Mixing and Matching: Using Qualitative Methods to Improve Quantitative Impact Evaluations (IEs) and Systematic Reviews (SRs) of Development Outcomes

Emmanuel Jimenez [1], Hugh Waddington[1],Neeta Goel[1], Audrey Prost[2], Andrew Pullin[3], Howard White[4], Shaon Lahiri[1] and Anmol Narain[1] with the assistance of Raag Bhatia[1]

[1] International Initiative for Impact Evaluation
[2] London School of Hygiene and Tropical Medicine
[3] Bangor University
[4]Campbell Collaboration

Suggested Citation: Jimenez E, Waddington H, Goel N, Prost A, Pullin A, White H, Lahiri S, Narain A, Bhatia R, 2018 Mixing and Matching: Using Qualitative Methods to Improve Quantitative Impact Evaluations (IEs) and Systematic Reviews (SRs) of Development Outcomes CEDIL Inception Paper 5: London

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by UKAID through DFID. The mission of the centre is to develop and promote new impact evaluation methods in international development.

Corresponding Author: Dr Emmanuel (Manny) Jimenez email:ejimenez@3ieimpact.org

Table of Contents

Abstract

Recent books about evaluations to study the quantitative impact of development programs and projects typically devote a chapter or two of the need to complement the analysis with other methods – specifically qualitative techniques. They often cite how qualitative techniques help explain the reason for positive or negative quantitative results. This is key if the one is to draw conclusions for accountability or for learning to improve future program design. Or they explain how qualitative work is critical to make sure that quantitative data are collected in the right way. Despite these textbook recommendations, there has been a wide range of experiences in how using both quantitative and qualitative methods have affected the overall quality of evaluations. In many cases, the qualitative analysis consists mostly of quotes to justify findings from the quantitative work. While this helps provide context, there is not much value-added beyond making an otherwise 'dry' quantitative presentation more interesting.

Some recent evaluations have begun to change this practice and have arguably improved the quality of impact evaluations in terms their relevance, the inferences that are drawn from them and their applicability to policy makers and programme implementers. This includes the use of innovative techniques to form the specific evaluative questions being asked and tested, to gather the right type of data and information on outcomes and intermediating variables, to explain findings and to disseminate them to the appropriate decision-makers. This paper will review this work. It will canvass a purposeful sample of experts from a variety of disciplines to gather the success stories, and where apparently well-planned approaches have failed to add the value expected of them. It will then draw lessons for future evaluations as a basis for guidance on use of mixed methods.

Section 1

Executive Summary

Recent evaluations have begun to use qualitative data in a manner that arguably helps improve the quality of these studies in terms of their relevance, the inferences that are drawn from them and their applicability to policy makers and programme implementers. These evaluations have used innovative techniques to form specific evaluative questions, to gather insight into the outcomes and intermediate variables, and to explain findings.

This paper reviews this work and identifies good practices to integrate qualitative methods into quantitative impact evaluations (IEs) and systematic reviews (SRs). Using recent literature on the characteristics of such practices, we developed two tools to assess the methodological rigour and mixed methods integration of 40 IEs and 7 SRs, drawing upon previous approaches. The tools, while designed specifically for each type of study, are closely related because the evidence generated by IEs flow naturally into SRs.

To capture a broad breadth of international development interventions while still making our task manageable, we assessed studies in seven sectors – agriculture, climate change, education, financial inclusion, health, infrastructure and governance. We drew upon systematic reviews and impact evaluation repositories in each of these sectors to identify mixed methods impact evaluations (MMIEs) and mixed methods systematic reviews (MMSRs).

Our main findings are that quantitative impact evaluations that successfully integrated qualitative methods were those that:

- Were rigorous in applying each method. Studies that scored highly on quantitative and qualitative rigour also tended to score highly on integration;
- Provided a clear rationale for the integration of methods. The top scoring studies made clear the rationale for integration at key stages of the evaluation, and often noted the value-added of doing so;
- Deployed multidisciplinary teams: Given epistemological, ontological and methodological differences in quantitative and qualitative methods, multidisciplinary teams can offer a substantive treatment to mixed methods and their synthesis. However, successful integration should be based on a common premise behind ideas, concepts and evaluation approach;
- Provided adequate documentation: A common element among our exemplar studies is the provision of adequate documentation, be it within a report, or through supplementary reports and/or appendices;

- Acknowledged limitations: Acknowledging the limits of integrating qualitative and quantitative findings facilitates a better understanding of the transferability of findings, and their implications in the policy domain.

Successful integration tended to improve mixed methods impact evaluations by:

- Collecting better data: At the level of data collection, integrating qualitative and quantitative lines of enquiry lies in the use of different methods of data collection, and how they inform study design and findings;
- Validating findings through integration: In several studies that score highly on integration, when qualitative and quantitative findings diverged, the authors discussed the reasons and implications for the differences, which informed their reported interpretation of findings;
- Contextualising quantitative results: In some cases, consolidating qualitative data obtained by different methods helped situate the findings in the local context. Mixed methods impact evaluations can be especially useful in fragile and conflict-affected settings which pose a particular challenge in terms of data collection from individuals;
- Contributing to forming policy recommendations: Successful integrations can inform policy recommendations by making contextually relevant policy recommendations. All of the exemplar studies report how mixed methods data influenced their policy recommendations.

In systematic reviews, there has been an evolution in approaches to incorporating qualitative evidence over time. Early cases typically used theory-based approaches to open up the intervention 'black box' and present outcomes along the causal chain. However, mixed-methods systematic reviews have been increasingly inclusive in incorporating qualitative evidence, including by undertaking additional searches for qualitative studies linked to the included quantitative studies or by conducting full searches for qualitative studies to answer specific review questions.

Our main findings from the assessment of mixed methods SRs are as follows:

- Mixed methods reviews can go beyond the 'sum of their parts' to provide holistic answers about development effectiveness. In some cases, the key contribution of integration was the identification of impacts that informed the approach to analysing the quantitative data, ensuring a comprehensive consideration of the evidence, even in areas where little or no quantitative evidence was found. In others, the contribution has been to provide evidence on the scalability of intervention. Integrated synthesis has also enabled reviews to provide evidence on unintended adverse outcomes for vulnerable groups.
- On conduct and reporting, the SRs we assessed specified answerable review questions and undertook searches over an appropriate time period. But they tended not to report deviations from the protocol (e.g. indicating whether analyses were pre-specified or undertaken post hoc), fully articulate study designs for review sub-questions, avoid bias through double-coding, or use

methods to avoid 'double-counting' evidence from dependent findings across quantitative and qualitative studies;
- On quantitative evidence appraisal and synthesis, the SRs tended to use appropriate methods including synthesis of effect sizes (i.e. avoiding vote counting) and reporting of heterogeneity but were more limited in exploring heterogeneity in findings.
- On the qualitative evidence side, reviews used appropriate methods to obtain evidence (usually through separate systematic searches) and synthesise it, but there were concerns about reporting of critical appraisal (including indicating which evidence was of higher confidence);

We conclude with the following recommendations:

- Create a repository of studies that integrate qualitative and quantitative methods successfully. This repository could highlight studies where the integration of methods has led to a better appreciation of how and why an intervention works (or does not work), or where policy uptake has been high. These studies could serve as exemplars and guides for researchers who want to expand their scope of work beyond quantitative methods or qualitative methods alone.
- Develop and report a common minimum understanding of what constitutes a "mixed methods impact evaluations and systematic review." Our findings suggest that there are a variety of definitions for "mixed methods." To provide a measure of consistency to funders and researchers, we suggest common minimum definitions for both MMIE and MMSR.
- Develop reporting guidelines for mixed methods impact evaluations and systematic reviews. This would ensure researchers are reporting key elements such as the rationale for integrating mixed methods, how the qualitative and quantitative data led to specific inferences or policy recommendations, the limitations of the integration, and how the researchers' personal views and opinions played into the process.
- Devote adequate time, monetary and human resources to designing mixed methods studies at the outset of impact evaluations and systematic reviews. Our findings suggest that there needs to be careful planning of the integration of methods, ideally at different stages of the evaluation or review. Mapping the qualitative and quantitative components to the theory of change/hypothesised causal chain of a programme or intervention is a good way to conceptualise how and when both lines of inquiry should be integrated. It is also crucial that teams adopt a transdisciplinary approach, transcending their disciplinary standpoints and working together within a common framework of ideas, concepts and evaluation approaches. This is especially important to avoid "tokenistic" uses of qualitative methods, and achieve value for money when undertaking mixed methods impact evaluations and systematic reviews.

- Explore innovations in combining qualitative and quantitative data. A number of recent mixed methods approaches have been employed to integrate qualitative and quantitative lines of enquiry. These include realist reviews, contribution analysis, qualitative component analysis, Bayesian reviews, and others. It is worth exploring these methods to understand their added value in serving particular evaluation questions or contexts. In this way, they can be added usefully to a methodological tool belt for researchers.

Section 2

Introduction

Recent books on quantitative evaluations of development policies typically devote a chapter or two to mixed methods, or the need to complement quantitative analysis with other methods – specifically qualitative techniques (see, for example, Gertler et al., 2016; White and Raitzer, 2017). They cite how qualitative techniques can help explain the reason for positive or negative quantitative results, or in exploring heterogeneity in outcomes for particular sub-groups and contexts. This is key if one is to draw conclusions for accountability or for learning to improve future programme design. These publications also explain how qualitative work is critical to making sure that quantitative data are collected in an informed manner, such as through the articulation of relevant causal pathways and choice of outcomes, including unintended ones.

Analysts, such as Bamberger (2015), have reported how the use of mixed methods improves the overall quality of evaluations, including equity and gender-focused evaluation. However, they also say that the qualitative data incorporated into many impact evaluations tend to be limited to narrative quotes that supplement findings from the quantitative work. Similarly, in systematic reviews, qualitative findings are often brought in only during the final discussion section. While this may help provide context, there may not be much value added beyond making an otherwise 'dry' quantitative presentation more contextual or interesting.

Some recent evaluations and reviews have begun to change this superficial use of qualitative data and arguably improved the quality of evaluations in terms of their relevance, the inferences that are drawn from them and their applicability to policy-makers and programme implementers. These evaluations used innovative techniques to form the specific evaluative questions being asked and tested, to gather the right type of data and information on outcomes and intermediate variables, to explain findings and to disseminate them to appropriate decision-makers.

This paper reviews this work and identifies good practices to integrate qualitative methods into quantitative impact evaluations (IEs) and systematic reviews (SRs).[1] Using

[1] We recognize that the primary focus on attributable evidence is a limitation of the scope of the paper which does not mean to suggest that quantitative analysis ranks ahead of qualitative techniques. .

recent literature on the characteristics of such practices, we developed two tools to identify such IEs and SRs, and then applied these tools to samples of studies from impact evaluation and systematic review databases. We combined the assessments of IEs and SRs into this paper due to the similarity in the objectives of mixed methods approaches in quantitative IEs and SRs of quantitative studies. Hence, we developed the two critical appraisal tools simultaneously. The similarities in ontology, epistemology and methodology in quantitatively-driven IEs and SRs means that our approach is sufficiently consistent, and we also recognise that appraisal of broader mixed-methods approaches are left for separate research. We also take this opportunity to compare how the application of mixed-methods techniques may differ in each type of study.

While this paper draws upon the literature on the integration of quantitative and qualitative methods, it largely focuses on how qualitative techniques and insights can add value to quantitative impact evaluations and systematic reviews. We do not review the pros and cons of quantitative and qualitative techniques in development evaluation, as there are many such analyses already (see, for example, Morra Imas and Rist, 2009). Nor do we discuss how evaluations that use primarily qualitative techniques would benefit from quantitative analysis. This includes an emerging literature on qualitative comparative analysis which attempts to make conclusions about attributable effects even when small sample sizes preclude the use of statistically valid comparative methods (Befani, 2016), and the Qualitative Impact Assessment Protocol which provides a guide for generating causal evidence from impact narratives (Bath Social & Development Research, n.d.), We recognize that our approach narrows our field of enquiry to those questions that are inherently amenable to quantitative analysis. However, we see this approach as a useful contribution to the literature on mixed methods impact evaluations (MMIEs) and mixed methods systematic reviews (MMSRs), given many examples of evaluations in which qualitative components appear to be 'subordinate' to quantitative components. Our approach is a step toward addressing this pervasive quantitative bent by exploring the complementarity of both methods in mixed methods impact evaluations and systematic reviews.

In Part One, we describe the development of a tool to assess the integration of qualitative methods into quantitatively-driven mixed methods impact evaluations and apply it to a sample of studies chosen from multiple IE databases (Sections II & III). In section IV, we discuss the implications of the findings, citing examples of successful integration. The complete tool for assessing the rigour and integration of mixed methods impact evaluations can be found in Appendix I. Figures for Part One are in Appendix II. A full list of scores for mixed methods IEs in our sample can be found in Appendix V.

In Part Two, we report a similar exercise for systematic reviews (SRs) and describe the assessment tool (section VI). In sections VII and VIII, we present summary findings, again using examples of successful integrations in SRs. The complete tool for assessing the rigour and integration of mixed methods systematic reviews can be found in Appendix III. Figures for Part Two can be found in Appendix IV. A full list of scores for mixed methods SRs in our sample can be found in Appendix VI.

In Part Three, we discuss the findings from the first two parts and present implications for future CEDIL workstreams.

Section 3

PART ONE: IMPACT EVALUATIONS

I. Identifying good practices for the integration of qualitative methods into quantitatively-driven impact evaluations

In order to identify examples of well-integrated IEs, determine their characteristics, and draw lessons from them, we developed a tool to assess rigour and integration practices for mixed methods IEs. In this section, we first clarify what we mean by quantitative and qualitative techniques in impact evaluations. We then describe an assessment tool to identify good practices and apply it to a sample of mixed methods IEs.

A. Definitions of quantitative and qualitative techniques in impact evaluations.

In this paper, we define quantitative impact evaluations as "analyses that measure the net change in outcomes for a particular group of people that can be attributed to a specific programme using the best methodology available, feasible and appropriate to the evaluation question that is being investigated and to the specific context" (3ie, n.d.b). These analyses measure an outcome as:

> "the difference in the indicator of interest (Y) with the intervention (Y1) and without the intervention (Y0). That is, impact = Y1 – Y0. An impact evaluation is a study which tackles the issue of attribution by identifying the counterfactual value of Y (Y0) in a rigorous manner" (White, 2010).

The main methodological issue is that programme participants may differ from non-participants. These differences may not be due to random variation but to purposeful selection, by either participants or programme designers. It is thus difficult to determine if simple observed changes in outcomes are due to the exposure to the programme or to the participants' inherent characteristics. To correct for these possible 'biases,' quantitative impact evaluations rely on experimental and quasi-experimental techniques. The former use randomisation to assign participation, as in Randomised Controlled Trials (RCTs); the latter use a range of statistical techniques that try to correct for the inherent biases (see Gertler et al. 2016 for a comprehensive introduction). All of these methods require the generation of data in sufficient quantity to assure statistical robustness.

Even when quantitative outcomes have been successfully measured, there often remain important questions about the conclusions that can be drawn from evaluations (White, 2009; Bamberger et al., 2010). Quantitative methods alone may be insufficient to understand contextual factors that may limit generalisability or unintended programme

impacts, which are important for decision-makers and implementers. In this regard, qualitative data collection and analyses can be especially useful.

What do qualitative methods include? Woolcock (2018) writes that: "Qualitative methods, such as those of mainstream anthropology, focus on understanding the intricate details of the processes and meanings associated with social interactions within and between particular groups. As such, qualitative methods (interviews, observations, textual analysis) tend to be associated with qualitative data (words, images); less concern is given to demonstrating whether emergent findings (e.g., from a single village) are 'representative' of the larger population from which they are drawn (e.g., a region or country) since such claims are rarely made or expected."

Qualitative data can be collected through existing documentary sources (e.g. journals), interviews (with individuals and/or in groups). Qualitative data can help understand context-specific meanings and processes that are not easily captured in quantitative surveys:

> "Qualitative methods are especially useful when the interventions to be evaluated increase in complexity (i.e., require many discretionary and face-to-face transactions, and are contentious), when the 'context' itself is highly variable (and perhaps volatile), when the quality and availability of existing data is poor, and when insights are sought on specific types of impacts on specific groups (e.g., the effectiveness of a project for ethnic minorities, informal firms or illegal immigrants, who may not be adequately represented in formal surveys). Qualitative methods can also be useful when evaluating small-N interventions such as regulatory reforms at the national level, or automation of procedures in one single agency" (Woolcock, 2018).

