time, the study protocol and the other documents we have provided)
- Respond within one month to Geoff *only* by hitting the reply button with your suggestions.
- Wait while we analyse all the responses and build the draft statements
- Respond to the ROUND 2 email (expected early September 2015) within one month by looking at the statements and entering your scores for each (we'll give you a link to an online questionnaire)
- Wait again while we analyse the data and send you back your scores
- If needed and you want to, join in an email discussion on how we might amend the statements
- Repeat the last three steps for ROUND 3 (expected late November 2015)

This Delphi panel is part of the wider RAMESES II project, which has three workstreams: [a] produce quality and reporting standards for realist evaluations; [b] support teams undertaking realist evaluations; and [c] develop, deliver and evaluate training materials and information resources for realist evaluations. The RAMESES II

study protocol is appended (the protocol has bee accepted for publication in BMJ Open but it is in press so please do not circulate it)

Authorship policy

We want to acknowledge the input of everyone who contributes to RAMESES II. We propose two levels of authorship:

a. People who contribute materially and significantly to conceptualising the study, undertaking the research, analysing the data or writing up will be named as co-authors alongside us on publications. The format of the author list will be "Smith A, Jones B, Bloggs D on behalf of the RAMESES II group".
b. Members of the Delphi panel who do not fulfil the above criteria will be acknowledged in any publication in the following format: "We want to express our gratitude to the Delphi Panel members who so generously gave their time and input into the project:: Aaron Aardvark (Anthill University), Bob Boggs (Peat Institute) …etc to Zoe Zindel (Last Foundation)".

Please let us know if you are looking for a formal authorship role or if at any stage you believe you deserve to join the author list. We will also be alert to input from Delphi panel members above and beyond what is expected of an ordinary participant. It is quite possible that the RAMESES II publication standards will have a large number of authors and we are comfortable with that.

Whatever your level of input to this project, you won't get paid unless you were costed on the grant application. Nevertheless your input is greatly valued.

Briefing on realist evaluations

Background

Many of the problems confronting researchers today are complex. For example, in the health sector, much health need results from the effects of smoking, suboptimal diets (including obesity), alcohol excess, inactivity or adverse family circumstances (e.g. partner violence) – all of which in turn have multiple causes operating at both individual and societal level. Interventions or programmes designed to tackle such problems are themselves both complicated - having multiple, interconnected components delivered individually or targeted at communities or populations and complex - with non-linear causation and emergent properties. Their success depends both on individuals' responses and on the wider context in which people strive (or not) to live meaningful and healthy lives. What works in one family, or one organisation, one city or one country may not work in another. Similar complexity exists in many – or perhaps most – other domains in which evaluators work.

Similarly, the 'wicked problems' of contemporary health services research – how to improve quality and assure patient safety consistently across the service; how to meet rising need from a shrinking budget; and how to realise the potential of information and communication technologies (which often promise more than they deliver) – require complex delivery programmes with multiple, interlocked components that engage with the particularities of context. What works in hospital A may not work in hospital B. Again, similar complexities exist in all domains.

One increasingly popular approach to addressing these problems is realist evaluation. A form of theory-driven evaluation based on realist philosophy (1), it aims to advance understanding of why these complex interventions work, how, for whom, in what context, in what respects and to what extent – and also to explain the many situations in which a programme fails to achieve the anticipated benefit.

Realist evaluation assumes both that social systems and structures are 'real' (because they have real effects) and also that human actors respond differently to interventions in different circumstances. To understand how an intervention might generate different outcomes in different circumstances, realism introduces the concept of *mechanisms* – underlying changes in the reasoning and behaviour of participants that are triggered in particular contexts.

Methodological issues in realist evaluations

Realist evaluation was developed by Pawson and Tilley in the 1990s to address the question "what works for whom in what circumstances and how?" in complex social interventions (2). A realist approach assumes that programmes are 'theories incarnate'. That is, whenever a programme is implemented, it is testing a theory about what 'might cause change', even though that theory may not be explicit. One of the tasks of a realist evaluation is therefore to make the theories within a programme explicit, by developing clear hypotheses about how, and for whom, programmes might 'work'. The implementation of the programme, and the evaluation of it, then tests those hypotheses. This means collecting data, not just

about programme impacts or the processes of programme implementation, but about the specific aspects of programme context that might impact on programme outcomes, and about the specific mechanisms that might be creating change.

Pawson and Tilley also argue that a realist approach has particular implications for the design of an evaluation and the roles of participants. For example, rather than comparing changes for participants who have undertaken a programme with a group of people who have not (as is done in randomised controlled or quasi-experimental designs), a realist evaluation compares context-mechanism-outcome configurations within programmes. It may ask, for example, whether a programme works more or less well, and/or through different mechanisms, in different localities (and if so, how and why); or for different population groups (for example, men and women, or groups with differing socio-economic status). Further, they argue that different stakeholders will have different information and understandings about how programmes are supposed to work and whether they in fact do so and data collection should be tailored to reflect this. Data in a realist evaluation is used both to determine whether and for whom a program 'works', and to refute or refine theories about how and for whom the programme 'works'.

Summary of published examples of realist evaluations

With the help of a specialist informaticist/librarian (Nia Roberts), we identified a sample of 152 published papers which claimed to be realist evaluations. 137 of these were in health related topics and 15 in non-health topics. We did not analyse in detail all 152 realist evaluations, as the purpose of the exercise was to use these to help inform us as to; [a] what might be important to include in reporting standards; and [b] identify the methodological challenges evaluators faced when undertaking realist evaluations. The former helped us to develop the briefing materials for this Delphi panel and we will use the latter to inform quality standards for realist evaluations. We chose to work 'backwards', starting with analysis of the most recent (and thus current) published examples of realist evaluations (i.e. from 2015 'backwards'). After we had analysed a total of 37 realist evaluations (32 in health related topics from 2015 to 2014 and 5 in non-health from 2015 to 2012) we had reached thematic saturation. These were all examined in detail by Geoff Wong, and aspects of his analysis checked by the rest of the project team.

