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 <u>how and why</u> 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.