While recognising that there are many definitions of qualitative research, in this paper, we use the following:

> "Qualitative research is a situated activity that locates the observer in the world. It consists of a set of interpretive, material practices that make the world visible. These practices transform the world. They turn the world into a series of representations, including field notes, interviews, conversations, photographs, recordings, and memos to the self. At this level, qualitative research involves an interpretive naturalistic approach to the world. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them" (Denzin et al., 2018).[2]

[2] This definition is consistent with that of Creswell (2014) who defines 'qualitative research' as "a means for exploring and understanding the meaning individuals or groups ascribe to a social or human problem. The process of research involves emerging questions and procedures; collecting data in the participants' setting; analysing the data inductively; building from particulars to general themes; and making interpretations of the meanings of the data."

Quantitative and qualitative methods are not substitutes for each other. Rather, they must be viewed as complementary in enabling decision-makers to assess development programmes. This has long been recognized (Shaffer, 2013). A common approach to integrating qualitative data collection in impact evaluations involves using these data to triangulate quantitative results on effects or mechanisms described on a causal pathway, checking for mechanisms that are harder to capture through quantitative measurements, and documenting any unintended intervention consequences (Schulte-Mecklenbeck et al., 2011). Additionally, development outcomes often hinge on human behaviour, which is driven by a constellation of latent factors. Qualitative methods can help shape more meaningful surveys if they are integrated into the quantitative findings. In analysis, quantitative techniques that measure observed effects are less effective when one is trying to understand processes, including *why* these effects are observed.

Despite these methodological innovations, the use of qualitative methods to inform the conclusions drawn from quantitative impact evaluations is still uncommon. A recent systematic review of interventions to improve schooling in low- and middle-income countries concluded that:

> "We found that few qualitative research and process evaluations are being carried out as part of impact evaluations. High-quality qualitative research and process monitoring data are important to provide explanations for programme failures or attempts to replicate successful interventions" (Snilstveit et al. 2016).

While we acknowledge that not all mixed methods impact evaluations are, or should be, quantitatively driven, for the purposes of this paper, we consider a "mixed methods impact evaluation" as a quantitatively driven impact evaluation, as defined above, using one or more qualitative methods of data collection and/or analysis.

B. Identifying quantitative impact evaluations that have successfully integrated qualitative methods: our approach

What does it mean to have a successful integration of mixed methods techniques? There have been several interpretations of this (See for example, Maxwell et al., 2015). A useful summary is provided by White (2008), who identified three main ways to combine quantitative and qualitative approaches. The first is about integrating methodologies. For example, researchers can use quantitative survey data to identify which individuals/communities could be invited to take part in a qualitative study, and use results of the survey to inform the interview guide for the qualitative work. Conversely, researchers can use qualitative data to inform the stratification of a quantitative sample, the design and acceptability of the survey questionnaire (pp 4-5). The second type of integration involves "confirming/reinforcing, refuting, enriching, and explaining the findings of one approach with those of the other" (p.4). This includes verifying quantitative results through the qualitative approach, using qualitative work to identify issues or obtain information on variables not obtained by quantitative surveys, generating hypotheses from qualitative work to be tested through the quantitative approach, and using qualitative work to understand unanticipated results from quantitative data. A final and third type of integration involves merging the findings of

the two approaches into recommendations to improve policies and programmes. In this paper, we focus on all three types of integration.

In successful integration, quantitative and qualitative methods inform each other. For example, Maxwell and colleagues (2015) stress that beyond joint collection of data, the methods must be used to test conclusions, draw a comprehensive picture of the context in which evaluations are situated, and provide examples of how this is done in a number of sectors. Additionally, Rao and Woolcock (2003) call for "participatory econometrics" in developing a participatory approach in which qualitative work leads to the construction of a quantitative questionnaire. In this way, there are a number of approaches and rationales for the integration of mixed methods in IEs.

C. Developing and applying a tool to assess qualitative and quantitative practices

Our primary goal was to develop a tool to assess the successful integration of mixed methods in impact evaluations. However, reliable evidence is a key ingredient of successful impact evaluations (Masset et al., 2018), and assessing the rigour of individual evaluation components provides insights into the credibility of the evidence generated by mixed methods evaluations. Additionally, this allows us to explore whether studies that pass the rigour test also integrate mixed methods well. Therefore, our tool also assesses the rigour of qualitative and quantitative methods in mixed methods impact evaluations.

To assess rigour in quantitative methods, our tool largely draws upon the evaluation criteria from Miles and Huberman (1994), Bamberger and colleagues (2012), Langer (2017) and 3ie's risk of bias tool (Hombrados and Waddington, 2012). 3ie's risk of bias tool comprises indicators to assess the quality of attribution methods (confounding and sample selection bias); the extent of spillovers in comparison groups; outcome and analysis in reporting bias and other sources of bias (Hombrados and Waddington, 2012). The tool developed by Bamberger and colleagues (2012) evaluates quantitative and qualitative rigour largely through separate tools, while Langer (2017) includes questions specific to the integration of methods, such as the type of integration (sequential explanatory, sequential exploratory or convergent)[3] as well as the defensibility, credibility, rigour and reflexivity of research.

We built upon criteria used by mixed methods researchers such as confirmability, credibility, utilization and others, to focus on our scope of mixed methods impact evaluations. Because our scope was defined, it was possible to refine open-ended criteria to specific constituents. For example, we were able to add criteria on whether the causal chain is elaborated using a programme/logic model, whether limitations to mixed method integrations were clearly defined, etc. Our tool also included descriptive sections to explore and add nuance to our inferences; however, these sections were not scored. For example, although we included the integration type and the stage(s) at which qualitative evidence was incorporated, we did not score these sections to avoid

[3] Definitions of sequential explanatory, sequential exploratory and convergent designs can be found in Section III.

unfairly privileging any particular design or stage. Including these factors, however, provided valuable insights into how, where and when studies used mixed methods more comprehensively in successful evaluations.

This quantitative portion of our tool includes criteria such as objectivity (confirmability), internal design validity (reliability, dependability, credibility, and authenticity), statistical conclusion validity, construct validity, external validity and utilisation (Bamberger et al. 2011). This part (see Section A in appendix I) covers areas related to the clarity of methods and procedures (A1); different aspects related to how the study was conducted (A2); selection bias and confounding in RCTs (A3); addressing bias in quasi experimental designs (A4); post-intervention biases (A5 and A6); threats to external validity (A7); and reporting statistical power (A8).

We note that causal attribution is a complex activity driven by a number of factors, including but not limited to internal design validity, and external validity covering the themes of transferability and utilisation. The tool covers risks to all these themes across qualitative and quantitative domains, and scores studies based on what is reported in published or grey literature.

To assess rigour in qualitative methods, our tool also drew upon dimensions of internal validity, external validity, and utilisation (Bamberger et al. 2012), as well as the thematic framework and criteria developed by Miles and Huberman (1994). These criteria cover the domains of confirmability, credibility, transferability and utilisation. Transparency in the documentation of the analysis process, and the logical sequence leading to the interpretation of results are also key elements to the conduct of such studies. (Mmari et al., 2006). The tool contains questions covering the clarity of the study's methods and procedures (B1); the rigour of the qualitative study in conduct (B2); reporting how researcher biases affected the study (B3); methods used to address bias (B4)[4]; the transferability of findings to other contexts (B5); and an assessment of whether qualitative data situated the findings within the political, institutional, cultural or social context of the study (B6)[5].

An example of rigour in qualitative inquiry lies in the reporting of frameworks for qualitative collection, analysis and interpretation to account for how authors plan to

[4]Bias is commonly understood to be a concept drawn from the quantitative research paradigm, and incompatible with the philosophical underpinnings of qualitative enquiry (Creswell, 2014; Thorne et al., 2016; Davies and Dodd, 2002). Instead, qualitative researchers agree that concepts such as rigour and trustworthiness are more applicable to the subjective nature of qualitative research. Our tool was therefore developed with the understanding that the nature of qualitative approaches may be iterative and inductive, and evolve with changing contexts. It incorporates these concepts of rigour and trustworthiness by themes of appraisal espoused by Creswell (2014), Greene(1989), Miles and Huberman(1994), Pluye(2011), Langer(2017), Critical Appraisal Skills Program (2018) and more.

[5] Section B5 covers the description of the context and conditions under which phenomena of interest occur, and the scope and limitations of data presented to enable generalisation to other settings. The term 'thick descriptions' is typically used in ethnographies, and we erred on the side of caution by not privileging one method over the other in the scoring criteria.

explore qualitative research questions and interpret results transparently. This is true for grounded theorists as well; while it is acknowledged that theorists cannot "shop their disciplinary stores for preconceived concepts and dress their data in them" (Charmaz, 2000, p. 511), the use of relatively neutral questions for gathering and analysing data could provide valid participant perspectives, within which, the logic of the line of inquiry can be progressively traced from initial interview questions to the final ones (Elliott and Higgins, 2012). Another such factor influencing the validity of findings is the reporting of researcher biases and ideological preferences (Creswell, 2014; Denzin et al, 1998). In the words of Patti Lather (1993, p 697), validity is 'multiple, partial, endlessly deferred." This does not mean, however, that anything goes (Norris, 1997).

In order to address the integration of methods, we relied on appraisal frameworks put forth by authors in the field of mixed methods evaluations such as Greene, Miles, Huberman, Bamberger, Creswell and others. Greene's (1989) conceptual framework of triangulation, complementarity, development, initiation, and expansion, informed the appraisal criteria put forth by Miles and Huberman (1994), and the development of methodological criteria by many researchers since. Our integration tool represents an amalgamation of these approaches. This section covers the integration of a theory of change/programme or logical model into the evaluation (C1), the study design (C2), the interpretation of findings (C3), the limitations of integration (C4), the stage at which qualitative evidence was incorporated in study design (C5) and the type of mixed methods integration (C6).

The tool was reviewed by a few subject matter experts (mixed methods research) and underwent rigorous testing by independent reviewers and feedback from experts in the field prior to finalisation. The tool is summarised in Table 1 below, and the full tool is reported in Appendix I.

We used the tool to assess mixed methods impact evaluations in our sample (described in the next section). To extract data from studies, we read through all available project documents, counting them as one study (e.g. if one study had separate reports for the qualitative and quantitative components, we counted these as one study). Where necessary, we contacted the authors to request additional details or reports on the study. For our qualitative scoring, we did not use a pre-established list of themes (in fact, not all studies even reported these themes) used within the study, as our analysis focused on the factors reported in the study rather than the primary data. Each study was scored independently by two reviewers, and discrepancies (greater than three points) in unweighted scoring were discussed and resolved. We identified the top scoring studies first by their total integration score (Section C) and then sorted the studies by overall rigour (Sections A and B together). We consider the top twelve studies (scoring 4 points or more out a maximum of 6 points on Integration) as the ones that excelled at integrating mixed methods. We describe these in the next section.

Table 1. Summary of mixed methods impact evaluation appraisal tool

Section	Types of questions	Scoring
Preliminary information	Coder name Reference to IE report Sector focus Donor Classification of impact evaluation design (RCT/Quasi/Non-experiment)	N/A
Section A: Quantitative Aspect	A1.Clear description of the study's methods and procedures A2. Rigour of the quantitative study in conduct A3. Selection bias and confounding (if an RCT) A4. Selection bias and confounding (if a natural experiment/quasi experiment/non-experiment) A5. Post intervention biases (motivation of participants) A6. Post intervention biases (analysis and reporting) A7. Threats to construct and external validity A8. Reportage of statistical power	o Yes o No o Not Applicable
Section B: Qualitative Aspect	B1. Clear description of the study's methods and procedures B2. Rigour of the qualitative study in conduct B3. Reportage of assumptions, values, biases B4. Attempts to address biases B5. Transferability of results B6. (Unscored) Data situated within political, institutional, cultural or social context	o Yes o No
Section C: Integration of Mixed Methods	C1. Integration theory of change/programme or logic model explored through mixed methods C2. Integration of methods to inform study design C3. Integration of methods to inform the interpretation of findings	o Yes o No

C4. Limitations of integration

C5. (Unscored) Stage(s) at which qualitative evidence is incorporated into the study.

C6. (Unscored) Categorisation of the type of mixed methods study into sequential exploratory, sequential explanatory and/or convergent design.

I. Findings: Characteristics of IEs with Mixed Methods.

In this section, we describe our sample of 40 MMIEs, findings related to quantitative and qualitative rigour, and findings related to the integration of the two methods.

A. Sample summary

In our review of studies, we found that there we found that there was no standard definition of "mixed methods impact evaluation," and the integration of methods was conducted and reported in diverse ways. Therefore, our inclusion criteria comprised quantitatively driven impact evaluations (as defined in Section II of this paper) that reported the use of at least one qualitative method (as defined in Section II). There were no geographical nor date restrictions. Only English language studies were included.

Sectors: We purposively selected studies to represent international development sectors with a large impact evaluation evidence base. As noted in Figure 3, the sectoral distribution was roughly equal, with a higher proportion of studies from the education sector (30%), followed by financial inclusion (25%), governance (23%) and health (23%).

Sources: Studies were identified from impact evaluation repositories (DFID, 3ie, World Bank and JPAL) as well as systematic reviews. Our search was also influenced by recommendations from sectoral experts. Within the stated repositories, we selected all studies that matched our inclusion criteria and were in the identified sectors note above. Due to the limited number of studies matching these parameters, all studies were selected, and no sub-sampling was conducted. Where studies cited related publications or reports (such as standalone baseline reports or separate publications), these were considered to be components of the same study.

Funders*:* A quarter of the studies in our sample were supported by the UK's Department for International Development (DFID). Other funders included 3ie (20% of studies), USAID (13% of studies) and others (33% of studies).

Geographical distribution: Our sample studies spanned 20 countries, of which the majority (95%) were conducted in low and middle-income countries (L&MICs), and one high income country. Five of the studies were conducted in fragile and conflict-affected contexts (four in the Democratic Republic of Congo and one in Liberia), as defined by the World Bank[6].

Methodology: Studies in our sample (see Figure 1) comprised randomised controlled trials (62%), quasi-experimental designs (35%), and a combination of both (3%). Only 10% studies specified a qualitative research design (e.g. ethnography). A majority of studies (58%) employed multiple techniques (Figure 2). Focus groups discussions were the most commonly used technique (55%), followed by key informant interviews (48%) and in-depth interviews (35%). Other techniques included life histories, contribution analysis, and social mapping.

Sequence: Qualitative methods were mainly used after an intervention in order to explore the implementation and its impacts (73% of studies). Studies also reported

[6]The World Bank's harmonised list of fragile situations for 2018 is available at: http://pubdocs.worldbank.org/en/189701503418416651/FY18FCSLIST-Final-July-2017.pdf

incorporating qualitative methods during an intervention (28%), before an intervention is rolled out (23%) and after endline analysis (20%). Table 3 describes the stages at which qualitative evidence is incorporated into the studies.

Table 2. Stages at which qualitative evidence is incorporated into mixed methods IEs.

At what stage is the qualitative evidence incorporated into the study?	No. of studies	Percentage of studies
Pre-intervention	9	23
During intervention	11	28
Post-intervention	29	73
After end line analysis	8	20

Note: Some studies incorporated qualitative evidence at multiple stages in study design.

Classification: Using Greene's classification (1989), the majority of studies used a sequential explanatory design (68%), followed by a convergent design (38%). Fewer studies employed a sequential exploratory design (13%). These designs are not mutually exclusive, and 15% of studies used more than one design (Figure 10). In this paper, we do not privilege any classification as they can vary by study design and the evaluation questions being explored.

A Sequential explanatory design is used to obtain an in-depth understanding of intervention effects and unintended consequences. It takes the form of a quantitative component, followed by a qualitative component. For example, Bonilla and colleagues (2017), conducted in-depth interviews after the conclusion of endline data collection to evaluate an unconditional cash transfer program designed to empower women. The explicitly stated objectives of the interviews were to triangulate findings from the quantitative evaluation and to critique the use of decision-making indicators as proxies for women's empowerment in the quantitative surveys. Similarly, Evans and colleagues (2014) estimated the effects of a low-cost early stimulation and parenting education program in Mexico. This quasi-experimental study incorporated focus group discussions after endline analysis to explore quantitative findings.

A *sequential exploratory* design can be used to explore, develop and test an instrument (or taxonomy), or a conceptual framework/theoretical model. It takes the form of a qualitative component followed by a quantitative component. For instance, qualitative methods used during formative research can inform quantitative survey design, and/or intervention design. Yeager and colleagues (2002) used in-depth interviews to design quantitative survey tools as well as an intervention to promote the sanitary disposal of child faeces in Peruvian slums.

A *convergent design* can be used to examine the same phenomenon by using qualitative and quantitative evidence to answer the same research questions. It takes the form of qualitative and quantitative components conducted at the same time. For example, an impact evaluation of a teacher training program on gender norms (Chinen et al, 2016) included evaluation questions explored through quantitative methods (e.g. a quantitative teacher survey), and questions about fidelity and uptake answered through qualitative methods (e.g. semi-structured interviews). However, after endline analysis, both qualitative and quantitative methods were used to ascertain if attitudes had changed as a consequence of the intervention. Similarly, an impact evaluation of community-based conditional cash transfers in Tanzania used participatory qualitative techniques to provide complementary information on program impacts at the household level (Evans et al. 2014).