As expected, the 37 evaluations covered a range of complex topic areas (e.g. education, implementation of programmes, chronic disease management and criminal justice). Most were published after 2009, and we know of several more evaluations which are ongoing or in press. We considered that 7 of our sample of 37 were "true" realist evaluations. Our classification of these evaluations was based on our judgment of whether [a] a realist analysis (the application of realist logic) had been undertaken and [b] realist concepts (especially mechanisms) had been appropriately conceptualised. A further 7 of the evaluations appeared to "almost" meet these criteria – either having partially used a realist logic of analysis or having mis-conceptualised one or more realist concepts. 21 papers described as realist evaluations did not meet even these fairly loose criteria. It was unclear in 2 papers as to whether they were realist evaluations.

Preliminary thoughts on publication standards for realist evaluations

Our analysis of these published evaluations, along with our discussions with evaluation teams who are currently undertaking realist evaluations and from the discussions that have occurred in the RAMESES JISCMail, have surfaced the following issues and implications for the RAMESES II project. These are preliminary – we hope the Delphi panel members will add to and/or challenge them.

1. TERMINOLOGY. Key terms were misunderstood or used inconsistently by evaluators (especially 'mechanism', despite recurrent discussions and explanations in different sources – e.g. books, methodological pieces, RAMESES JISCMail and in training workshops).

   => *We need a glossary and set of definitions.*

2. PHILOSOPHICAL BASIS OF REALIST EVALUATION. The philosophical assumptions of realist evaluation (e.g. the form of realism set out by Pawson and Tilley) appear to be widely misunderstood or ignored. Misunderstanding or undervaluing the importance of the philosophical basis of realist evaluation and its implications appeared to lead to mis-application of the method.

   => We need to find ways of making the philosophy accessible and its implications clear.

3. CLASSIFICATION. Some evaluators did not appear to understand the fundamental differences between a realist evaluation and other approaches to evaluations. Two common observations we made were that realist evaluation was seen as a type of qualitative method or a means of combining qualitative and quantitative. In these cases, a realist logic of analysis was either not or partially used and/or the philosophical basis of realist evaluation misunderstood or ignored.

   => We need to include very clear criteria for classifying an evaluation as a 'realist evaluation' and an alert that the term is sometimes misused.

4. TITLE. Some but not all realist evaluations were described as such in the title.

   => We need to encourage authors to do this.

5. RATIONALE FOR USING REALIST EVALUATION. Some published realist evaluations clearly and in some detail explained; [a] what the purpose was of their evaluation; [b] why the approach was suitable for their topic area and; [c] the scope of their evaluation. In other cases, the rationale provided was brief and mentioned that it was because the intervention was "complex" or because they saw realist evaluation as a way to address 'how', 'for whom', 'in what context' and (to a lesser extent) 'to what extent' a programme or

intervention 'works', but without applying a realist logic of analysis, understanding and/or ignoring the philosophical basis of realist evaluation.

=> We need to encourage evaluators to clearly explain why realist evaluation is the appropriate approach for the purpose, topic area, focus and questions they seek to answer. We also need to highlight when realist evaluation might be UNsuitable.

6. METHODS. Some evaluators provided detailed descriptions of the processes they employed in their realist evaluation. In a minority of cases, it was possible to see how these processes had been operationalised in their evaluation. A common observation was that evaluators reported that they would apply a realist logic of analysis in their methods section, but then it was not evident in the publication that this had indeed been done. In some cases, though a realist logic of analysis had been applied, evaluators appear to have 'slipped out' of a realist approach when (for example) they assumed that a realist mechanism is the same thing as an intervention strategy. This suggests that some journal editors and peer reviewers are unable to judge whether the methods reported are being followed or not. Some evaluators described their evaluation to be 'based on' or a 'modified' realist evaluation but did not say how and why they modified it.

   => We need to include techniques for confirming that the methods reported were actually followed. We need to include the instruction that if evaluators modify the approach, they have to say how and why they modified it.

7. DATA COLLECTION METHODS. Many realist evaluations had used suitable data collection methods to provide data with which to test their programme theory (or theories) and support their knowledge claims. We did however notice that not all would collect the data needed to test theory programme theory and/or support their knowledge claims. For example, in some evaluations, a claim would be made that a programme has been successful but such a claim was only based on self-reported change and not corroborated by any other data gathered. Another observation we made was that data collection methods were rarely changed to collect additional data on specific aspects of a programme theory that required further testing. For example, once a semi-structured qualitative interview schedule had been developed it would not be changed. Reasons for this were unclear.

   => We need to encourage evaluators to collect an appropriate mix of data to develop and refine their realist programme theory. We also need to point out that changes in the nature of the data collected may be entirely justifiable in a realist evaluation and that if this was not done reason(s) are reported.

8. PROGRAMME THEORY. A number of realist evaluations did not either understand what a realist programme theory is and/or develop one. Often terms like "conceptual framework" or "model" were used instead of

programme theory. Only a minority of realist evaluations demonstrated they understood the purpose and value of a realist programme theory.

=> We need to help those using realist evaluations to understand the purpose and value of a realist programme theory. If a realist programme theory is not developed and refined, such a decision should be justified.

9. FINDINGS. Some review teams did not provide sufficient detail to support the inferences in their findings section. A particular common issue was that only some evaluations clearly 'labelled' their findings as a context, mechanism or outcome and/or provided detailed context-mechanism-outcome configurations (CMOCs). Many more provided tables with unconfigured contexts, mechanisms and outcomes. In some evaluations, the findings would have been more coherent and plausible if the relationships between their CMOCs and programme theory had been reported.

   => We need to include clear guidance on how we expect evaluators to present and justify their findings in a way that allows others to judge their coherence and plausibility.