A few studies used combinations of mixed methods sequencing at differences phases of a study (Figure 10). For instance, Langford and Panter-Brick (2013), employed qualitative methods for two distinct purposes in the impact evaluation – developing an intervention through formative research, and understanding the implementation and consequences of the intervention with an embedded ethnography throughout the evaluation. In this example, the authors used a sequential exploratory design (formative study leading to the design of survey instruments and an intervention), followed by a convergent design (quantitative surveys and an embedded ethnography).

Our tool also assessed studies on research transparency practices. While there is broad agreement among both funders and researchers on principles of transparency and open data access, the use of pre-analysis plans or providing open access to data is still not common in the international development sector. Our sample is reflective of this, as the majority of studies did not report a study protocol or pre-analysis plan (73%). It is possible that teams did have pre-analysis plans or protocols, but these were not mentioned due to word limits or space constraints. However, the lack of an ex-ante study analysis protocol or pre-analysis plan represents a possible risk to the credibility of findings (Olken, 2015). Similarly, very few studies reported publicly available study data (10% and 3% respectively for quantitative and qualitative data).

D. Rigour

Our tool included criteria for assessing how mixed methods impact evaluations address various sources of bias in both quantitative and qualitative components. In the quantitative section, we assessed if studies defined research questions justified their choice of methods, and addressed bias. We found that most studies (81% of RCTs and 60% of quasi-experimental studies) presented balance tables to demonstrate comparability between treatment and control groups. Nearly half of the RCTs describe the process of randomisation used to allocate participants to treatment and control groups. Other sources of bias, where applicable, were addressed by fewer studies: 33% of studies addressed instances of bias due to non-adherence, 8% addressed recall bias, 18% addressed social desirability bias and 10% addressed Hawthorne effects. Figure 7 provides average scores for rigour in quantitative methods.

While methods to account for bias were generally well described for quantitative components of the impact evaluations, fewer studies demonstrated comparable

thoroughness with the qualitative components. For instance, only 20% of studies reported on the analytical framework[7] for qualitative data; in contrast, 80% of studies described a framework for quantitative analysis. Only 38% of the studies presented information on their qualitative sampling (such as the composition of focus group participants), compared with 90% of studies describing a rationale for quantitative sampling. Figure 8 provides average scores for rigour in qualitative methods.

Qualitative methods can involve interpreting the thoughts, perceptions and beliefs of participants, it is important for researchers to acknowledge how their own backgrounds and opinions might influence data collection and analysis at different stages of the evaluation (Bamberger et al., 2012; Qin, 2016). Doing so provides the reader with an understanding of the assumptions influencing the study decisions, and the logical sequence leading to the interpretation of results (Mmari et al., 2006). However, less than 5% of studies acknowledged how researchers' positionality which represents a departure from best practice in qualitative enquiry. We also note that social science researchers are not entirely 'removed' from the participants of their research, and the notion of reflexivity is emblematic of the dynamic processes within the researcher-participant relationship. In this regard, situating oneself as a researcher within this relationship also extends to quantitative research (See Ryan and Golden, 2006 for an example of reflexivity applied to quantitative research). For example, reflexive quantitative authors can describe how their personal views and backgrounds influence how survey questions and variable are framed). None of the included studies demonstrated researcher reflexivity for the quantitative components. As a measure of reflexivity for this review, we used the SR tool described in Section V to critically appraise our own use of mixed methods. The results are summarised in Box 3.

Similarly, there are several approaches to qualitative validation, such as triangulation, member checking and intercoder agreement (Creswell and Clark, 2011). However, only 20% studies specifically reported any form of validity checks for their qualitative findings. In summary, studies mostly do better on quantitative rather than qualitative rigour.

E. Integration

A key appraisal criterion to assess the quality of mixed methods integration was the presentation of specific inferences linked to both qualitative and quantitative data. A majority of studies (75%) provided separate data for qualitative and quantitative areas of enquiry and brought the two together to inform study findings. However, only half of the studies provided a clear rationale for the integration of qualitative and quantitative methods. Reporting the limitations of the integration is also important, as it helps to understand what it is useful for, and what is beyond its scope to answer. However, we observe only 13% of our sample-reporting limitations to the integration of methods. Figure 9 provides average scores for the integration of mixed methods.

[7]By analytical framework, we are referring to whether or not the study reported the themes, coding and analysis procedures.

About a quarter of the studies reported a divergence between qualitative and quantitative findings[8]. Almost half of the studies (43%) report how the integration of methods influenced their policy recommendations. Figure 5 summarises the integration of findings.

We identified 12 studies, which scored highly on the integration indicators, defined as scoring four or more out of six points[9]. These 12 studies included all quantitative designs and used a variety of qualitative techniques. Half were RCTs, and a majority (83%) used multiple qualitative techniques in their evaluation. The majority (92%) of the top scoring studies were situated in low- and middle income countries (L&MICs), with two being located in fragile and conflict-affected countries (the Democratic Republic of Congo). Nearly half of the top scoring studies were from the financial inclusion sector (42%), followed by education (33%), governance (17%) and health (8%). In the next section, we identify a few characteristics of these studies.

The average weighted quantitative score was 54 points higher (SD= 7.2) than the average weighted qualitative score. This is unsurprising and is in line with *a priori* assumptions that qualitative components would be less rigorously reported, and perhaps conducted, compared with the quantitative components. The average combined weighted quantitative and qualitative rigour score for the sample was 139 out of a possible 260 points. On overall rigour, the IEs in our sample were not very well reported according to the criteria in our tool. On integration, the IEs were fairly evenly spread with a standard deviation of 1.4, and an average score of 3 points out of a possible 6. The scoring summary is presented in Table 2 below, and full scores for all mixed methods IEs are presented in Appendix V.

[8] For an example of such divergences, refer to section IV, part B.

[9] Integration indicators cover six domains, which include the provision of logic or programme models explored through mixed methods, the use of mixed methods to inform components of study design, and to inform the interpretation of findings, as well as limitations to the integration of methods. For more information, refer to section C of the tool in the appendix, and part II, section C.

Table 3. Scoring summary for mixed methods IEs.

	Total quantitative rigour score (out of 130)	Total qualitative rigour score (out of 130)	Overall rigour score (out of 260)	Total integration score (out of 6)
Average	96	43	139	3
Minimum	60	0	98	0
Maximum	125	91	203	6
Standard Deviation	16.4	23.6	30.9	1.4

II. Discussion: Implications of the findings for the use of qualitative methods in quantitative impact evaluations.

In this section, we discuss the key characteristics of successful integration, and how it can improve quantitative IEs. Because numerical comparisons of the characteristics of studies, which score higher or lower on integration can only take us so far, we examine examples from the highest scoring studies in detail to provide additional information about successful integration.

A. What makes for successful integration of qualitative and Quantitative Methods?

Being rigorous in applying each method. In general, we found that studies which scored highly on quantitative and qualitative rigour also scored highly on integration. For instance, 13 out of 16 studies scoring above the median integration score of 3.0 also scored above the median overall rigour (quantitative + qualitative rigour) score of 133.5. When qualitative rigour was high, it was easier to discern how well a study had integrated qualitative and quantitative components. This is unsurprising given that our sample comprised quantitatively driven impact evaluations, many of which give little credence to the qualitative component.

However, our sample also included exceptions to this pattern. For instance, one study in the top 12 scored relatively low in both qualitative and quantitative rigour, but high on integration. In this evaluation, the authors used a combination of data from surveys, secondary data, interviews and focus group discussions. In the quantitative component, the authors did not address biases affecting participant behaviour, and in the qualitative component, the authors did not describe the scope, limitations and context within which the evaluation was situated. Similarly, participant perspectives were not well placed within personal contexts. Despite these methodological reporting omissions, the study integrated its qualitative and quantitative lines of enquiry well at different stages of the evaluation. For instance, they provided a clear description of how their quantitative data influenced qualitative data collection, present a clear rationale

for the integration, how mixed methods evidence led to specific inferences and policy recommendations, and the limitations of their integration (i.e. what the mixed methods integration can answer, and what it cannot). This represents a robust engagement with both strands of data at key stages of the evaluation.

In contrast, another evaluation scored highly on qualitative and quantitative rigour, but this did not translate into high scores for integration. Perhaps because of the complexity of evaluation components, the integration of the qualitative and quantitative lines of enquiry was less thoroughly reported as compared to the qualitative and quantitative components individually. The relationship, or level of influence between qualitative and quantitative lines of enquiry, was not described for the data collection process (either through the process evaluation or impact evaluation), though the different components all related to the same objectives of the study. Similarly, the authors did not make efforts to explain divergent findings between their different evaluation components, nor explain the limitations of their integration (i.e. what the integrated data answer, and what they cannot answer inherently to the methods used to obtain data). As a result, this study scored less highly on integration than it did on methodological rigour. This represents a missed opportunity to engage more thoroughly with a mix of different data strands, which may have informed practical recommendations for the development of this government programme. In this way, there was no standard formula for a successful mixed methods study with regard to quantitative or qualitative rigour. Our results seem to confirm the distinction between rigour and integration.

Approximately 23% of studies across all scoring strata had a programme/intervention theory of change or logical framework. Doing so provided clarity on how studies integrated qualitative and quantitative components. For instance, Nisbett and colleagues (2016), one of the top 12 scoring studies, mapped each input and output of the causal chain to the means through which they would be investigated, with clear indicators of the respective quantitative and qualitative methods to be used. We see this as best practice, as it demonstrates a clear link between mixed methods and the hypothesised causal chain of a programme or intervention.

Providing a clear rationale for integration: The presentation of a clear rationale for the integration of qualitative and quantitative methods was another characteristic the exemplar studies shared. Studies that scored low on integration often included focus group discussions or key informant interviews as part of their impact evaluation, without stating a clear rationale. A few studies did not present the sampling, study design nor description of the qualitative component. So, although they claimed that qualitative methods were used to corroborate quantitative findings, no analyses were presented to support this claim. In these instances, the authors often did not report qualitative results comprehensively, but only make a mention in passing that qualitative findings confirmed the quantitative findings. For example, a study evaluating the effectiveness of an educational intervention on school learning outcomes did not present the rationale, research questions or objectives for qualitative enquiry, and only made a passing reference (in the Conclusion section) that interviews were conducted to assess program uptake.

Conversely, the top scoring studies made clear the rationale for integration at key stages of the evaluation and often noted the value-added of doing so. For example, an evaluation of conditional community cash transfers in Tanzania explicitly outlined how qualitative and quantitative approaches complement each other as important characteristics of the evaluation design (Evans et al 2014). Other studies used qualitative techniques as part of formative work to develop an intervention or quantitative tools, while others only conducted qualitative data collection and analysis after they had completed their quantitative endline analysis. Further, all top twelve-studies reported both quantitative and qualitative findings substantively and also highlighted how these data contributed to specific inferences or recommendations.

Using multidisciplinary teams: To achieve fully integrated mixed methods research, putting together a multidisciplinary team, with each member working from a discipline-specific knowledge base may not be enough. Certainly, any poorly managed team has the potential to waste resources and engender conflict. Given epistemological and ontological differences between disciplines, and between the traditional 'camps' of quantitative and qualitative researchers, a multidisciplinary team may involve interdisciplinary conflict. In this regard, it is important to convene teams, in which "researchers work jointly using a shared conceptual framework drawing together disciplinary-specific theories, concepts, and approaches to address common problems" (Rosenfield, 1992). Using a shared framework with a delineation of boundaries that transcend individual disciplines can help bridge gaps, and lead to more robust, fully integrated mixed methods research. Among our top 12 studies, Nielsen et al. (2010) describe the composition of the evaluation team including the team's sectoral, quantitative, qualitative and mixed methods expertise. Similarly, other top scoring studies (American Institutes for Research, 2017; Bonilla et al, 2017; Chinen et al, 2016) were conducted by teams with expertise in quantitative and qualitative research methods and from different disciplines.

Providing adequate documentation: Confronted by space or word limit constraints (especially in journal articles), authors often neglect to report various details on methodology, transparency or other elements. However, from a reader's perspective, this presents an incomplete picture, as it is unclear why certain study decisions were made, or what the rationale for integrating methods might be. Hence, one common element among our exemplar-studies is the provision of adequate documentation. This could be within a report, or through supplementary reports and/or appendices. A study by Merttens and colleagues (2013) is notable in that the team produced four separate reports on different components of the same impact evaluation, each with multiple appendices providing a wealth of information on various aspects of the evaluation. However, this does not imply that journal articles, which are subject to word and page limits, are always incomplete. For example, one of the top 12 studies, Bonilla and colleagues (2017), describe an evaluation of the Government of Zambia's Child Grant Program in a journal article. Despite being subject to space limitations (as defined by the journal), the study provides the information needed to understand the context and the design of both qualitative and quantitative components, as well as their integration.

Acknowledging limitations: Acknowledging the limits of integrating qualitative and quantitative findings facilitates a better understanding of the transferability of findings,

and their implications in the policy. This involves explicating the 'boundaries' of utility for both qualitative and quantitative lines of enquiry in serving the evaluation questions or research objectives. Limitations to the integration of mixed methods are reported in 41% of the top 12 studies. For example, an evaluation of a cash transfer programme in Zambia acknowledges limitations to qualitative work, such as the absence of a 'before and after' panel of respondents and elaborates on limitations on the quantitative front as well, such as the need for larger sample sizes to detect program impacts (Bonilla et al. 2017). Another study assessing the impact of teacher training on gender norms in Northern Uganda refers to limitations in quantitative methods such as self-reported surveys, which may suffer from courtesy and social desirability bias (Chinen et al. 2016).

B. How has successful integration contributed to better IEs?

The value of a successful integration to evaluators lies in its ability to strengthen data collection, analysis, interpretation and policy recommendations. In addition to describing the value added by successful integrations, we also present two case illustrations in Boxes 1 and 2, which detail what high scoring studies did well, and how they did it.

Collecting better data: At the level of data collection, integrating qualitative and quantitative lines of enquiry lies in the use of different methods of data collection, and how they inform study design and findings. For example, a study evaluating the impact of humanitarian cash transfers used participatory techniques of data collection in conflict affected communities to identify of target beneficiaries, which, in the absence of qualitative data might not have led to nuanced findings in the Democratic Republic of Congo (American Institutes for Research, 2017). Another study (Nisbett et al. 2016) estimating the impact of nutrition and livelihood related interventions on the nutritional status of children explicitly makes mention of how qualitative and exploratory data collection complemented the purely quantitative portion of the study. The mixed methods approach afforded the authors a toolset to investigate processes of change and underlying causal mechanisms, as well as the contextual factors that helped explain how, why, and the conditions under which the intervention may have led to significant impacts.

Validating findings through integration: By virtue of combining two or more epistemologically different approaches, studies may have to deal with divergences in findings for the same phenomena studied, and 67% of the top scorers attempted to explain contradictory observations and/or findings, if applicable.

In several studies that score highly on integration, when qualitative and quantitative findings diverged, the authors discussed the reasons and implications for the differences, which informed their reported interpretation of findings (Evans et al, 2014; Chinen et al, 2016; Bonilla et al, 2017; American Institutes for Research, 2017). Beyond reporting and discussing divergences, higher scoring studies tend to acknowledge the epistemological differences between qualitative and quantitative lines of enquiry in order to specify the limits of the integration for generalisability. For instance, the study by the American Institutes for Research (2017) offered a discussion of the limits of their quantitative and qualitative data, acknowledging the inherent qualities of each and their roles in evaluations.

Divergent results between qualitative and quantitative components can be critical to understanding intervention/programme effects. This is because in such cases, addressing the divergence of study findings often results in more nuanced interpretations that might be afforded by using a single method alone. They also provide authors with a sense of the limitations of certain methods for addressing evaluation questions, and how these could be mitigated through the use of complementary methods. For instance, quantitative results suggested that a cash transfer program has no impact on local savings institutions (Haynes and Merttens, 2017). However, qualitative research suggested that beneficiaries actively participated in savings groups as a direct result of the transfer. In an attempt to reconcile these findings, the study team noted that the implementing agency promoted savings groups in the intervention areas among both beneficiaries and non-beneficiaries, thereby potentially undermining program impact.

Contextualising quantitative results: The use of qualitative methods can enhance the understanding of quantitative results by providing the context or background necessary to situate the findings. In cases where findings (across qualitative and quantitative methods) converge, they enhance confidence in study results. In other cases, they provide insights into the validity and limitations of quantitative findings. For instance, Langford and Panter-Brick (2013) use data obtained by ethnographic observations and in-depth interviews to understand the impact of a handwashing intervention on women. In contrast with the quantitative findings, which demonstrated a 40% reduction in child diarrhoea, the triangulated qualitative findings highlight an important negative impact of the intervention. The qualitative findings suggest that the ultra-poor in the sample were not only unable to take-up the intervention due to structural constraints and competing priorities, but also suffered social censure from those in the sample who participated in the intervention. Rather than assuming a net positive impact of the intervention, as suggested by quantitative data, the authors offer a nuanced critique of the intervention on health equity, highlighting areas where interventions can have large impacts but insufficient reach to improve lives of the most vulnerable segments of the population.

Mixed methods impact evaluations can be especially useful in fragile and conflict affected settings. These settings pose a particular challenge in terms of data collection from individuals. During a crisis for instance, respondents may not accurately remember specific pre-crisis conditions, making the measurement of human welfare impacts over time difficult (Puri et al., 2017). Additionally, self-reporting errors may be correlated with the severity and frequency of crisis conditions. In such settings, the importance of utilising multiple methods of quantitative and qualitative data collection become crucial to mitigate the inherent challenges to information bias in humanitarian settings.

In our sample, a study by the American Institutes for Research (2017) is particularly illustrative of the value of mixed methods in fragile contexts. The authors evaluated the impacts of a cash transfer program in the DRC, and one aspect their evaluation sought to capture was the program's impact on the social dynamics of the context, including the influence of the programme on community relationships, gender relations and decision-making. Utilising focus group discussions, in-depth interviews and key-

informant interviews allowed the authors to capture a deeper understanding of the perceptions of key stakeholders within the context that would have been afforded by quantitative methods alone. The authors' own rationale for doing so was to acknowledge that qualitative research is the "ideal methodology for investigating processes because of its descriptive and discursive nature" (p.36).

Similarly, in an evaluation of a programme to improve security in the DRC by Palladium (2015), the authors acknowledge that quantitative measures alone to assess subjective concepts such as perceptions of police and security are risky, as it may unintentionally raise the subjective expectations of police performance among the beneficiaries. In this way, any objective improvements in police capacity may fall short of raised subjective expectations of the populace. This represents a potential unintended consequence of a purely quantitative approach to assessing police performance, which the authors mitigated through the use of questions aimed at understanding changing attitudes and perceptions of police performance. Additionally, given that the program ended earlier than intended due to mitigating circumstances, the urgency to provide an in-depth understanding of the program's impacts became more severe, bolstering the utility of contribution analysis to consolidate evidence of impact from a variety of sources.

Additionally, integrating qualitative methods of enquiry also helped the authors to explore unintended consequences of the intervention. For instance, in the study by the American Institutes for Research (2017), qualitative data collection indicated instances in which a cash transfer to a household reinforced male power dynamics, as some husbands used the money for alcohol or prostitutes. However, such findings were not evident in the quantitative data, potentially because respondents felt more comfortable revealing personal details in in-depth interviews rather than in quantitative surveys in the presence of other household members. The urgency of understanding the appropriateness and reception of cash-based assistance in this context was particularly salient given the heterogeneity of household needs as a result of protracted conflict. Divergences between qualitative and quantitative findings in such settings become especially important as they have implications for immediate aid programming.

Contributing to forming policy recommendations: Successful integrations can inform policy recommendations by making contextually relevant policy recommendations. All of the exemplar studies report how mixed methods data influenced their policy recommendations. For example, an evaluation of a nutrition programme in Bangladesh links quantitative and qualitative findings to specific policy conclusions. Since the quantitative methods are not able to detect significant impacts of the intervention, and qualitative evidence points to specific nodes in the intervention pathway that did not lead positive outcomes, policy recommendations focus on resolving those issues (Nisbett et al. 2016).

Box 1: Illustration of a successful mixed methods impact evaluation in Zambia

Bonilla and colleagues (2017) conducted a mixed methods impact evaluation of the Government of Zambia's unconditional cash transfer programme on women's decision making and empowerment. In addition to performing well on methodological appraisal criteria for qualitative and quantitative techniques at the collection, analysis and interpretation stages, the study integrates methods in the following ways:

- The authors report how quantitative data influenced the collection of qualitative data, with a clear description of the rationale for the integration methods provided in the report.
- The presentation of qualitative and quantitative findings, and reportage of how qualitative and quantitative data led to integrated inferences or interpretations. Quantitatively, the authors find only modest increases in decision making among women in beneficiary households, compared with the control group. However, qualitatively, women in beneficiary households actually felt more empowered as a result of the intervention, which seemed to increase overall well-being for women. By considering both strands of data, the authors find that entrenched gender norms may have been the reason for increasing empowerment in only five out of nine domains.
- The authors report how mixed methods presents an opportunity for improving the measurement of empowerment, including women's decision-making indicators. In this regard, the authors' combined data led to them to question empirical measures for capturing the concept of empowerment through decision-making indicators.
- Given their mixed results, the authors explain how these diverging strands of evidence informed their policy recommendations. In particular, authors refer to the potential for unconditional cash transfers to improve the status of female beneficiaries, with the help of additional design components and the need for transformational change.

Box 2. An evaluation of teacher training program in Uganda.

Chinen and colleagues (2016) evaluate the short-term impact of teacher training and SMS messages on teachers' knowledge, attitudes and practices around gender equality and gender socialisation in the Karamoja region of Uganda. In addition to scoring highly on quantitative and qualitative rigour, the study integrates both lines of inquiry well in the following ways:

- The authors use a stakeholder-informed theory of change to map each evaluation question to elements of the causal chain, along with how they would be investigated (ie. through particular quantitative or qualitative methods).
- To provide further rationale for their integration, the authors explained that the qualitative component was used to provide in-depth analysis and insights into the impact of teacher training.
- The authors make explicit their intention to triangulate qualitative and quantitative data, offering a substantive and integrated treatment of both lines of enquiry. Results are structured according to outcome, in separate sections for qualitative and quantitative data. For instance, results on teacher knowledge present a graphical illustration of effect sizes for two different knowledge outcomes, followed by a narrative for qualitative findings from interviews and discussions.
- The interpretations of data are also complementary, in that the authors use both qualitative and quantitative components to make specific inferences. For example, in considering the programme's impact on teachers' sense of self-efficacy to solve the most pressing problems of the school, the authors present a table showing the programme's impact, along with standard error and effect size. They then present information from qualitative activities, demonstrating long-held embedded concepts of masculinity and femininity which seemed to be in opposition with the teacher training. In this way, the authors demonstrated the absence of a programme impact, and probed it further through qualitative data.
- In closing, the authors acknowledge the limits of their integration (i.e. what their integration can and cannot answer). They describe limitations in their sampling, in that they were not able to visit all treatment schools, and that the methods of administering the surveys was not uniform throughout the data collection process.

Section 4

PART TWO: SYSTEMATIC REVIEWS

V. Identifying good practices for the integration of qualitative methods into systematic reviews of effects

Systematic reviews locate, critically appraise and synthesise all the high-quality evidence from multiple contexts. According to the Campbell Collaboration, "[t]he purpose of a systematic review is to sum up the best available research on a specific question. This is done by synthesizing the results of several studies." We define systematic reviews as having "(1) a well-defined question for the review, (2) an explicit search strategy,

(3) clear criteria for the inclusion or exclusion of studies, (4) systematic coding and critical appraisal of included studies, and (5) a systematic synthesis of study findings" (White and Waddington, 2012).

Historically, systematic reviews have used either quantitative or qualitative methods to synthesise evidence on a particular question. Most systematic reviews are still limited to questions about *effects* drawing on impact evaluations (Higgins et al., 2011). Methods for synthesising qualitative evidence, such as meta-ethnography (Noblit and Hare, 1988) and thematic synthesis including of 'barriers and facilitators' (Thomas et al., 2004) are being increasingly used by international development researchers (e.g. Munro et al., 2007; Skalidou and Oya, 2018). Snilstveit, Oliver and Vojtkova (2012) provide a useful summary of approaches to the synthesis of qualitative evidence.

Petticrew and Roberts (2006, p.191) stated: "Qualitative work has tended to be allocated a rather small (or no) place in many systematic reviews in the past, but this is changing." Increasing numbers of reviews on international development topics incorporate mixed methods in 'parallel review modules' (Snilstveit, 2012). These studies may combine quantitative and qualitative evidence to answer different questions relating to *programme effectiveness* along the causal pathway from programme design, implementation and targeting through to intermediate outcomes and endpoint outcomes; to collect information on participant or practitioner views; to explore heterogeneity in findings by context, programme design and implementation; or to analyse adverse outcomes (Waddington, Masset and Jimenez, 2018).

In much the same way as primary studies, qualitative evidence is often incorporated into systematic reviews of effects by drawing on a programme theory or logic model, where the qualitative evidence is used to open up the intervention 'black box' or assess underlying assumptions along the causal pathway, or to articulate different causal mechanisms (White, 2009; White, 2018; Kneale et al., 2018; Waddington, Masset and Jimenez, 2018). More recent approaches to incorporating mixed methods include qualitative comparative analysis (Candy et al. 2011, 2013).

In this section, we describe the approach to developing a tool to appraise mixed methods systematic reviews, by which we mean systematic reviews of effects that

incorporate synthesis of qualitative evidence. We then present findings of piloting the tool on a number of systematic reviews that incorporated qualitative and quantitative evidence systematically.

A. Approach to assess incorporation of qualitative evidence in SRs of effects

There have been many calls for incorporating programme theory and use of a broad range of quantitative and qualitative evidence into systematic reviews over the years, including in the development of theory (Pawson, 2002, 2006; Oliver, 2008; van der Knapp et al., 2008) and in testing it (e.g. Davies, 2006; Anderson et al., 2011; Snilstveit, 2012; Waddington et al., 2012; Kneale, Thomas, and Harris, 2015; Maden et al., 2017). The importance of using theory and mixed methods to develop relevant review questions, structure evidence collection, and present findings are well-recognised in systematic reviewing including in international development (e.g. Anderson et al., 2011; Snilstveit, 2012; Waddington, Masset and Jimenez, 2018). Snilstveit (2012) convincingly argues that mixed methods reviews are necessary to answer the important 'what is the what?' question for a review's construct validity (external validity of the review to the issues at hand), by articulating at the very least the intervention design and process, which a 'bare bones' review drawing only on impact evaluation reports and journal articles is usually unable to do. Noyes et al. (2011) indicate reviews of effects can incorporate qualitative evidence in the following ways:

- "informing reviews" in order to define the question and ensure the review includes relevant outcomes (e.g. informing the theory of change);
- "enhancing reviews" by incorporating qualitative evidence contained in the impact evaluation reports, or "extending reviews" through additional searches for evidence from qualitative studies, in order to address questions about effects (e.g. exploring heterogeneity in findings); and
- "supplementing reviews" in order to answer different questions through "a stand-alone, but complementary, qualitative review to address [different] questions."

Snilstveit (2012) further defines enhancing and extending versus supplementing reviews as 'effectiveness plus' and 'effectiveness plus with parallel review modules,' respectively. According to Snilstveit (2012: 396), "the focus of [the effectiveness plus] approach is to provide a more detailed analysis of the causal chain. Questions that this approach aims to answer include: How does the intervention work? What are the key intervention components? Did outcomes vary by context? If so, how?" Including a separate but linked review module, enables the review to answer a broader range of questions: "For instance, beneficiaries' opinions, attitudes and knowledge [which] will be helpful in analysing the behaviour-change process that usually underpins and functioning social intervention, and can identify potential barriers and facilitators of intervention effectiveness" (p.396).

We present Noyes's (2011) and Snilstveit's (2012) categories (Table 2) alongside those of Julia Greene (1989; cited in Shaffer, 2013), whose categories are also mentioned above in Part One's discussion of impact evaluations (Section III.A). We note that reviews may use multiple designs hence these categories need not be applied mutually exclusively.

Furthermore, it appears common that reviews incorporating mixed methods plan to do so at the same time in the study protocol, even though the methods may be applied sequentially in study implementation itself. Therefore, we define the sequential designs as those where review modules are based on stand-alone protocols and simultaneous designs as those where the review protocol incorporates both quantitative and qualitative review components. In the following, we present examples of each approach for SRs.

- The purpose of *sequential explanatory design* in systematic reviews of effects is to explain quantitative results using qualitative findings. For example, the quantitative analysis is followed by qualitative evidence synthesis to explore participant views (e.g. Thomas et al., 2004), explain null findings (e.g. King et al., 2010), or assess the applicability of findings in particular contexts (e.g. Piza et al., 2016).
- Similarly, the purpose of *sequential exploratory design* in the context of a systematic review of effects is for the qualitative findings to inform the quantitative data collection instruments and/or approach to synthesis. For example, this can take place in the development of the theory of change or determination of relevant outcomes at the systematic review protocol stage. We are not aware of any examples of this exploratory phase being applied systematically in international development or elsewhere, although it appears quantitative reviews frequently draw on qualitative evidence non-systematically in informing the study design and theory of change.
- It seems more common for mixed-methods systematic reviews in international development to design the quantitative and qualitative review modules simultaneously in the protocol, even if the implementation of study modules is done sequentially. Simultaneous *triangulation designs* in a systematic review of effects might triangulate the evidence through further iterations of the theory of change at the analysis phase (e.g. Carr-Hill et al., 2018), or by undertaking meta-regression analysis drawing on moderators identified from qualitative evidence synthesis (e.g. Phillips et al., 2016).
- Whereas, the purpose of simultaneous *embedded/convergent design* in the context of a systematic review of effects is to better understand a specific issue found in the quantitative meta-analysis or causal chain synthesis, using a qualitative sub-study. For example, a systematic review of effects with sub-component examining participant views (e.g. Brody et al., 2016) or implementation processes (barriers and facilitators/enabler analysis) (de Buck et al., 2017). Some reviews contain embedded and triangulation designs (e.g. Carr-Hill et al., 2018; Waddington et al., 2014).

Table 4 Ways of incorporating qualitative evidence into quantitative SRs

Review stage	Greene (1989)	Noyes (2011)	Snilstveit (2012)
Formative	Sequential exploratory design	Informing reviews	'Effectiveness plus' (theory of change development)
Process	Sequential explanatory design	Enhancing reviews Extending reviews	'Effectiveness plus' (articulate intervention components and assess variation in outcomes by context)
Throughout	Convergent design including embedded and triangulation	Supplementing reviews	'Effectiveness plus with parallel review modules' to answer different questions about effectiveness (e.g. barriers and enablers synthesis)

B. Developing a tool to assess quality of mixed methods SRs

Drawing on these frameworks for mixed methods research, and existing systematic review critical appraisal tools and checklists (3ie, n.d.a; Shea et al., 2017; Langer, 2017), we developed a tool to categorise and critically appraise systematic reviews that incorporate quantitative and qualitative evidence to answer different questions about the effectiveness of development interventions (Appendix III).

After the preliminary information section, the appraisal tool is split into five sections, the first three providing critical appraisal questions for quantitative and qualitative aspects of the review, the fourth categorising the approach to integrating the quantitative and qualitative evidence, and a final section providing a summary and overview of the full critical appraisal (Table 3). The critical appraisal approach for identifying and including studies (Part 1 Section A) and methods used to analyse findings in quantitative analysis (Part 1 Section B) draw strongly on the tool used by 3ie (3ie, n.d.a) itself drawing on Lewin et al. (2009) as well as Shea et al. (2017) on reporting deviations from protocol. Methods used to include and analyse findings in qualitative analysis (Section C) and methods used to analyse the causal chain and reach conclusions (Section D) are new sections. Section C draws on questions from Sections A and B. Part A includes 8 questions, parts B and D include 7 questions and part C includes 6 questions.

All of these sections are scored (Yes, Partially, No, Can't tell, Not applicable) based on explicit decision rules. Part 2 of the tool, draws on the previous sections to describe the methods of integrating the quantitative and qualitative evidence (Section D) and provide an overall rating of the review's reliability and assessment of the use of mixed methods (Section E). Reviews are given an overall rating of overall confidence in conclusions about effects: low confidence reviews are those in which there are major methodological limitations; medium confidence reviews are those with important limitations, and high confidence reviews are those with minor limitations.

Two authors piloted the tool and revised questions and decision rules accordingly.