10. CONCLUSIONS. Some but not all teams provided a clear line of reasoning linking findings to conclusions and recommendations.

    => We need to require conclusions should be 'traceable' back to detailed presentation of findings.

11. RECOMMENDATIONS. Few evaluations contained sufficient detail on the contextual influences on outcomes and the mechanisms involved. The explanations in realist evaluations are highly dependent on contextual influences. It follows that recommendations must be contingent (for example only under certain contexts will a particular mechanism be triggered to generate the desired outcome) rather than a list of "dos and don'ts".

    => We need to stipulate the recommendations in a realist evaluation should be consistent with a realist view of the world (i.e. recommendations need to be contingent rather than a list of "dos and don'ts").

Reference List

(1) Pawson R. The Science of Evaluation: A Realist Manifesto. London: Sage, 2013.

(2) Pawson R, Tilley N. Realistic evaluation. London: Sage, 1997.

Appendix 3 'Paper' version of round 2 online Delphi panel survey

RAMESES II Delphi - Round 2

Introduction

Thank you for continuing to help us with the RAMESES II project.

In Round 1 of our Delphi process, we had asked panel members for suggestions of Items to include in the RAMESES II reporting standards realist evaluations. What we hope to produce are reporting standards rather than detailed guidance on how to conduct a realist evaluation. Your comments related to how to conduct realist evaluations have however been captured for later use when we develop our training materials. We hope to make our standards relevant to evaluators, researchers, journal editors, peer-reviewers and funders.

We have collated all your responses and compiled a list of potential Items for inclusion in the 'RAMESES II reporting standards for realist evaluations'. In Round 2, we would be grateful if you would please rate each Item for:

- **Relevance (should we include an Item on this theme/topic at all?)**
- **Content (should we word this Item like this?)**

There will be a free text box for you to make comments on any aspect of an Item. To help you understand why an Item has been included we have also provided a brief explanation. We would also appreciate any comments you may have regarding the order the Items have been presented in.

This survey will take you between 15 to 30 minutes to complete.

You may at any time stop and return to where you left off by clicking on the unique web link you were sent inviting you to take part in this survey. You may also go back to previous items if you wish.

We would be most grateful if you would please try to complete the survey by 8th Novmeber 2015 at the latest.

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 2

Item 1: Title

Item 1: Title

In the title, identify the document as a Realist Evaluation.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Our background searching has shown that some realist evaluations are not flagged as such in the title and may also be inconsistently indexed, and hence are more difficult to locate. There are also some evaluations that use a different realist approach (e.g. such as critical realism). Researchers, policy and decision makers and other knowledge users may wish to be able to locate reports using these different realist approaches.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 2: Summary or Abstract

Item 2: Abstract

A summary of abstract should be as informative but brief as possible. At the very least a summary should contain information about the following aspects of a realist evaluation: purpose of the evaluation; setting and participants; description of the overall evaluation strategy; data collection methods used; key findings and; implications of findings. If the evaluation is published in a more formal way the publication outlet (e.g. journal) will often stipulate the format of the abstract. As far as possible taking account of journal-specific formatting and content requirements, the abstract should contain brief details of the study context, evaluation question(s) and/or objective(s); data gathering method(s) used, nature and number of participants, recruitment/sampling approach, data documentation processes, data analysis and synthesis processes; results; and conclusions/implications.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Apart from the title, an abstract is often the only source of information accessible to searchers unless the full paper is obtained. Many busy knowledge users will often not have the time to read an entire evaluation report or publication and only access the summary or abstract. The information in it must allow the reader to decide if the evaluation is a realist evaluation and relevant to their needs.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Introduction section

The following Items in this section are topics for consideration in the Introduction section for the RAMESES II publication standards for realist evaluations.

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 2

Item 3: Rationale for evaluation

Item 3: Rationale for evaluation

Explain why the evaluation was done and the implications of the purpose on the focus and broad design of the evaluation.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Evaluations are conducted for multiple purposes (e.g. to develop a programme theory/logic, assess the process of delivering a programme or the cost of a programme). The purpose has significant implications for the focus of work, the nature of questions, the choice of methodology and the design. In some commissioned evaluations a background section is often expected. Where this is the case, it should: [a] explain what is already known; [b] what the evaluators considered to be the 'knowledge gaps'; [c] why the evaluation was done and; [d] what the implications were of the purpose on the focus and broad design of the evaluation.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 4: Programme theory

Item 4: Programme theory

Describe the programme theory (or theories) that underpin the programme or initiative.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Realist evaluations set out to develop, test and refine realist programme theory (or theories). All programmes or initiatives will (implicitly or explicitly) have a programme theory or theories (which may or may not be realist in nature) and these should be articulated here. As an evaluation progresses, a programme theory that is not realist in nature will need to be developed, and refined so that it becomes a realist programme theory.

Programmes are theories incarnate. Within a realist evaluation, a programme theory (or theories) can serve many functions. One of its functions is to describe and explain (some of) how and why, in the 'real world', a programme 'works', for whom, to what extent and in which contexts. Other functions include focusing an evaluation, identifying questions, and determining what type of data need to be collected and from whom.

As the evaluation progresses, any initial programme theory should be iteratively developed, tested and refined. At the start of an evaluation, any initial programme theory may need additional development. Different processes can be used for developing programme theory in different circumstances, including literature review, programme documentation review, and interviews and/or focus groups with key informants. The processes used to develop the programme theory are usually different from those used later to refine it. The programme theory development processes need to be clearly reported as this may enable judgements to be made on its adequacy, coherence and plausibility. The processes used for programme theory development may be reported here or in Item 14 – Data analysis.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 5: Evaluation questions, objectives and focus

Item 5: Evaluation questions, objectives and focus

State the research question(s) and specify the objectives for the evaluation. Define and justify the scope of the evaluation – with particular reference to the roles played by the programme theory.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Realist evaluation questions contains some or all of the elements of 'What works, how, why, for whom, to what extent and in what circumstances, in what respect?' and applies realist logic to address the question (see Item 14 – Data analysis). Specifically, realist evaluation questions need to reflect the underlying purpose of realist evaluation – that is to explain (how and why) rather than only describe outcome patterns.