Table 5 Summary of mixed methods systematic reviews appraisal tool

Section	Types of questions	Scoring
Preliminary information	Reference (SR report, protocol, summary and journal article) Date of appraisal Coder name Date of the last search in review	N/A
Part 1: Critical appraisal of systematic review conduct and reporting		
Section A: Methods used to identify and include studies	Reference to a protocol Transparent reporting of review questions Specification of PICOS (separately by review question) Comprehensiveness of search Restrictions by time period/date Reporting of inclusion decisions Dependent findings	o Yes o Partially o No o Can't tell
Section B: Methods used to analyse the findings in quantitative analysis	Reporting of included studies Risk of bias (RoB) assessment Methods of analysis including effect size calculations Description of heterogeneity in findings Synthesis of findings Reporting of findings by RoB status Exploration of heterogeneity	o Yes o No o Partially o Not applicable (e.g. no included studies)
Section C: Methods used to include and analyse qualitative evidence	Searches for qualitative evidence Use of qualitative evidence Critical appraisal Reporting and analysis of findings Reporting of findings by critical appraisal	o Yes o No o Partially o Not applicable (e.g. no included studies)
Part 2: Integration of evidence and overall assessment		
Section D: Methods used to analyse the causal chain and reach conclusions	Use of programme theory Incorporation of qualitative evidence in review design Analysis of outcomes along causal chain	o Yes o No o Partially o Not applicable

	Use of qualitative evidence in causal chain analysis Use of qualitative in other aspects Integration of qual and quant Method of reaching implications Type of mixed methods study	
Section E: Overall assessment of the reliability of the review and the incorporation of qualitative evidence	Describe the confidence in the review findings Describe how qualitative evidence is used in the review	N/A

VI. Findings: Characteristics of SRs with Mixed Methods

A. Sample summary

We coded and appraised seven systematic reviews that incorporate quantitative and qualitative evidence covering agriculture, infrastructure, climate change and decentralised governance:

- King et al. (2010) on community development approaches (governance)
- Berg and Denison (2012) on interventions against female genital mutilation/cutting (FGM/C) (public health)
- Watson et al. (2012) on energy services (infrastructure)
- Pullin et al. (2013) on protected areas (climate change)
- Waddington et al. (2014) on farmer field schools (agriculture)
- Hulland et al. (2015) on water, sanitation and hygiene (WASH) programmes (infrastructure)
- De Buck (2017) on hygiene and sanitation behaviour change (infrastructure).

We chose these reviews to illustrate the breadth of mixed methods approaches that have been applied over the years, coinciding with the advent of relatively large funding programmes for systematic reviews in international development (notably DFID and 3ie) and supported by a range of bodies (3ie, the Campbell Collaboration, the Collaboration for Environmental Evidence, and the Evidence for Policy and Practice Information and Coordinating Centre, EPPI-centre). The reviews represent a range of review scopes, including reviews of single intervention types (e.g. farmer field schools, protected areas) and comparative reviews drawing on multiple interventions (community development, energy services, FGM/C, WASH). All reviews combined a quantitative component under a review question specifically asking about programme effects with a component drawing on qualitative evidence, whether that evidence was taken from included quantitative studies or searched for separately using explicit search and inclusion protocols. All reviews are limited to programmes implemented predominantly in, and evidence collected from, low- and middle-income countries (L&MICs). We would usually expect

reviews of single interventions to be able to conduct qualitative evidence synthesis (or integrate that synthesis with quantitative evidence) at a greater depth of analysis, than comparative reviews of multiple interventions, hence conduct more comprehensive simultaneous designs. We discuss this point below. It is difficult to draw firm conclusions for a small sample.

We coded 'yes', 'partially' and 'no' responses as 1, 0.5 and 0, respectively. Percentages of totals for parts A through D are given in Appendix IV Figure 11. Overall, all included reviews were assessed as being of 'medium' or 'high confidence' in informed decisions, and no reviews were assessed as being of 'low confidence'.

One study used a sequential explanatory design (King et al., 2010). A second study used a simultaneous convergent design but also used a sequential exploratory design element for determining quantitative outcome categories for synthesis (Pullin et al. 2013). The remaining studies used simultaneous designs (Watson et al., 2012; Waddington et al., 2014; Hulland et al., 2015; de Buck et al., 2017). Integration of qualitative and quantitative evidence was usually based on theory of change framework and in some cases also incorporated further iterations to the theory of change or triangulation through additional quantitative analysis drawing on evidence from the qualitative synthesis (Waddington et al., 2014).

B. Methodological rigour of quantitative and qualitative analyses

All reviews passed basic systematic review conduct and reporting factors including use of systematic searches of published and grey literature, explicit inclusion criteria specifying eligible populations, interventions, comparisons or contexts, outcomes and study designs (PICOS), critical appraisal of included evidence and synthesis of effects, and most satisfy internal quality assurance standards (double coding). The included reviews incorporate evidence to answer different questions along the causal chain including questions about intervention design or implementation (e.g. King et al. 2010; Waddington et al. 2014; de Buck et al. 2017), and intermediate and endpoint outcomes (e.g. participant or practitioner views).

Appendix IV Figures 13 to 16 provide average scores across reviews for general conduct and reporting (part A), quantitative synthesis (part B), qualitative synthesis (part C) and integration (part D). As we might expect, given that incorporation of mixed methods is a fairly new phenomenon in SRs, we found the average scores across reviews to decrease from 86 percent for part A (SR standards), 78 percent for part B (quantitative methods), 71 percent for part C (qualitative methods) and 66 percent for part D (integration).

On systematic review conduct and reporting, reviews tended to score best on specifying review questions and determining appropriate PICOS, and undertaking searches over an appropriate time period. Reviews scored marginally less well on reporting a deviation from the protocol (if any), articulating study designs for review sub-questions, avoiding bias through double-coding (at least for a random sample of included studies) and accounting for dependent findings consistently across quantitative and qualitative studies (Figure 13).

On the quality of quantitative evidence appraisal and synthesis, reviews tended to score highest on effect size calculation and reporting of heterogeneity. Reviews performed

worse on critical appraisal (using appropriate risk of bias assessment), synthesis methods (including reporting findings by bias categories), and worst on reporting characteristics of included studies (usually due to lack of independent coding by two reviewers) (Figure 14).

On the quality of qualitative evidence appraisal and synthesis (Figure 15), reviews tended to score best on methods for obtaining qualitative evidence (usually through separate systematic searches) and synthesis methods (often thematic synthesis). Reviews performed less well on reporting the findings of qualitative studies accurately (due to the single coding of evidence) and reporting evidence appropriately (indicating which evidence was of higher quality).

There also seems to have been an evolution in approaches to incorporating qualitative evidence over time. Early cases undertook 'enhanced reviews' (Noyes et al. 2011) by incorporating qualitative evidence contained in the studies eligible for the quantitative review of effects (King et al., 2010). This approach could be a model for evidence synthesis if the impact evaluations on which reviews of effects are based typically used theory-based approaches to open up the intervention black box and present outcomes along the causal chain. However, due to early realisation that this was typically not the case (White, 2009; Snilstveit, 2012) mixed-methods systematic reviews have been increasingly inclusive in incorporating qualitative evidence, including by undertaking additional searches for qualitative studies linked to the included quantitative studies (Watson et al., 2012) or by conducting full searches for qualitative studies to answer specific review questions (Waddington et al., 2014).

C. How have SRs integrated qualitative analysis?

Reviews have used different methods to integrate qualitative evidence. From our small sample of studies, it seems that earlier reviews used sequential explanatory designs to incorporate qualitative evidence after the quantitative synthesis had been conducted (King et al., 2010; Berg et al., 2012) (Appendix IV Figure 12). However, the authors used different methods of synthesis. King et al. adopt a theory-based approach drawing on a logic model/ theory of change and presenting evidence on implementation and processes according to that model. Berg and Denison (2010) use a realist synthesis approach to integrate the quantitative and qualitative evidence, without drawing on an explicit logic model or theory of change.

With regards to the coding tool, the average scores for integration of quantitative and qualitative evidence are lower (Figure 16). We found that reviews tended to score most highly on incorporating qualitative evidence in the design (usually through parallel review modules) and analysis (relating to specific parts of the causal chain), and integration of quantitative and qualitative evidence (e.g. through iterations of the logic model or theory of change, or through formal statistical testing). Reviews performed marginally less well on causal chain analysis of outcomes (it is still common practice that reviews do not collect evidence on intermediate and 'endpoint' outcomes consistently), although the extent to which reviews should necessarily do this as best practice is debatable. For example, it is not clear whether a review of hygiene and sanitation behaviour change as a primary outcome (de Buck et al., 2017) would necessarily need to incorporate outcomes further along the causal chain as primary outcomes as well. The

answer depends on the relevant question being asked in the review. Other questions in the tool relating to the use of qualitative evidence in other ways, for example to assess the applicability of evidence (D.5), may also need to be incorporated into previous questions (D.4) as it is doubtful whether reviews should necessarily provide this analysis. Finally, the methods in which implications for policy and practice are drawn from the quantitative and qualitative syntheses are usually limited (use of a summary of findings tables, GRADE or cerQUAL approaches).

VII. Discussion: Implications for the use of Qualitative Methods in SRs.

Reviews of qualitative evidence can answer questions about effectiveness early in the intervention causal chain, such as on intervention design and implementation or participant views. Reviews of quantitative impact evidence can answer questions about the change in net outcomes resulting from the intervention. Mixed methods reviews can go beyond the 'sum of their parts' to provide holistic answers about development effectiveness. In some cases, the key contribution of integration was the identification of impacts that informed the approach to analysing the quantitative data, ensuring a comprehensive consideration of the evidence, even in areas where little or no quantitative evidence was found (Pullin et al., 2013). In others, the contribution has been to provide evidence on the scalability of interventions (Waddington et al., 2014). In other cases, the integration has enabled reviews to provide evidence on unintended adverse outcomes for vulnerable groups. To take some examples of reviews not included in this assessment, Brody et al. (2017) find increases in gender-based violence due to economic self-help groups in the short term, and Lawry et al. (2017) find the displacement of women as a result of tenure reform in Africa.

We note here that the use of mixed methods in systematic reviews should be done as appropriate to the questions being asked and the underlying evidence base being reviewed (and also cognisant of the resources available to the review team). For example, it may be possible to explain heterogeneity in impacts across studies (i.e. answer the why question) drawing solely on quantitative impact evaluations, if (and only if) these studies collect outcomes along the causal chain (an example is Welch et al., 2016). This approach, however, requires a minimum number of SRs to be identified, which we know is still not available on many topics (Cameron et al., 2016). In some cases, it may be the case that sequential rather than simultaneous designs are sufficient to answer the questions being asked (*a priori* by helping develop the review questions and programme theory, *a posteriori* by helping explain findings such as impact heterogeneity). This might be the case, for example, for large comparative reviews of multiple interventions. But in other cases, such as where reviews aim to answer broader questions about a particular intervention, such as about implementation and then link that evidence with evidence on effects, simultaneous designs are optimal.

Indeed, it is at these initial stages of the review process that formal guidance is most lacking on effective mixed methods approaches, especially convening the study team and constructing the initial conceptual framework to support the integration of qualitative and quantitative evidence. Establishing teams with appropriate qualitative and quantitative skills, preferably drawing on broad academic disciplines, is usually needed for high quality mixed methods reviews to be done efficiently. All of the reviews

presented here included authors with quantitative and qualitative skills from academic disciplines including social sciences (e.g. anthropology, economics, policy science, public health and sociology) and environmental science. The reviews also drew on explicit programme theory (logic model, theory of change) to identify at what points qualitative and quantitative evidence provided the most valuable contributions to understanding the causal chain/pathways.

Section 5

PART THREE: CONCLUDING REMARKS AND RECOMMENDATIONS

VIII. Concluding Remarks

Our paper explores the contributions of qualitative methods to quantitative impact evaluations and systematic reviews. As described in our paper, the use of mixed methods can enhance the quality of the impact evaluations by strengthening data collection, analysis and interpretation. More specifically, qualitative methods can help discern how and for whom an intervention had impacts, the mechanisms that translated inputs into certain outcomes, the trajectory of identified impacts (linear or nonlinear), and also in identifying unforeseen consequences. Additionally, the complementarity between qualitative and quantitative methods can be usefully applied to systematic reviews at least to inform the questions being asked and the programme theory or logic model used, if not systematically throughout the data collection and analyses stages.

We note several important limitations to this study. First, our sample was purposive and limited to 40 studies across four sectors. Hence, our results may have limited applications for other sectors or impact evaluations in general. Second, while our tool underwent intensive internal testing and refinement, including consultations with thematic experts, like any quality appraisal tool, ours could be improved with further testing. However, it builds upon a number of existing validated tools, while adding elements (such as integration) that were not available in the original tools. Further, while we attempted to limit subjectivity by having the studies coded by two independent reviewers, it is unlikely that we eliminated it. Given the inherent subjectivity involved in quantifying the rigour and integration of MMIEs, we are confident that our credibility checks and reflexivity exercise increase the transparency of our review. Additionally, our analysis was limited to information reported in the study, which may have resulted in some inadvertent gaps. For instance, if a study did not refer to a study protocol or a pre-analysis plan, its absence would be noted in our coding sheet. However, it is possible that the study team did write a study protocol, but neglected to report it. In either case, we are confident that our recommendations will be useful to improve study reporting and conduct.

Similarly, in the systematic review section too, the sample was purposive and limited to 7 studies. It would be useful to develop further the tool that was piloted here by undertaking critical appraisals of more systematic reviews, preferably using double coding, and further refining critical appraisal questions. Further development of the tool should take into account the needs of decision makers.

Finally, we noted earlier how reflexive quantitative authors can describe how their personal views and backgrounds influence how survey questions and variable are framed. As a measure of reflexivity for this review,[10] we used the SR tool described in Section V to critically appraise our own use of mixed methods. The results are summarised in Box 3.

Box 3. How do we rate our own use of mixed methods in this review?

We used the mixed methods systematic review tool, described in section V,b, to appraise the methodological rigour and integration of our review.[1] Given that our paper is not a systematic review, our paper scored lower than average due to a lack of a systematic search, meta-analysis and other hallmarks of SRs. However, this exercise helped us to improve the quality of our reporting, such as explicitly stating the review questions, method of analysis and providing clear inclusion criteria for our search. Perhaps most usefully, this was an entry point for us to reflect on the mixed methods nature of our review.

In this paper, we developed and used two tools to quantify the rigour of qualitative and quantitative lines of inquiry, and integration of qualitative and quantitative methods within impact evaluations and systematic reviews. We also extracted common themes across studies for our discussion and recommendations. For this reason, we see our paper as a mixed methods review given that it utilises empirical and interpretive approaches to data collection and analysis. We agree with the point made by Morgan (2016) that recognising the blurry distinction between qualitative and quantitative research can help highlight the strengths of different methods in serving evaluation questions.

Our departure point for this review was to examine the role of qualitative methods in quantitatively-driven impact evaluations. Researchers undertaking these studies may have a limited understanding of the contributions of qualitative research. This is one of the reasons why undertaking this review is an important first step to understanding how the complementarity of methods can benefit quantitatively-driven IEs. A logical next step is to examine this complementarity in qualitatively-driven IEs.

IX. Recommendations

Based on our findings, we have the following recommendations for institutions like CEDIL, which are promoting innovations to improve evaluations. Our recommendations contain elements of 'best practice' guidance on mixed methods integration by others (See for example, Creswell et al., 2011). Where we see our contribution as unique in the

[10]We are grateful to James Copestake for suggesting this exercise.

literature is our focus on MMIEs, and specific guidance for reporting the integration in ways that may reduce the variation in how mixed methods are conceived, implemented and reported.

First, establish a common minimum understanding of mixed methods impact evaluations (MMIEs) and mixed methods systematic reviews (MMSRs) to establish semantic consistency in the way these studies are conceived.

Based on our review, we suggest the following definitions:

> A mixed methods impact evaluation (MMIE) assesses the net change in an outcome or set of outcomes, attributed to a specific programme or intervention by comparison with a counterfactual, using complementary strands of empirical and interpretative methods to serve the evaluation question(s) being investigated. MMIEs have a clear rationale for integrating methods and do so in at least one stage of the evaluation process in order to inform the interpretation of results.
>
> A mixed method systematic review (MMSR) uses transparent procedures to find, evaluate and synthesis the results of quantitative and qualitative evidence, in order to answer different questions. These can include questions relating to *programme effectiveness* along the causal pathway from programme design, implementation and targeting through to intermediate outcomes and endpoint outcomes; to collect information on participant or practitioner views; to explore heterogeneity in findings by context, programme design and implementation; or to analyse adverse outcomes (Waddington, Masset and Jimenez, 2018).

Second, establish common minimum reporting guidelines for MMIEs and MMSRs to ensure key elements related to the integration of methods are reported. Similar to extensions to the CONSORT guidelines, we recommend publishing a set of key elements which MMIEs and MMSRs should report. For MMIEs, we suggest the following integration elements be reported as best practice:

- A clear rationale for the integration of methods as they relate to the evaluation question(s) and/or study objective(s) at each stage of the evaluation;
- The composition of study teams, describing the process (e.g. boundary framework) for establishing a common premise behind the value of mixed methods to serve the particular evaluation question(s) and/or study objective(s);
- Adequate documentation required to understand the context of research decisions behind the integration of methods, and each component within qualitative and quantitative lines of inquiry;
- An acknowledgement of the 'boundaries' or limitations of both qualitative and quantitative approaches in serving the evaluation question(s) and/or study objective(s);
- An acknowledgement of how researcher backgrounds, thoughts, opinions, values, and/or perspectives fit within the research process (i.e. report reflexivity);
- The framework used for quantitative analysis and a plan for analysing/interpreting qualitative data. Note: even if the authors are using a

grounded theory approach which does not necessitate hypothesising, they should report this; and

- How specific findings (and policy recommendations, if applicable) were influenced by qualitative and quantitative data.