Because a particular evaluation will never be able to address all potential questions or issues, clarification of the scope of the evaluation has to take place. This important process may involve discussion and negotiation with (for example) context experts, funders and/or users. The processes used to establish purposes, scope, questions, and/or objectives should be described. The role of the programme theory in determining the scope of the evaluation should be clearly articulated.

In the real world, the programme being evaluated does not sit in a vacuum. Instead it is thrust into a messy world of pre-existing programmes, a complex policy environment, multiple stakeholders and so on. All of these may have a bearing on (for example) the research questions, focus and constraints of the evaluation. Provide information to the reader of the policy and other circumstances that may have influenced the purposes, scope, questions, and/or objectives of the evaluation.

Given the iterative nature of realist evaluation, if the purposes, scope, questions, objectives, programme theory and/or protocol have changed, it should either be reported here or in Item 17 – Main findings.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 6: Ethics

Item 6: Ethics

State if the realist evaluation has gained ethical approval from the relevant authorities. Provide enough detail to enable independent checks that the evaluation has been conducted in accordance with local regulatory requirements and professional standards. If ethical approval was not needed, explain why.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Realist evaluation is a form of primary research and will usually involve human participants. It is important that evaluations are conducted ethically with relevant and necessary attention to the well-being of the participants. Evaluators come from a range of different professional backgrounds and work in diverse fields. This means that different professional ethical standards and local ethics regulatory requirements are likely to apply. Evaluators should ensure that they aware of and comply with their professional obligations and local ethics requirements during the evaluation project.

Specifically, a challenge that realist evaluations may face is that as the evaluation evolves legitimate changes may need to be made to the methods used and participants recruited. Anticipating that such changes may be needed is important when seeking ethics approval. Flexibility may need to be built into the project and explained to those who provide ethics approvals.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Methods section

The following questions cover potential Items for inclusion in the Methods section of the RAMESES II reporting standards for realist evaluations

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 2

Item 7: Rationale for using realist evaluation

Item 7: Rationale for using realist evaluation

Explain why a realist evaluation approach was used.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Realist evaluation is a theory-driven approach that is firmly rooted in a realist philosophy of science. It places particular emphasis on understanding causation (in this case, understanding how programmes and policies generate outcomes) and how causal mechanisms are shaped and constrained by social, political, economic (and so on) context. This makes it particularly suitable for evaluations of certain topics and questions – for example, complex social programmes that involve human decisions and actions. It also makes realist evaluation less suitable than other evaluation approaches for certain topics and questions – for example those which seek primarily to determine the average effect size of a simpler intervention administered in a limited range of conditions. The most common limitation of published 'realist' evaluations is inadequate engagement with the philosophical principles of the realist approach and the implications these have, firstly, for understanding policies, programmes and initiatives and how they work, and secondly, for cumulating evidence and explanation.

Published evaluations demonstrate that some evaluators have deliberately adapted or been 'inspired' by the approach as first described by Pawson and Tilley. The description and rationale for any adaptations made or what aspects of the evaluations have been 'inspired' by realist evaluation should be provided. Such information will allow criticism, debate and counter criticism amongst evaluators and users on suitability of those adaptations for the particular purposes of the evaluation.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 8: Protocol or evaluation design

Item 8: Protocol or evaluation design

The final protocol or evaluation design (i.e. the account of what was planned) should be reproduced, at least in summary form, in the document which presents the main findings. If this is not done, the omission should be justified and a reference or link to the protocol or evaluation design given. It may also be appropriate to publish or make freely available (e.g. online on a website) the original protocol or evaluation design (e.g. as set out in the commissioned proposal or developed in the early stages of the evaluation).

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

The design for a realist evaluation may differ significantly from other evaluation approaches. As noted above (in Item 4 - Evaluation questions, objectives and focus), the evaluation question(s) and scope (and, by implication, many subsequent steps) of a realist evaluation may evolve over the course of the evaluation. An accessible summary of what was planned in the protocol or evaluation design, in what order, and why is essential for interpreting the evaluation. Comparing the original protocol or evaluation design with the final account of what was done may provide transparency on how the evaluation's processes have evolved in its bid to build understanding of policy, programme or initiative (i.e. the evaluand - that which is being evaluated, such as policies, programmes and initiatives).

Sometimes evaluations can involve a large number of steps and processes. Providing a diagram or figure of the overall structure of the evaluation may help to orient the reader.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 9: Setting(s) of the evaluation

Item 9: Setting(s) of the evaluation

Describe the setting in which the evaluation is taking place.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Explain and describe the setting(s) in which the policy, programme or initiative is being evaluated. These may (for example) include details about the policy landscape, stakeholders, service configuration and availability and funding and so on. Such information enables the reader to make sense of the relevant surrounding complexities and contexts at differing levels (e.g. meso and macro).

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 10: Nature of the programme being evaluated

Item 10: Nature of the programme being evaluated

Describe the nature of the programme being evaluated.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Realist evaluation may be used in a wide range of sectors (e.g. health, education, natural resource management, education, climate change), by a wide range of evaluators and on diverse evaluands. It should not be assumed that the reader will be familiar with the nature of the evaluand. The evaluand should be adequately described: what does it consist of, what is it supposed to achieve, and so on.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 11: Recruitment process and sampling strategy

Item 11: Recruitment process and sampling strategy

Describe and justify the recruitment process of the individuals who were approached to provide information to the realist evaluation that enables theory testing - how were they recruited, why and where?

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Specific kinds of information are required for realist evaluations. Data are used to develop and refine theory about how, for whom, and in what circumstances programs generate their outcomes. This implies that any process used to recruit individuals needs to find those who are able to provide information about contexts, mechanisms and outcomes, and that the sample needs to be structured appropriately to test the program theory. Describing the recruitment process enables judgements to be made about whether the process used is likely to recruit individuals who were likely to have the information needed to test the program theory.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 12: Data gathering approaches

Item 12: Data gathering approaches

Describe and justify the data gathering approaches used and how they were used to test programme theory.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Because of the nature of realist evaluation, a broad range of data may be required and a range of approaches may be necessary to collect it. Commonly, realist evaluations use more than one data gathering approach to gather data about contexts, mechanisms and outcomes and the relationships between them. Data collection tools and processes may need to be adapted to suit realist evaluation. The specific techniques used (e.g. realist interviewing) or adaptations made should be described in detail. Judgements can then be made on whether the approaches chosen, instruments used and adaptations made are capable of capturing the necessary data, in formats suitable for realist analysis.

For example, if interviews are used, the nature of the data collected must change from accessing respondents' interpretations of events, or 'meanings' (as is often done in constructivist approaches) to identifying causal processes (i.e. mechanisms) or relevant elements of context – which may or may not have anything to do with respondents' interpretations.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 13: Data documentation

Item 13: Data documentation

State and explain the rationale underlying the processes used to document the data collected in the evaluation.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

It is important that it is possible to judge if the processes used to document the data used in a realist evaluation are rational and applied consistently. For example, a realist evaluation might report that all data from interviews were audio taped and transcribed verbatim and numerical data were entered into a spreadsheet, or collected using particular software.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 14: Data analysis

Item 14: Data analysis

Describe in detail the analysis processes for all the data gathered. This section should include information on the constructs that are analysed, describe the analytic process, explain how the programme theory was developed, tested and refined and document and justify any changes in this process as the evaluation unfolded.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

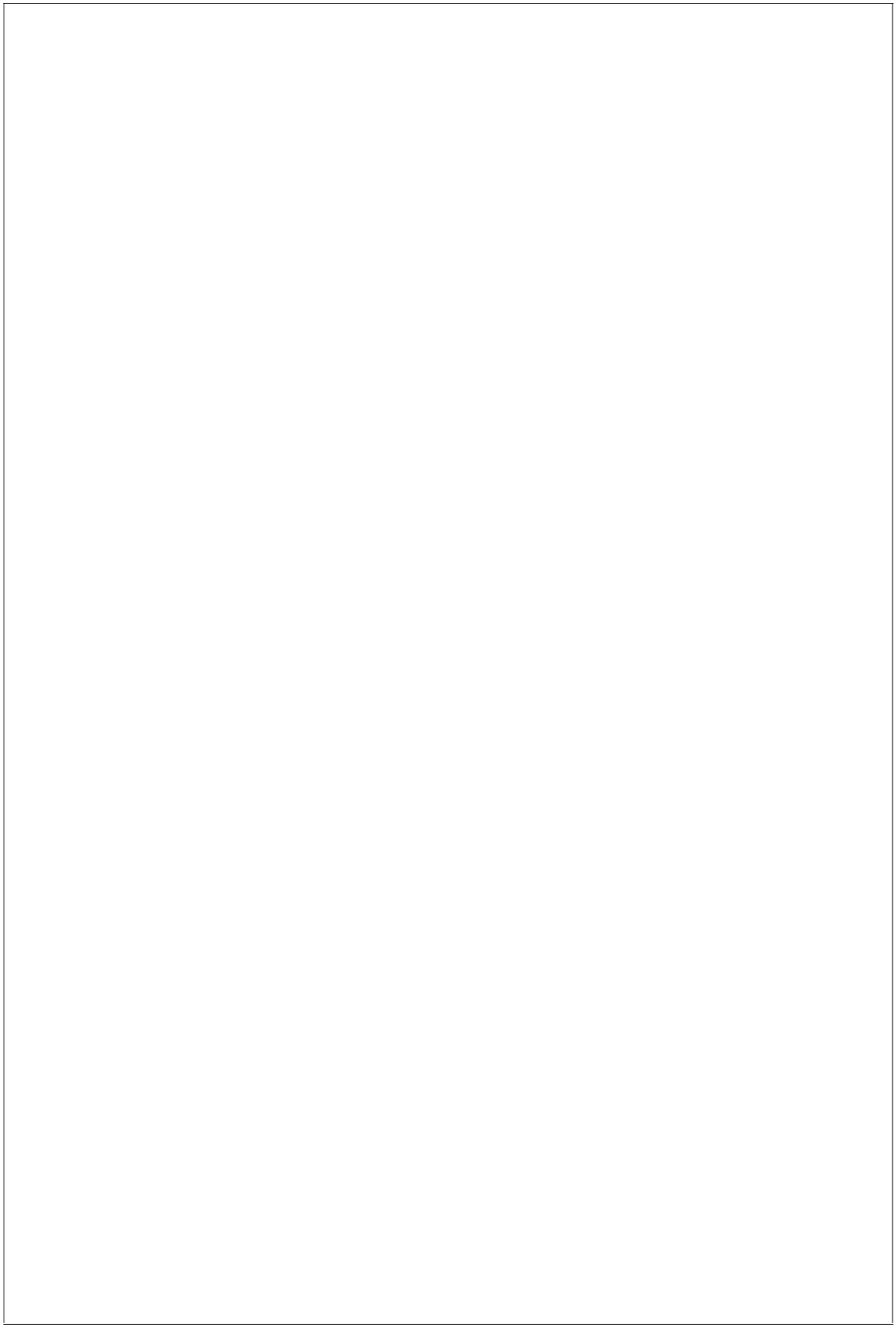
In a realist evaluation, the analysis process occurs iteratively. Realist evaluation is usually multi-method or mixed-method. The strategies used to analyse each method of data collection and integrate them should be explained. How these data are then used to develop, test and refine programme theory should also be explained. For example, if interviews were used, how were the interviews analysed? If a survey was also conducted, how was the survey analysed? In addition, how were these two sets of data integrated? The data analyses and may be sequential or in parallel – i.e. one set of data may be analysed first and then another or they might be analysed at the same time.