For MMSRs, we suggest the following elements be reported as best practice:

- The rationale for integrating mixed methods as they relate to the review questions, including acknowledging the limitations of qualitative and quantitative approaches in serving the evaluation question(s) and/or study objective(s);
- A theory of change unpacking black boxes of both intervention and outcomes, as well as articulating underlying assumptions, contexts and stakeholders, in order to guide the data collection and presentation;
- Search study flow diagrams indicating the approach to sourcing quantitative evidence eligible for inclusion, and qualitative evidence eligible for inclusion.
- Critical appraisal for each included quantitative and qualitative study, assessed using appropriate tools for quantitative and qualitative evidence.
- Separate reporting of results of quantitative and qualitative synthesis, followed by, where possible, an integrated synthesis drawing on the theory of change or other methods of analysis and presentation.
- Transparent reporting of the approach used to draw conclusions (especially implications for policy and practice) from the results – e.g. summary of findings tables drawing on GRADE or QUAL appraisal.

Third, develop and populate a repository of MMIEs and MMSRs that integrate qualitative and quantitative methods successfully. This repository could highlight studies where the integration of methods has led to a better appreciation of how and why an intervention works (or does not work), or where policy uptake has been high. These studies could serve as exemplars and guides for researchers who want to expand the scope of their studies beyond quantitative methods or qualitative methods alone. The repository would be helpful to draw upon best practice in MMIEs and MMSRs across the methodological, epistemological and ontological spectrum to inform what works and why in different evaluation contexts.

Fourth, devote adequate time, monetary and human resources to designing MMIEs and MMSRs. Our findings suggest that several times, quantitative studies may include qualitative methods to explore (or corroborate) quantitative findings. However, qualitative research can play a much more significant role in enhancing the quality (methodologically or even in terms of the findings) of the project. However, this requires careful planning of the integration of methods, ideally at different stages of the evaluation. Mapping the qualitative and quantitative components of an evaluation to the theory of change/hypothesised causal chain of a programme or intervention is a good way to conceptualise how and when both lines of enquiry should be integrated. Similarly, conducting an evaluability assessment can be useful to determine entry points and points of convergence for multidisciplinary teams within a mixed methods impact

evaluation (See Davies, 2013 for more on evaluability assessments). It is also recommended that multidisciplinary teams become truly transdisciplinary in working from a common framework of concepts, ideas, definitions and boundaries for robust mixed methods research. This is especially important to avoid "tokenistic" uses of qualitative methods, and achieve value for money when undertaking mixed methods impact evaluations.

Additionally, incorporating mixed methods into reviews of effects with parallel modules incorporating qualitative evidence requires adequate time, monetary and human resources from the outset. Commissioners of reviews, and review authors need to be aware of the challenges in undertaking mixed methods SRs. There is now a large library of systematic reviews on international development topics, many of which demonstrate the possibilities of applying 'single-method' review designs. Many of these existing reviews are now in need of updated searches, and a case can be made for also updating scope to ensure the reviews answer relevant questions for decision makers (Waddington, Masset and Jimenez, 2018). A promising and potentially efficient, way to assess the value addition of mixed methods reviews would be through a programme of work to undertake review updates on a set of priority existing reviews to trial different approaches to incorporating mixed methods. These might be standalone reviews or reviews produced alongside thematic collections of studies on particular topic areas.

Fifth, further develop the tools to assess MMSRs and MMIEs piloted here. We recommend that the critical appraisal of methodological rigour and integration tools developed here be used by other researchers in order to increase the reliability of the tool. Additionally, we recommend using the tool to assess qualitatively-driven impact evaluations (e.g. contribution analysis) to provide a logical complement to the scope of this paper. Beyond further refinements to the tool, this exercise would allow us to understand the complementarity of mixed methods across a broader spectrum of impact evaluation.

Further, it would be useful to conduct a survey of existing systematic reviews in international development to assess the extent to which they 1) use programme theory explicitly, and 2) engage with qualitative as well as quantitative evidence. 3ie's Systematic Reviews Repository would be a good starting ground for this work since the searches for that database were updated in 2017 and some initial work has begun to collect information on these categories for reviews.

Sixth, explore innovations in combining qualitative and quantitative data. A number of recent mixed methods approaches have been employed to integrate qualitative and quantitative lines of enquiry. For instance, Humphreys and Jacobs (2015) propose a unified analytical framework to aggregate across findings derived from quantitative analysis and process-based observations. In this approach, inferences from quantitative analysis may be supplemented by inferences from a smaller number of case studies by experts and adjusted to see how much they are likely to change if the expert opinions are informative or not. Glynn and Ichino (2014) suggest that when outcomes are difficult to measure, qualitative information can be converted into ordinal measures of outcomes within matched sets to reduce p-values. Additionally, QCA may be particularly useful in determining the causal constituents of complex interventions,

through a fully integrated and systematic method (See also, Intervention Component Analysis by Sutcliffe et al., 2015). Similarly, realist reviews are also useful innovations which seek to understand the particularities of intervention constituents in relation to their differential effectiveness for particular populations (Pawson et al., 2005). It is worth conducting a review of these methods to understand 1) how they interpret causal attribution in MMIEs, and 2) how they can add value in serving particular evaluation and review questions or contexts. In this way, they can be added usefully to a methodological tool belt for researchers and commissioners of MMIEs and MMSRs.

References

3ie, n.d.a Checklist for making judgements about how much confidence to place in a systematic review of effect (adapted version of SURE checklist). London: International Initiative for Impact Evaluation (3ie).

3ie, n.d.b Principles of Impact Evaluation. New Delhi: International Initiative for Impact Evaluation (3ie).

Abdel-tawab, N., Lūzā, S. F. & Zaki, A. 2008. Helping Egyptian women achieve optimal birth spacing intervals through fostering linkages between family planning and maternal/child health services. Citeseer.

Aker, J. C. 2017. Comparing cash and voucher transfers in a humanitarian context: Evidence from the Democratic Republic of Congo. The World Bank Economic Review, 31, 44-70.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A. & Tobias, J. 2012. Targeting the poor: evidence from a field experiment in Indonesia. American Economic Review, 102, 1206-40.

Anderson LM, P. M., Rehfuess E, Armstrong R, Ueffing E, Baker P, Francis D, Tugwell P. "Using logic models to capture complexity in systematic reviews." Res Synth Methods. 2011 Mar;2(1) doi: 10.1002/jrsm.32. Epub 2011, 33-42.

American Institutes for Research. (2017). *Humanitarian Cash Transfers in the Democratic Republic of the Congo: Evidence from UNICEF's ARCC II Programme.* Washington, DC: Author.

Andrade, E. L. 2012. Thinking Outside the Soapbox: Evaluating the Effectiveness of a Community-based Hygiene Promotion Intervention in Santa Clara, El Salvador. the George Washington University.

Bamberger, M. 2015. Innovations in the Use of Mixed Methods in Real-world Evaluation. Journal of Development Effectiveness, 7 (3), 327-335.

Bamberger, M., Rao. V, & Woolcock, M. 2010. Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development. Policy Research Working Paper Series 5245. Washington D.c.: the World Bank.

Bamberger, M., Rugh, J. & Mabry, L. 2012. Realworld Evaluation: Working Under Budget, Time, Data, and Political Constraints - 2nd ed. SAGE Publications.

Banerjee, A., Duflo, E., Imbert, C., Mathew, S. & Pande, R. 2015. Can E-governance Reduce Capture of Public Programs? Experimental Evidence from a Financial Reform of India's Employment Guarantee. In: 3ie (ed.) Impact Evaluation Report Series. New Delhi: 3ie.

Banerjee, A., Duflo, E., Imbert, C., Pande, R., Walton, M. & Mahapatra, B. 2014. An Impact Evaluation Of Information Disclosure On Elected Representatives' Performance: Evidence From Rural And Urban India, 3ie Impact Evaluation Report 11, August 2014. New Delhi: International Initiative For Impact Evaluation (3ie).

Befani, B. 2016. Pathways To Change: Evaluating Development Interventions With Qualitative Comparative Analysis (QCA). Elanders Sverige Ab, Stockholm, Sweden.

Berg, R.C. & Denison E. 2012. Interventions To Reduce The Prevalence Of Female Genital Mutilation/Cutting In African Countries. Campbell Systematic Reviews 2012:9 Doi: 10.4073/Csr.2012.9.

Brody, C., De Hoop, T. Vojtkova, M., Warnock, R., Dunbar, M., Murthy P. & Dworkin, S. 2017. Can Economic Self-help Group Programs Improve Women's Empowerment? A Systematic Review. Journal Of Development Effectiveness, 9 (1), 15–40. Doi:10.1080/19439342.2016.1206607.

Bonilla, J., Zarzur, R. C., Handa, S., Nowlin, C., Peterman, A., Ring, H. & Seidenfeld, D. 2017. Cash For Women's Empowerment? A Mixed-methods Evaluation Of The Government Of Zambia's Child Grant Program. World Development, 95, 55-72.

Bose, R. 2010. CONSORT Extensions for Development Effectiveness: guidelines for the reporting of randomised control trials of social and economic policy interventions in developing countries. Journal of Development Effectiveness, 2 (1), 173-186.

Brooker, S. & Halliday, K. 2015. Impact Of Malaria Control And Enhanced Literacy Instruction On Educational Outcomes Among School Children In Kenya.

Cameron, D., Mishra, A. & Brown, A. 2016. The growth of impact evaluation for international development: how much have we learned? Journal of Development Effectiveness, 8 (1), 1-21.

Candy, B., King, M., Jones, L. & Oliver, S. 2011. Using qualitative synthesis to explore heterogeneity of complex interventions. *BMC Medical Research Methodology*, 11:124 doi:10.1186/1471-2288-11-124.

Candy, B., King, M., Jones, L. and Oliver, S. 2013. Using qualitative evidence on patients' views to help understand variation in effectiveness of complex interventions: a Qualitative Comparative Analysis. *Trials 14:179* doi:10.1186/1745-6215-14-179.

Cardenas, S., Evans, D. & Holland, P. 2017. Estimating The Effects Of A Low-cost Early Stimulation And Parenting Education Programme In Mexico.

Carr-Hill, R., Rolleston, C., Schendel, R. and Waddington, H., 2018. The effectiveness of school-based decision making in improving educational outcomes: a systematic review. *Journal of Development Effectiveness, 10(1)*, pp.61-94.

Charmaz, K. 2000. Grounded theory: Objectivist and constructivist methods. In: *The Handbook of Qualitative Research*. Edited by N. K. Denzin and Y. Lincoln. Thousand Oaks, CA: Sage Publications, Inc.

Chinen, M. & Elmeski, M. 2016. Evaluation Of The Transformative Potential Of Positive Gender Socialization In Education For Peacebuilding. Washington, D.C.: Air.

Clements, T., Suon, S., Wilkie, D. S. & Milner-gulland, E. 2014. Impacts Of Protected Areas On Local Livelihoods In Cambodia. World Development, 64, S125-s134.

Coalition For Evidence-based Policy. 2010. Checklist For Reviewing A Randomized Controlled Trial Of A Social Program Or Project, To Assess Whether It Produced Valid

Evidence [Online]. Available: Http://Coalition4evidence.Org/Wp-content/Uploads/2010/02/Checklist-for-reviewing-a-rct-jan10.Pdf [Accessed].

Collaboration, S. 2011. Sure Checklist For Making Judgements About How Much Confidence To Place In A Systematic Review. Sure Guides For Preparing And Using Policy Briefs. Version, 2.

Creswell, J.W. 2014. Research Design: Qualitative, Quantitative And Mixed Methods Approaches. London: Sage Publications.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L. & Smith, K. C. 2011. Best Practices For Mixed Methods Research In The Health Sciences. Bethesda (Maryland): National Institutes Of Health, 2094-2103.

Critical Appraisal Skills Program, 2018. CASP Qualitative Checklist. Available at: https://casp-uk.net/wp-content/uploads/2018/01/CASP-Qualitative-Checklist.pdf.

Crouch, L., Korda, M. & Mumo, D. 2009. Improvements In Reading Skills In Kenya: An Experiment In The Malindi District. Research Triangle Park, Nc: Rti International. Retrieved From: Http://Ddp-ext. Worldbank. Org/Edstats/Kendprep09. Pdf.

Davis, P. 2011. Exploring The Long-term Impact Of Development Interventions Within Life-history Narratives In Rural Bangladesh. Journal Of Development Effectiveness, 3, 263-280.

Davies, D. and Dodd, J., 2002. Qualitative research and the question of rigor. Qualitative health research, 12(2), pp.279-289.

Davies, P. 2006. What Is Needed From Research Synthesis From A Policy-making Perspective? In: J. Popay, Ed. Moving Beyond Effectiveness In Evidence Synthesis - Methodological Issues In The Synthesis Of Diverse Sources Of Evidence. London: National Institute For Health And Clinical Excellence.

Davies, R. 2013. Planning evaluability assessments: A synthesis of the literature with recommendations. *In:* DFID (ed.) *DFID Working Paper 40.* London: UK: DFID.

De Buck, E., Van Remoortel, H., Hannes, K., Govender, T., Naidoo, S., Avau, B., Vande Veegaete, A., Musekiwa, A., Vittoria, L., Cargo, M., Mosler, H.-j., Vandekerckhove, P. & Young, T. 2017. Approaches To Promote Handwashing And Sanitation Behaviour Change In Low- And Middle-income Countries: A Mixed Method Systematic Review. Campbell Systematic Reviews 2017:7 Doi: 10.4073/Csr.2017.7

De Hoop, T., Van Kempen, L., Linssen, R. & Van Eerdewijk, A. 2014. Women's Autonomy And Subjective Well-being: How Gender Norms Shape The Impact Of Self-help Groups In Odisha, India. Feminist Economics, 20, 103-135.

De Wilde, C. K., Milman, A., Flores, Y., Salmerón, J. & Ray, I. 2008. An integrated method for evaluating community-based safe water programmes and an application in rural Mexico. Health Policy and Planning, 23, 452-464.

Denzin, N.K. and Lincoln, Y.S., 1998. The landscape of qualitative research: Theory and issues.

Denzin, N. K. & Lincoln, Y. S. 2018. *The Sage Handbook Of Qualitative Research*. Fifth

Edition. Los Angeles:Sage.

Dickin, K. & Griffiths, M. 1997. Designing By Dialogue: A Program Planner's Guide To Consultative Research For Improving Young Children Feeding.

Elliott, N. and Higgins, A., 2012. Surviving grounded theory research method in an academic world: Proposal writing and theoretical frameworks. *The Grounded Theory Review*, *11*(2), pp.1-12.

Evans, D., Hausladen, S., Kosec, K. & Reese, N. 2014. Community-based Conditional Cash Transfers In Tanzania: Results From A Randomized Trial, World Bank Publications.

Fearon, J., Humphreys, M. & Weinstein, J. 2008. Community-driven Reconstruction In Lofa County: Impact Assessment. Unpublished Manuscript.

Fetters, M. D., Curry, L. A. & Creswell, J. W. 2013. Achieving Integration In Mixed Methods Designs—principles And Practices. Health Services Research, 48, 2134-2156.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L.B., Vermeersch, C.M.J. 2016. *Impact Evaluation in Practice,* Washington, DC, World Bank Group.

Glenton, C., Colvin, C., Carlsen, B., Swartz, A., Lewin, S., Noyes, J. & Rashidian, A. 2013. Barriers And Facilitators To The Implementation Of Lay Health Worker Programmes To Improve Access To Maternal And Child Health: Qualitative Evidence Synthesis (Protocol). Cochrane Database Of Systematic Reviews 2013, Issue 2. Art. No.: Cd010414. Doi: 10.1002/14651858.Cd010414.

Glynn, A.N. and Ichino, N., 2015. Using qualitative information to improve causal inference. *American Journal of Political Science*, *59*(4), pp.1055-1071.

Greene, J. C., Caracelli, V. J. & Graham, W. F. 1989. Toward A Conceptual Framework For Mixed-method Evaluation Designs. Educational Evaluation And Policy Analysis, 11, 255-274.

Guiteras, R., Jina, A. & Mobarak, A. M. 2015. Satellites, Self-reports, And Submersion: Exposure To Floods In Bangladesh. American Economic Review, 105, 232-36.

Harden, A. & Thomas, J. 2005. Methodological Issues In Combining Diverse Study Types In Systematic Reviews. International Journal Of Social Research Methodology, 8 (3), 257-271.

Haynes, A. & Merttens, F. 2017. Evaluation Of The Uganda Social Assistance Grants For Empowerment (SAGE) Programme (No.354). International Policy Centre for Inclusive Growth.