Specifically, at the centre of any realist analysis is the application of a realist philosophical 'lens' to data. A realist analysis of data seeks to analyse data using realist concepts. Specifically, realism adheres to a *generative* explanation for causation – i.e. an outcome (O) of interest was generated by relevant mechanism(s) (M) which was triggered by, or could only operate in, context (C). Within or across the data sources, recurrent patterns of outcomes and their associated mechanisms and contexts (CMO configurations) are likely to occur.

During analysis, the data gathered is used to iteratively develop and refine any initial programme theory (or theories) into one or more realist programme theories for the whole programme or initiative. This purpose has implications for the type of data that needs to be gathered – i.e. the data that needs to be gathered must be capable of being used for programme theory development, testing and refinement. These data must not only contain information that enables the evaluators to make inferences about whether something in the data is a context, mechanism or outcome, but also about the relationships between the contexts, mechanisms and outcomes. In other words the data gathered needs to contain information that enables evaluators to make inferences about the configuration of contexts, mechanisms and outcomes (i.e. Context-Mechanism-Outcome configurations or CMOCs). Other data gathered may have other functions in that they may be used to corroborate, refine or refute the assignment of a conceptual label to data (e.g. 'in this aspect of the analysis, this element is functioning as context) or inferences made about relationships within a CMOC. Data gathered will also be required to make inferences (and later corroborate or refute) the relationships between CMOCs – i.e. the location and interactions between CMOCs within a programme theory.

Ideally a description should be provided on who played which functions in the evaluation overall and if the data analysis processes evolved as the evaluation took shape.

Optional - Please comment on item, including wording and/or item order:



RAMESES II Delphi - Round 2

Item 15: Processes used to ensure quality

Item 15: Processes used to ensure quality

State the processes used to ensure quality during the evaluation.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Evaluations require a range of processes over a number of stages. For the findings of an evaluation to be credible, it is important for the reader to know that: a) the appropriate processes were used in an evaluation and; b) these were applied as described.

Items 11 to 14 above outline the guidance on the reporting of methodological processes. This item provides guidance on the reporting of the processes used to ensure that the evaluation was conducted to a high standard. We acknowledge that there is no universally accepted 'quality' standard against which all evaluations should be conducted. Evaluators should design their evaluations to meet three types of standards: the standards set by a relevant Evaluation Society; the standards required for high quality in the particular design (high quality ethnographic evaluation has to do different things well than does high quality survey-based evaluation); and the standards required to ensure that the evaluation is realist.

The processes used to design and implement the evaluation, and to ensure that high quality is maintained throughout the process and (where necessary) across all members of the evaluation team should be reported here or included within the relevant items above.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Results section

The following questions cover potential Items for inclusion in the Results section of the RAMESES II reporting standards for realist evaluations.

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 2

Item 16: Characteristics of participants

Item 16: Characteristics of participants

State the characteristics of the participants and describe the nature of the data they provided and how they contributed to programme theory testing.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

One important source of data in a realist evaluation comes from participants (e.g. clients, patients, service providers, policy makers and so on). To ensure transparency and to enable judgements about the probative value of the data provided, it is important that details are provided on who (anonymised if necessary) provided what type of data.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 17: Main findings

Item 17: Main findings

Present the key findings, including how they related to the programme theory and were used to refine it.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

The defining feature of a realist evaluation is that it is explanatory rather than simply descriptive, and that the explanation is consistent with a realist philosophy of science. That is, the realist programme theory is used to explain how and why patterns of outcomes occur for different groups or in different contexts. In other words, any such explanation should also include a description and explanation of the behaviour of key mechanisms under different contexts in generating outcomes.

Mechanisms are contingent: they are causal processes that have a tendency to occur in a particular set of conditions, but which do not always occur (because the circumstances have to be right for any particular mechanism to operate, and because many mechanisms can operate concurrently, sometimes cancelling each other out, sometimes contributing in different ways to a particular outcome).

At the start or in the early stages of a realist evaluation, the programme theory may be very rough and sketchy and not necessarily realist in nature. A major focus of any realist evaluation is to use the data to gradually refine the programme theory – gradually turning it into a realist programme theory. Ideally, in realist evaluations, this process of gradual refinement should be explicitly reported.

The findings in a realist evaluation necessarily include inferences about the links between context, mechanism and outcome and the explanation that accounts for this links. The explanation may draw on formal theory or program theory, or may simply comprise inferences drawn by the evaluators on the basis of the data available. It is important that where inferences are made this is clearly articulated. It is also important to include as much detailed data as possible to show how these inferences were arrived at. These data provided may (for example) support inferences about a factor operating as a context within a particular Context-Mechanism-Outcome configuration (CMOC). The theories developed within a realist evaluation often have to be built up from multiple inferences made on data collected form different sources. Providing the details of how and why these inferences were made may require that (where possible) additional files are provided, either online or at request from the evaluation team.

When reporting findings it is worth remembering that programme theories are usually 'middle-range' – that is, specific enough to generate propositions that can be tested against data but sufficiently abstract to be applicable to other contexts or other programmes using the same underlying theories.

Where relevant, disagreements or challenges faced by the evaluators in making any inferences should be reported here.

Transparency of the evaluation processes can be demonstrated, for example, by including such things as a detailed worked example, verbatim quotes from primary sources, or an exploration of disconfirming data (i.e. findings which appeared to refute the programme theory but which, on closer analysis, could be explained by other contextual influences).

When reporting context-mechanism-outcome configurations, evaluators should be clearly label what they have categorised as context, what as mechanism and what as outcome within the configuration.