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L. & Sterne, J. A. 2011. The Cochrane Collaboration's Tool For Assessing Risk Of Bias In Randomised Trials. British Medical Journal, 343, D5928.

Higgins, J.P.T. & Green, S. (Editors). 2011. Cochrane Handbook For Systematic Reviews Of Interventions Version 5.1.0 [Updated March 2011]. The Cochrane Collaboration, 2011. Available From: Www.Handbook.Cochrane.Org.

Hombrados, J.G. & Waddington, H. 2012. Tool To Assess Risk Of Bias And Internal Validity Of Social Experiments And Quasi-experiments. Mimeo. London: International

Initiative For Impact Evaluation (3ie).

Hulland, K., Martin, N., Dreibelbis, R., Debruicker Valliant, J. & Winch, P. 2015. What Factors Affect Sustained Adoption Of Safe Water, Hygiene And Sanitation Technologies? A Systematic Review Of Literature. London: Eppi-centre, Social Science Research Unit, Ucl Institute Of Education, University College London.

Humphreys, M. & Jacobs, A. M. 2015. Mixing methods: A Bayesian approach. *American Political Science Review,* 109**,** 653-673.

Humphreys, M., Sanchez De La Sierra, R. & Van Der Windt, P. 2014. Social And Economic Impacts Of Tuungane: Final Report On The Effects Of A Community-driven Reconstruction Programme In The Democratic Republic Of Congo. In: 3ie (Ed.) 3ie Impact Evaluation Report Series. New Delhi: International Initiative For Impact Evaluation (3ie).

National Democratic Institute, 2013. Constituency Dialogues And Citizen Engagement In Cambodia: Findings From A Mixed Methods Impact Evaluation. Washington, DC: National Democratic Institute.

Kerr, D., Keating, A., Poet, H., Spielhofer, T., Lopes, J. & Mundy, E. 2011. Evaluation Of The Schools Linking Network-final Report.

Kerwin, J. T. & Thornton, R. 2015. Making The Grade: Understanding What Works For Teaching Literacy In Rural Uganda.

King, E., Samii, C. & Snilstveit, B. 2010. Interventions To Promote Social Cohesion In Sub-saharan Africa. Synthetic Review 002. International Initiative For Impact Evaluation, New Delhi.

Kneale, D., Thomas, J. & Harris, K. 2015. Developing And Optimising The Use Of Logic Models In Systematic Reviews: Exploring Practice And Good Practice In The Use Of Programme Theory In Reviews. Plos One, 10 (11), E0142187. Doi:10.1371/Journal.Pone.0142187

Kneale, D. Gough, D., Bangpan, M., Waddington, H. & Thomas, J. 2018. Causal Chain Analysis In Systematic Reviews Of International Development Interventions. Unpublished Manuscript. Centre Of Excellence For Development Impact And Learning (Cedil), London.

Langer, L. 2017. Mixed-methods Critical Appraisal Tool. In: Evidence, A. C. F. (Ed.). Johannesburg: Africa Centre For Evidence.

Langford, R. & Panter-Brick, C. 2013. A Health Equity Critique Of Social Marketing: Where Interventions Have Impact But Insufficient Reach. Social Science & Medicine, 83, 133-141.

Lather, P., 1993. Fertile obsession: Validity after poststructuralism. The sociological quarterly, 34(4), pp.673-693.

Lawry, S., Samii, C., Hall, R., Leopold, A., Hornby, D. & Mtero, F. 2017. The Impact of Land Property Rights Interventions on Investment and Agricultural Productivity in Developing Countries: A Systematic Review. Journal of Development Effectiveness, 9 (1): 61–81. doi:10.1080/19439342.2016.1160947.

Lewin, S., Oxman, A.D., Lavis, J.N. & Fretheim, A. 2009. Support Tools For Evidence-informed Health Policymaking (Stp) 8: Deciding How Much Confidence To Place In A Systematic Review. Health Research Policy And Systems 7 (Suppl 1), S8.

Maden, M., Cunliffe, A., Mcmahon, N., Booth, A., Carey, G.M., Dickson, R. & Gabbay, M. 2017. Use Of Programme Theory To Understand The Differential Effects Of Interventions Across Socio-economic Groups In Systematic Reviews-a Systematic Methodology Review. Systematic Reviews, 6 (1), 266. Doi: 10.1186/S13643-017-0638-9.

Masset, E., Rathinam, F., Nath., M. & Wood, B. 2018. Successful impact evaluations: lessons from DFID and 3ie. *CEDIL Inception Paper*.

Maxwell, J., Chmiel, M. and Rogers, S. (2015). Designing Integration in Multimethod and Mixed Methods Research. In: S. Hesse-Biber and R. Johnson, ed., The Oxford Handbook of Multimethod and Mixed Methods Research enquiry, 1st ed. New York: Oxford University Press.

Merttens, F., Hurrell, A., Marzi, M., Attah, R., Farhat, M., Kardan, A. & Macauslan, I. 2013. Kenya Hunger Safety Net Programme Monitoring And Evaluation Component. Impact Evluation Final Report. Oxford Policy Management.

Miles, M. B. & Huberman, A. M. 1994. Qualitative Data Analysis: An Expanded Sourcebook, Sage.

Mmari, K. and T. J. H. University (2006). Using Qualitative Methods for Monitoring and Evaluation. Fundamentals of program evaluation, The Johns Hopkins University.

Mo, D., Bai, Y., Boswell, M. & Rozelle, S. 2016. Evaluating The Effectiveness Of Computers As Tutors In China. In: 3ie (Ed.) 3ie Impact Evaluation Report Series. New Delhi: International Initiative For Impact Evaluation (3ie).

Mollinga, P. P. 2010. Boundary work and the complexity of natural resources management. *Crop Science*, 50, S-1-S-9.

Morgan, D.L., 2016. Living Within Blurry Boundaries: The Value of Distinguishing Between Qualitative and Quantitative Research. *Journal of Mixed Methods Research*, p.1558689816686433.

Morra Imas, L. G. & Rist, R. C. 2009. The Road To Results: Designing And Conducting Effective Development Evaluations, World Bank Publications.

Munro, S.A., Lewin, S.A., Smith, H.J., Engel, M.E., Fretheim, A. & Volmink, J. 2007. Patient Adherence To Tuberculosis Treatment: A Systematic Review Of Qualitative Research. Plos Med. 2007;4(7):E238.

Nielsen, N., Godden, K., Leguene, P., Ruegenberg, D. & Rüdiger, J. 2010. Wfp Cambodia School Feeding 2000–2010: A Mixed Method Impact Evaluation. Rome: World Food Programme Office Of Evaluation.

Nisbett, N., Longhurst, R., Barnett, I., Feruglio, F., Gordon, J., Hoddinott, J., Jahan, F., Karachiwalla, N., Roy, S. & Shah, V. 2016. Impact Evaluation Of The Dfid Programme To Accelerate Improved Nutrition For The Extreme Poor In Bangladesh.

Noblit, G.W. & Hare, R.D. 1988. Meta-ethnography: Synthesizing Qualitative Studies.

Qualitative Research Methods, Volume 11. Newbury Park, California: Sage Publications.

Norris, N., 1997. Error, bias and validity in qualitative research. *Educational action research*, *5*(1), pp.172-176.

Noyes, J., Popay, J., Pearson, A., Hannes, K. & Booth, A. On Behalf Of The Cochrane Qualitative Research Methods Group. 2011. Qualitative Research And Cochrane Reviews. Chapter 20 In Higgins, J.P.T. And Green, S. (Editors) Cochrane Handbook For Systematic Reviews Of Interventions Version 5.1.0. The Cochrane Collaboration.

Olken, B. A. (2015). "Promises and perils of pre-analysis plans." Journal of Economic Perspectives 29(3): 61-80.

Palladium, 2015. Independent Evaluation of the Security Sector Accountability and Police Reform Programme. Final Evaluation Report. London: Palladium.

Pandey, P., Goyal, S. & Sundararaman, V. 2011. Does Information Improve School Accountability?: Results Of A Large Randomized Trial, World Bank Washington, Dc.

Parthasarathy, R., Rao, V. & Palaniswamy, N. 2017. Unheard Voices: The Challenge Of Inducing Women's Civic Speech. The World Bank.

Patton, M. Q. 2003. Qualitative Evaluation Checklist. Evaluation Checklists Project.

Pawson, R. 2002. Evidence-based Policy: The Promise Of 'realist Synthesis'. Evaluation, 8 (3), 340-358.

Pawson, R., Greenhalgh, T., Harvey, G. & Walshe, K. 2005. Realist review - a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy,* 10**,** 21-34.

Pawson, R. 2006. Evidence-based Policy: A Realist Perspective. London: Sage.

Phillips, D., Waddington, H. & White, H. 2015. Better Targeting Of Farmers As A Channel For Poverty Reduction: A Systematic Review Of Farmer Field Schools Targeting. Development Studies Research, 1 (1), 113-136.

Piza, C., Cravo, T., Taylor, L., Gonzalez, L., Musse, I., Furtado, I., Sierra, A.C. & Abdelnour, S. 2016. Business Support For Small And Medium Enterprises In Low- And Middle-income Countries: A Systematic Review, 3ie Systematic Review 25. London: International Initiative For Impact Evaluation.

Pluye, P., Robert, E., Cargo, M., Bartlett, G., O'cathain, A., Griffiths, F., Boardman, F., Gagnon, M.-p. & Rousseau, M. 2011. Proposal: A Mixed Methods Appraisal Tool For Systematic Mixed Studies Reviews. Montréal: Mcgill University, 1-8.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A. & Artha, R. P. 2014. Improving Educational Quality Through Enhancing Community Participation: Results From A Randomized Field Experiment In Indonesia. American Economic Journal: Applied Economics, 6, 105-26.

Puri, J., Aladysheva, A., Iversen, V., Ghorpade, Y. & Brück, T. 2017. Can Rigorous Impact Evaluations Improve Humanitarian Assistance? Journal Of Development Effectiveness, 9, 519-542.

Qin, D. 2016. Positionality [Online]. Online. Available: Https://Onlinelibrary.Wiley.Com/Doi/10.1002/9781118663219.Wbegss619 [Accessed].

Rao, V. and M. Woolcock (2003). "Integrating qualitative and quantitative approaches in program evaluation." The impact of economic policies on poverty and income distribution: Evaluation techniques and tools: 165-190.

Rao, V., Ananthpur, K. & Malik, K. 2017. The Anatomy Of Failure: An Ethnography Of A Randomized Trial To Deepen Democracy In Rural India. World Development, 99, 481-497.

Bath Social and Development Research. n.d. Quip And The Yin/Yang Of Quant And Qual: How To Navigate Quip Visualisations. Available from: http://bathsdr.org/wp-content/uploads/2017/07/How-to-navigate-QuIP-visualisations.pdf.

Rosenfield, P. L. 1992. The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science & Medicine*, 35, 1343-1357.

Roy, S. K., Jolly, S. P., Shafique, S., Fuchs, G. J., Mahmud, Z., Chakraborty, B. & Roy, S. 2007. Prevention Of Malnutrition Among Young Children In Rural Bangladesh By A Food-health-care Educational Intervention: A Randomized, Controlled Trial. Food And Nutrition Bulletin, 28, 375-383.

Ryan, L. & Golden, A. 2006. 'Tick the Box Please': A Reflexive Approach to Doing Quantitative Social Research. *Sociology*, 40, 1191-1200.

Saha, S., Kermode, M. & Annear, P. 2015. Effect Of Combining A Health Program With A Microfinance-based Self-help Group On Health Behaviors And Outcomes. Public Health, 129, 1510-1518.

Samsi, K. N.D. Critical Appraisal Of Qualitative Research [Powerpoint Presentation]. In: The Policy Institute, K. S. C. (Ed.). London: King's College London.

San Antonio, D. M., Morales, N. S. & Moral, L. S. 2011. Module-based Professional Development For Teachers: A Cost-effective Philippine Experiment. Teacher Development, 15, 157-169.

Schulte-Mecklenbeck, M., Kühberger, A. and Johnson, J.G. eds., 2011. A handbook of process tracing methods for decision research: A critical review and user's guide. Psychology Press.

Schulz, K. F., Altman, D. G. & Moher, D. 2010. Consort 2010 Statement: Updated Guidelines For Reporting Parallel Group Randomised Trials. Bmc Medicine, 8, 18.

Shaffer, P. 2013. Q-squared: Combining Qualitative And Quantitative Approaches In Poverty Analysis. Oxford: Oxford University Press.

Shattuck, D., Kerner, B., Gilles, K., Hartmann, M., Ng'ombe, T. & Guest, G. 2011. Encouraging Contraceptive Uptake By Motivating Men To Communicate About Family Planning: The Malawi Male Motivator Project. American Journal Of Public Health, 101, 1089-1095.

Shea, B., Reeves, B.C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P.,

Welch, V., Kristjansson, E. & Henry, D.A. 2017. Amstar 2: A Critical Appraisal Tool For Systematic Reviews That Include Randomised Or Non-randomised Studies Of Healthcare Interventions, Or Both. British Medical Journal 2017;358:J4008.

Skalidou, D. & Oya, C. 2018. The Challenges Of Screening And Synthesising Qualitative Research In A Mixed-methods Systematic Review. The Case Of The Impact Of Agricultural Certification Schemes. Journal Of Development Effectiveness, 10 (1).

Snilstveit, B. 2012. Systematic Reviews: From 'Bare Bones' Reviews To Policy Relevance. Journal Of Development Effectiveness, 4 (3), 388-408.

Snilstveit, B., Vojtkova, M. & Oliver, S. 2012. Narrative Approaches To Systematic Review And Synthesis Of Evidence For International Development Policy And Practice. Journal Of Development Effectiveness, 4 (3), 409-430.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M. & Eyers, J. 2015. Interventions For Improving Learning Outcomes And Access To Education In Low- And Middle-income Countries: A Systematic Review, 3ie Final Review. London: International Initiative For Impact Evaluation.

Snilstveit, B, Stevenson, J, Menon, R,Phillips, D, Gallagher, E, Geleen, M,Jobse, H, Schmidt, T and Jimenez, E, 2016.The impact of education programmes on learning and school participation in low- and middle-income countries: a systematic review summary report,3ie Systematic Review Summary 7. London: International Initiative for Impact Evaluation (3ie)

Snilstveit, B., J. Stevenson, I. Shemilt, Jimenez, E. Clarke, & J. Thomas. 2018. Efficient, timely and living systematic reviews: opportunities in international development. Unpublished manuscript. London: Centre of Excellence for Development Impact and Learning.

Sutcliffe, K., Thomas, J., Stokes, G., Hinds, K. & Bangpan, M. 2015. Intervention Component Analysis (ICA): a pragmatic approach for identifying the critical features of complex interventions. *Systematic Reviews,* 4**,** 140.

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., Brunton, G. & Kavanagh, J. 2004. Integrating Qualitative Research With Trials In Systematic Reviews. British Medical Journal 2004;328;1010 Doi: Https://Doi.Org/10.1136/Bmj.328.7446.1010

Thorne, S., Stephens, J. and Truant, T., 2016. Building qualitative study design using nursing's disciplinary epistemology. *Journal of advanced nursing*, *72*(2), pp.451-460.

Torgerson, C., Wiggins, A., Torgerson, D., Ainsworth, H., Barmby, H., Hewitt, C., Jones, K., Hendry, V., Askew, M. & Bland, M. 2011. Every Child Counts: The Independent Evaluation. Report for the Department of Education. Research Report DFE-RR091a, 29 March. London: DFE.

Van Der Knapp, L.M., Leeuw, F.L., Bogaerts, S. & Nijssen, L.T.J. 2008. Combining Campbell Standards And The Realist Evaluation Approach: The Best Of Two Worlds? American Journal Of Evaluation, 29 (1), 48-57. Doi: 10.1177/1098214007313024

Vu, N., Van Velzen, R., Lensink, R. & Bulte, E. 2015. The Impact Of Gender And Business

Training For Female Microfinance Clients In Vietnam. Grantee Final Report. New Delhi: International Initiative For Impact Evaluation (3ie).

Waddington, H., Snilstveit, B., White, H. & Fewtrell, L. 2009. Water, Sanitation And Hygiene Interventions To Combat Childhood Diarrhoea In Developing Countries. Synthetic Review 001. New Delhi: International Initiative For Impact Evaluation.

Waddington, H., White, H., Snilstveit, B., Hombrados, J.G., Vojtkova, M., Davies, P., Bhavsar, A., Eyers, J., Koehlmoos, T.P., Petticrew, M., Valentine, J.C. & Tugwell, P. 2012. How To Do A Good Systematic Review Of Effects In International Development: A Tool Kit. Journal Of Development Effectiveness, 4 (3), 359-387.

Waddington, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Phillips, D., Davies, P. & White, H. 2014. Farmer Field Schools For Improving Farming Practices And Farmer Outcomes In Low-and Middle-income Countries: A Systematic Review. Campbell Systematic Reviews, 10.