Multiple sources of data might be needed to support an evaluative conclusion. It is sometimes appropriate to build the argument for a conclusion as an unfolding narrative in which successive data sources increase the strength of the inferences made and the conclusions drawn.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Discussion section

The following questions cover potential Items for inclusion in the Discussion section of the RAMESES II reporting standards for realist evaluations.

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 2

Item 18: Summary of findings

Item 18: Summary of findings

Summarise the main findings with attention to the evaluation questions, focus of the evaluation, and intended audience.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

In order to place the findings in the context of the wider literature and/or policy need, it is necessary to summarise briefly what has been found. This section should be succinct and balanced. Specifically for a realist evaluation, this section should summarise and explain the main findings and their relationships to the 'final' refine realist programme theory which emerged from the analysis. It should also highlight the strength of evidence for the main conclusions. This should be done with careful attention to the needs of the main users of the evaluation.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 19: Strengths, limitations and future research directions

Item 19: Strengths, limitations and future research directions

Discuss both the strengths of the evaluation and its limitations. These should include (but need not be limited to): [a] consideration of all the steps in the evaluation processes and; [b] comment on the adequacy and trustworthiness of the explanatory insights which emerged. In some evaluations, there may be an expectation to provide guidance on future research directions, programme implementation and/or programme design. The limitations identified may point to areas where further work is needed.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

Specifically for a realist evaluation, the strengths and limitations in relation to realist methodology and analysis should be included. Realist evaluations may be constrained by time and resources, by the skill mix and collective experience of the evaluators and/or by anticipated or unanticipated challenges in gathering the data or the data itself. These should be made explicit so that readers can interpret the findings in the light of them. Limitations imposed by any modifications made to the evaluation processes should also be reported and justified.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 20: Comparison with existing literature

Item 20: Comparison with existing literature

Where appropriate, compare and contrast the evaluation's findings with the existing literature on the same policy, programmes or initiatives.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Not all evaluations will be required to report on this item which is probably most relevant for peer-reviewed academic articles.

Comparing and contrasting the findings from an evaluation with the existing literature may help readers to put the findings into context. For example, this item might cover questions such as; how does this evaluation design compare to others (e.g. were they theory-driven?); what does this evaluation add, and which body of work in particular does it add to?; has this evaluation reached the same or different conclusion to previous evaluations?; and has it answered a question previously identified as important by leaders in the field?

Referring back to previous literature can be of great value in realist evaluations. Realist evaluations develop and refine realist programme theory (or theories) to explain observed outcome patterns. The focus on how mechanisms work (or don't) in different contexts potentially enables cumulative knowledge to be developed around families of policies and programmes or across initiatives in different sectors that rely on the same underlying mechanisms. Consequently, reporting for this item should focus on comparing and contrasting the behaviour of key mechanisms under different contexts.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 21: Conclusion and recommendations

Item 21: Conclusion and recommendations

List the main implications that are justified by the data. If appropriate, offer recommendations.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

A clear line of reasoning is needed to link the implications drawn from the findings with the findings themselves, as presented in the results section. If the evaluation is small or preliminary, or if the strength of evidence behind the inferences is weak, firm implications for practice and policy may be inappropriate.

If recommendations are given, these should be consistent with a realist approach. In particular, if recommendations are based on programme outcome(s), the recommendations themselves should take account of context. For example, if an evaluation found that a program worked for some people or in some contexts (as would be expected in a realist evaluation), it would be inappropriate to recommend that it be run everywhere for everyone. Similarly, recommendations for program improvement should be consistent with findings about how the program has been found to work (or not) – for example, to support the features of implementation that fire 'positive mechanisms' in particular contexts, or to redress features that prevent intended mechanisms from firing.

Optional - Please comment on item, including wording and/or item order:

RAMESES II Delphi - Round 2

Item 22: Funding

Item 22: Funding

Details should be provided for the funding source (if any) for the evaluation, the role played by the funder (if any) and any conflicts of interests of the evaluators.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	○	○	○	○	○	○	○
Content - (Item wording)	○	○	○	○	○	○	○

Explanation:

The source of funding for an evaluation and/or personal conflicts of interests may influence the evaluation questions, methods, data analysis, conclusions and/or recommendations. No evaluation is a 'view from nowhere', and readers will be better able to interpret the evaluation if they know why it was done and for which commissioner.

If an evaluation is published, the process for reporting funding and conflicts of interest as set out by the publication concerned should be followed.

Optional - Please comment on item, including wording and/or item order:

Appendix 4 'Paper' version of round 3 online Delphi panel survey

RAMESES II Delphi - Round 3

Introduction

Thank you for continuing to help us with the RAMESES II project.

In Round 2 of our Delphi process, we asked you to rate 22 potential items for the RAMESES II reporting standards realist evaluations. After analysing your ratings and comments and from discussions within the project team, only one item needs to be rated again.

In Round 3, we would be grateful if you would please rate Item 11 for:

- **Relevance (should we include an Item on this theme/topic at all?)**
- **Content (should we word this Item like this?)**

There will be a free text box for you to make comments on any aspect of the Item. To help you understand why the Item has been included we have also provided a brief explanation.

This survey will take you only a few minutes to complete.

We would be most grateful if you would please try to complete the survey by*17th January 2016* at the latest.

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 3

Methods section

The following question covers a potential Item for inclusion in the Methods section of the RAMESES II reporting standards for realist evaluations

Please click on the NEXT button below to proceed.

RAMESES II Delphi - Round 3

Item 11: Data collection methods

Item 11: Data collection methods

Describe and justify the data collection methods used - which ones were used, why and how they fed into developing, supporting, refuting or refining programme theory. Provide relevant details of the steps taken to enhance the trustworthiness/accuracy of data collection and documentation.

* Please rate this Item for:

	1 = Strongly Disagree	2	3	4	5	6	7 = Strongly Agree
Relevance - (Item inclusion)	◯	◯	◯	◯	◯	◯	◯
Content - (Item wording)	◯	◯	◯	◯	◯	◯	◯

Explanation:

Because of the nature of realist evaluation, a broad range of data may be required and a range of methods may be necessary to collect them. Data will be required for all of context, mechanism and outcome. Data collection methods should be adequate to capture intended and unintended outcomes, and the context-mechanism interactions that generated them. Where possible, 'objective' data about outcomes should be obtained. Where not possible, data about outcomes should be triangulated (at least using different sources, if not different types, of information).

Commonly, realist evaluations use more than one data method to gather data. Administrative and monitoring data for the programme or policy, existing data sets (e.g. census data, health systems data), photographs, videos or sound recordings, as well as data collected specifically for the evaluation may all be required. The only constraints are that the data should be relevant to the programme theory and to the purposes of and the questions for the evaluation.

Data collection tools and processes may need to be adapted to suit realist evaluation. The specific techniques used or adaptations made to instruments or processes should be described in detail. Judgements can then be made on whether the approaches chosen, instruments used and adaptations made are capable of capturing the necessary data, in formats that will be suitable for realist analysis.

For example, if interviews are used, the nature of the data collected must change from only accessing respondents' interpretations of events, or 'meanings' (as is often done in constructivist approaches) to identifying causal processes (i.e. mechanisms) or relevant elements of context – which may or may not have anything to do with respondents' interpretations.

Methods for recording data (for example, translation and transcription of qualitative data; choices between video or oral recording; and the structuring of quantitative data systems) are all theory driven. Explain the rationale for the methods used and their implications for data analysis.

It is important that it is possible to judge whether the processes used to collect and document the data used in a realist evaluation are rational and applied consistently. For example, a realist evaluation might report that all data from interviews were audio taped and transcribed verbatim and numerical data were entered into a spreadsheet, or collected using particular software.

Optional - Please comment on item, including wording and/or item order:

Appendix 5 Agenda and notes from public participant session

The RAMESES II Project:

Developing quality and reporting standards and training materials for realist evaluation

Date: 20th September 2016
Time: 2 pm to 3:30pm
Location: Meeting Room 2
Nuffield Department of Primary Care Health Sciences (NDPCHS)
Radcliffe Primary Care Building
Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG
(for directions please see: https://www.phc.ox.ac.uk/about/contact-us)

Session lead: Dr Geoff Wong, Clinical Research Fellow, NDPCHS
geoffrey.wong@phc.ox.ac.uk
+44 07973 818122

Agenda:

	Timing (minutes)	*Who's leading this*
Introductions	5	Geoff Wong
Outline purpose of the session	10	Geoff Wong
Refining the leaflet/document	60	Everyone
Ideas about other formats	10	Everyone
Summary and finish	5	Geoff Wong

Expenses and other claims process:
You are entitled to claim expenses for participating in the workshop and there is also a payment for your time.
1) Expenses and honorarium
You will be provided with an expense claim form on the day. You have to complete and sign this form and provide original receipts or tickets for public transport and parking so please come prepared to do that. If you can only provide such proof after the day please send your completed form with the receipts/tickets to me after the day. (Scanned versions are not acceptable – it must have a 'wet signature')

2) Payment
In order to make a payment to you for your time (honorarium) or indeed to pay your expenses, the University opts to make a BACS transfer so please come prepared with the following information:

- Name
- Email
- Address
- NI Number
- UK Bank sort code
- UK Bank account number
- Bank's address

If you have any queries please do not hesitate to contact me.
Notes from PPI meeting:
Location: as per agenda
Timings 14:00 50 13:30
Present:
Geoff Wong (Project PI)
Lynne Maddocks (Department PPI coordinator - Observer)

PPI participants:
Jean N
Roger E
Mathew LC
Jo LW
Maria C

Everyone introduced themselves
GW explained the background to the RAMESES II project and the purpose of the session - namely to produce generic text that could be adapted as needed by realist evaluators for use when recruiting participants.
Participants asked for clarifications (e.g. exact audience, purpose of document/text).

Participants read the 2 sides A4 documents we had drafted.

Feedback from them:
Clarification of when it is that people would need this information - I explained this was when (for example) an service was being evaluated (i.e. when doing research). Service development might not count as research and so consent would probably not be needed.

Text was "dense" and too detailed.

Sentences too long.

Quite technical language use - issue here is to bear in mind the average (low) reading age of the population. Advice was to use simpler words.

Glossary was useful, but definitions too long and most agreed that having it at the end of the document was probably not helpful as no one would bother to flick back and forth. Perhaps embed definitions (kept as simple as possible) into the main text.

Address potential participants directly (e.g. you / your) - rather impersonal at the moment.

Avoid some words as they might alarm some people - e.g. intervention / consent.

It probably does not matter to the person who is being recruited into a realist evaluation what exactly a realist evaluation is. In other words the detail of what a RE is or is not is not likely to matter to the potential participant. So much of the detail in the text is not needed. So the text should be short and kept to 1/2 a side of A4 or 1 side A5.
Suggested format:
- seek consent
- explain what is needed from the participant
- explain why we need their help
- explain why we are doing a RE and give examples. A discussion arose on whether or not participants actually care if they are taking part in a RE or not. I raised the issue that in a RE participants were more likely to be probed about their reasoning behind certain actions. Point was raised that this would not matter as long as it was done in a sensitive manner. In effect there was agreement that the research approach being used was not likely to be important to the participants.
- If people want to know more should direct them to a website.

Agreed actions:

GW would draft new shorter participant recruitment materials based on the feedback received form this meeting
Circulate to PPI members for their feedback.
Redraft as needed from their feedback.

NOTE FOR GW - thank PPI members and ask if it would be OK if we acknowledge their contribution by name in the PPI material and also for the project report.

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library