Waddington, H. Masset, E. & Jimenez, E. 2018. What Have We Learned From 10 Years Of Systematic Reviews In International Development? Journal Of Development Effectiveness, 10 (1).

Watson, J., Byrne, R., Opazo, J., Tsang, F., Morgan-jones, M. & Diepeveen, S. 2012. What Are The Major Barriers To Increased Use Of Modern Energy Services Among The World's Poorest People And Are Interventions To Overcome These Effective? CEE Review 11-004. Collaboration for Environmental Evidence. http://www.environmentalevidence.org/wp-content/uploads/2014/07/CEE11-004.pdf (last accessed 23/08/2018).

Welch, V. A., E. Ghogomu, A. Hossain, S. Awasthi, Z. A. Bhutta, C. Cumberbatch, R. Fletcher, et al. 2016. Deworming and Adjuvant Interventions for Improving the Developmental Health and Well-Being of Children in Low- and Middle-Income Countries: A Systematic Review and Network Meta-Analysis. Campbell Systematic Reviews 2016: 7. doi:10.4073/csr.2016.7.

White, H. 2008. Of Probits and Participation: The Use of Mixed Methods in Quantitative Impact Evaluation. IDS Bulletin, 39, 98-109.

White, H. 2009. Theory-based Impact Evaluation: Principles And Practice. Journal Of Development Effectiveness, 1 (3), 271-284.

White H (2010). A Contribution to Current Debates in Impact Evaluation. Evaluation 16(2): 153- 164

White, H. & Raitzer, D.A. 2017. Impact Evaluation Of Development Interventions: A Practical Guide. Manila: Asian Development Bank. Http://Dx.Doi.Org/10.22617/Tcs179188-2

White, H. 2018. Theory Based Systematic Reviews. Journal Development Effectiveness, 10 (1).

Woolcock, M. 2018. Reasons For Using Mixed Methods In The Evaluation Of Complex Projects. In: Nagatsu, M and Ruzzene A (eds.) Philosophy and Interdisciplinary Social Science: A Dialogue. London: Bloomsbury Academic.

Yeager, B. A., Huttly, S. R., Diaz, J., Bartolini, R., Marin, M. & Lanata, C. F. 2002. An Intervention For The Promotion Of Hygienic Feces Disposal Behaviors In A Shanty Town Of Lima, Peru. Health Education Research, 17, 761-773.

Appendix I. Complete tool for mixed methods impact evaluations.

First author last name, year	
Title of the study	
Sector focus:	
Donor:	
Classification of impact evaluation design: clarify design/ method of analysis	
RCT/Natural or quasi/Non-experimental	
Blue boxes are not scored: They are only intended to collect descriptive information.	
Part A: Quantitative aspect	
Scoring guide	**[2]-**All 6 lines are marked Yes||**[1]**-First 4 lines are marked yes||**[0]**-if above requirements are not met
A.1 Are the study's quantitative methods and procedures clearly described?	
o Research questions suited to quantitative enquiry	
o Methodology for quantitative evaluation *eg: RDD, PSM, RCT etc.*	
o Outcome(s) relevant to the quantitative evaluation *eg: outcomes relevant to the* quantitative *research question(s)*	
o Participants/ settings/ population specific to the quantitative evaluation *eg: demographics of sample and numbers*	
o Units of allocation/ treatment *e.g; level at which data was collected-individuals, groups, social organizations etc.*	
o Units of analysis *e.g; level at which data is being analyzed in the study-individuals, groups, social organizations etc.*	
Comments (note important limitations or uncertainty)	

Scoring guide	**[2]-**First 4 lines are marked Yes, All 5 lines should be marked Yes where the 5th point is applicable||**[0]**-if above requirements are not met
A.2 Is the quantitative study rigorous in conduct?	
o Rationale for quantitative assessment e.*g: reasons for using RCT, PSM, FE models, etc*	
o Rationale for quantitative sampling criteria *e.g; reasons for choosing a representative or appropriate sample*	
o Clear description of process by which data was collected *e.g: how, why and at what stages was data collected*	
o Clear framework for quantitative analysis *e.g: defined path for data interpretation and subsequent analysis*	
o Reasons for modifications to methods, if applicable *e.g; If the methods for analysis was changed at any point during the study, is there any rationale provided for modifications to methods?*	
Comments (note important limitations or uncertainty)	
Scoring guide	**[2]-**3 or 4 lines are marked Yes, 4 or 5 where the 5th line is applicable||**[1]**-2 lines are marked yes, 3 lines where the 5th line is applicable||**[0]**-if above requirements are not met
A.3 Selection bias and confounding (RCTs):	
Is there a clear description of the randomization sequence generation?	
o a random component in the sequence generation process is described (e.g. referring to a random number table) which was	

determined centrally (e.g. by researchers or through a public lottery)	
o the method was used to allocate all participants in the study at the start of the intervention	
o the authors present tables indicating balance in covariates at baseline	
o the authors present data on losses to follow-up (attrition) in intervention groups by covariates	
o random allocation is made to 'encouragement' to treatment (e.g. information campaign about the intervention) only	
Comments (note important limitations or uncertainty)	
Scoring guide	**[2]:** identify methodology and present balance tables; **[1]:** identify methodology but no balance tables provided; **[0]:** can't identify methodology
A.4 Other designs (natural experiment, quasi-experimental, non-experimental design):	
Is there a clear description of the design and methods used to control for selection bias and confounding?	
o all participants are allocated to intervention groups based on a quasi-random selection process at the start of the intervention (natural experiment) and present tables indicating balance in covariates at baseline	
o all participants are allocated to intervention groups based on a test score on a continuous variable measured at pre-test at the start of the study (regression discontinuity design) and present tables indicating balance in covariates at baseline	
o all participants are allocated to intervention groups non-randomly and methods are used to account for unobservable confounding using outcomes data collected pre-and post-intervention on a panel of observations (e.g. difference in differences or fixed effects estimation) and present tables indicating balance in covariates at baseline	

o all participants are allocated to intervention groups non-randomly and methods are used to account for confounders between groups (e.g. instrumental variables, regression adjustment, statistical matching) and present tables indicating balance in covariates at baseline	
o all participants are allocated to intervention groups non-randomly and methods are used to ensure groups are as similar as possible (e.g. naive matching) and present tables indicating balance at group level	
o all participants are allocated to intervention groups non-randomly without methods to ensure or any reporting of comparability across groups (e.g. outcome means are presented without balance assessment)	
Comments (note important limitations or uncertainty)	
Scoring guide	**[2]:** if all from applicable are addressed (yes)||**[1]:** if 1 or more from those applicable is addressed(yes)||**[0]:** if none addressed(No), or acknowledged but unclear on how it is addressed
A.5 Post-intervention biases: motivation of participants	
o Bias due to non-adherence (spillover and crossovers effects) (e.g. intention to treat or instrumental variables estimation used)	
o Recall bias (e.g. reporting over a sufficiently short period – usually two weeks or less, or tools used to help participants accurately recall longer recall periods)	
o Social desirability (courtesy) bias (e.g. blinding of participants to intervention, or outcomes data collected using hard measures, or efforts made to reduce biases in self-reporting)	
o Hawthorne/John Henry effects (e.g. blinding of participants to intervention is used, or researchers minimise the number of site visits and data collection periods)	
Comments (note important limitations or uncertainty)	

Scoring guide	**[2]-**Two scored lines are marked Yes||**[1]**-One of the scored lines is marked yes||**[0]**-if above requirements are not met
A.6 Post-intervention biases: analysis and reporting	
o mention a study protocol or pre-analysis plan	
o report findings for outcomes whether or not findings they are statistically significant	
o report findings at the study level	
o report findings for particular sub-groups of participants (e,g, men and women separately)	
o Study data is retained and available for reanalysis. (e.g datasets, analysis files etc.)	
Comments (note important limitations or uncertainty)	
Scoring guide	**[2]-**Two scored lines are marked Yes||**[1]**-One of the scored lines is marked yes||**[0]**-if above requirements are not met
A.7 Threats to construct and external validity:	
Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect. Reasons why inferences about how study results would hold over variations in persons, settings, treatments, and outcomes may not be correct	
o Outcome constructs used represent adequately the outcomes of interest in the study and not intermediate outcomes e.g attitude instead of the quality of life.	
o Study participants are a random sample of the underlying population of interest	
o The study reports on how applicable the findings are to 'real world' implementation contexts (if relevant)	

Comments (note important limitations or uncertainty)	
Scoring guide	**1=Yes 0=No**
A.8 Does the study report statistical power?	
Does the study report being underpowered to detect changes in the outcome?	
Overall score quantitative (out of 13)	
Part B: Qualitative component	
Scoring guide	**[2]-**All 4 lines are marked Yes||**[1]**-First 3 lines are marked yes||**[0]**-if above requirements are not met
B.1 Are the study's methods and procedures clearly described?	
o Research questions suited to qualitative enquiry *eg: questions not attributing impact but exploring questions relevant to the IE at the pre, mid or post-intervention stage*	
o Methodology for qualitative evaluation *eg: FGD, interview, observation, etc.*	
o Outcome(s) relevant to the qualitative evaluation *eg: outcomes relevant to the qualitative research question(s)*	
o Participants/ settings/ population specific to the qualitative evaluation *eg: demographics of the sample and numbers*	
Scoring guide	**[2]:** if all from scored and applicable are addressed (4 or 5 yes)**[0]:** if one or more from scored and applicable lines is not marked yes
B.2 Is the qualitative study rigorous in conduct?	
o Rationale for qualitative methods	

E.g.: reasons for using observation, case study or IDI	
o Rationale for qualitative sampling criteria *E.g.: sampling of participants with reasons*	
o Clear description of process by which data was collected *e.g.: for interviews, an indication of how they were conducted and the schedule used*	
o Evidence of how analytical categories have been generated for qualitative data collection and analysis *E.g.: for more than one data collection method such as FDGs + IDIs, a framework for categorising collection to make clear how they intend to measure the same thing. If both attempts to understand satisfaction, how are they defining its measure?*	
o Distinction between framework for data collection and analysis/interpretation *Eg: a study may provide a framework for collection such as an interview schedule, but not one for subsequent analysis and interpretation. How did they infer from qualitative findings, if there is no evidence on the framework used for analysis?*	
O Reasons for modifications to methods, if applicable *Eg: if the questions in an interview, or plan for the qualitative study design changed over the course of time, is there any rationale provided for modifications to methods?*	
Comments (note important limitations or uncertainty)	
Scoring guide	**[2]-**Three scored lines are marked Yes||**[1]**-Any two scored lines are marked yes||**[0]**-if above requirements are not met
B.3 Has the researcher been explicit and self-aware about assumptions, values and biases?	

o Acknowledgement of how bias might influence data collection	
o Acknowledgement of scope for personal opinions/values influencing analysis	
o Acknowledgement of how bias might influence interpretations and presentation	
Comments (note important limitations or uncertainty)	
Scoring guide	**[2]-**3 or 4 (where applicable)scored lines are marked Yes||**[1]**-2 or 3 (where applicable) scored lines are marked yes||**[0]**-if above requirements are not met
B.4 Were the methods used to address bias adequate?	
o Attempts to validate the credibility of findings E*.g.: Independent coding by at least two analysts/respondent validation, peer reviews (Langer VI, Kings college)*	
o Efforts made to explain contradictory or divergent observations/findings, if applicable *Eg: For more, please refer to checklists by the world bank, Langer and Kings college*	
o Explicit rules for the confirmation of hypotheses or research questions *Eg: clear sequence by which study design would lead to answers to research questions or hypotheses. For more, please refer to the world bank checklist.*	
o Researcher's reactions to key events, if applicable *Eg: If the study design had to be changed as a result of a shift in government or other contextual circumstances, was the researcher explicit about personal reactions to these events, and how the study changed as a result?*	
o mention of a study protocol or pre-analysis plan	
o report findings for particular sub-groups of participants (e,g, men and women separately)	

o mention of publicly available data	
Scoring guide	**[2]-**Two scored lines are marked Yes||**[1]**-One of the scored lines is marked yes||**[0]**-if above requirements are not met
B.5 Are the results transferable to other settings?	
o Descriptions of scope, limitations and context within which the evaluation is situated *Eg: description of the context and conditions under which phenomena of interest occur, and the scope and limitations of data presented to enable generalisation to other settings*	
o Participants perspectives placed in personal contexts *Eg: Clear mention of the contexts within which all sub-groups covered operate. For instance, a study on handwashing could present the handwashing findings for a different subgroup of the sample, along with explanations for specific characteristics of different findings*	
Comments (note important limitations or uncertainty)	
Scoring guide	**Yes** or **No** only
B6. Did qualitative data situate the findings within the political, institutional, cultural or social context?	
Overall score qualitative	
Section C: Integration of Mixed Methods (out of 6)	
C.1 Integration of theory of change/programme or logic model	**Mark 'yes' if applicable. Unscored.**
o The authors describe how the causal chain will be explored through mixed methods. *Eg: the authors may clearly link components of the intervention/programme theory of change with specific qualitative and quantitative methods of enquiry."To probe causal link X, we conductted focus group discussions and used open-ended questionnaires with Y sample; the impact of X on Y as seen in the theory of change was estimated through a difference-in-difference methodology." The main point is to link the mixed methods to the theory of change.*	

C.2 Study design	*Put overall numeric score here*
	[2]- Indicator 1 and/or 2 are marked Yes, and indicator 3 is marked yes. |[1]- One indicator is marked is yes. [0]-none of the indicators are marked yes.
o The authors refer to how qualitative data influenced quantitative data collection. *Eg: qualitative methods of enquiry through formative work could have informed the survey questionnaire used for quantitative data collection.*	
o The authors refer to how quantitative data influenced qualitative data collection. *Eg:, if quantitative data report that a particular subgroup is not benefiting from the intervention, the authors report using qualitative methods of enquiry to find out why this group was not reached by the intervention.*	
o The authors make clear the rationale for integrating quantitative and qualitative methods. *Eg: the authors explain that to investigate the impact of an intervention on handwashing, they conducted self-reported surveys to measure change in handwashing rates, in addition to structured observation to observe the behaviour directly. The main point here is to explain why mixed methods are used to evaluate the impact of a programme/intervention.*	
C.3 The interpretation of findings	*Put overall numeric score here*
Scoring guide	**[2] All applicable indicators are marked yes.|[1]- One or two applicable indicators are marked yes. [0]- none of the**

	indicators are marked yes.
o The authors make clear how **qualitative and quantitative** data led to specific inferences or interpretations of findings. *Eg: the authors make clear that quantitative findings suggest that the impact of programme X on Y outcome was limited, but qualitative findings revealed that beneficiaries actually benefited from programme X. Alternatively, they may report that focus group discussions with beneficiaries confirm the interpretation that programme X was in effective in terms of Y outcome. The main point here is to that the authors show how both types of data were used to inform the interpretation of findings.*	
o The authors report how mixed methods data influenced their policy recommendations if provided. *Eg: the authors mention that the focus group discussion and interview data from subgroup X have implications for the programme's reach. In this regard, they may frame recommendations on improving equity or access to the programme based on qualitative findings, while reporting the overall effect of the programme through quantitative findings. The main point here is that if the authors provide policy recommendations, they mention how the integration of methods led to the formulation of the recommendations.*	
o Efforts made to explain contradictory or divergent observations/findings, if applicable *Eg: the authors may find that their quantitative survey data demonstrate a positive impact of intervention X on outcome Y for population Z. However, they note that qualitative interviews suggested that intervention X actually had an uncaptured negative effect on population Z. The authors may determine that this occurred because their quantitative survey did not account for factor F, which only arose through the qualitative work.*	
o The authors make clear the value-added of mixed methods of informing the study findings.	
C.4 Limitations of Integration	*Put overall numeric score here*
Scoring guide	**[2]** - Yes **[0]** - No
O The authors report the limitations of integrating qualitative and quantitative data	

Eg: the authors mention that their qualitative data comprised a small sample whose findings diverged from the quantitative sample, and consequently are unlikely to be representative of the entire sample. In this case, they might mention that a particular subgroup in the qualitative sample merits further enquiry.The main point here is for the authors to delimit the usefulness or trustworthiness of the mixed methods integration. What is it useful for, and what is beyond its scope to answer?	
Scoring guide	*mark yes where applicable*
C.5 At what stage is the qualitative evidence incorporated into the study?	
o Pre-intervention (*e.g.in the design of the evaluation questions (e.g. theory of change) or in the design of the data collection tools (e.g. formatively to determine sample or design survey questions)*	
o During intervention (*e.g.to provide factual evidence on intervention design, targeting and implementation or to provide factual evidence on participant/non-participant adherence (uptake of treatment)*	
o Post-intervention (*e.g to provide evidence about unintended consequences or to provide evidence about heterogeneity for sub-groups of participants (e.g. those for whom quantitative sampling is insufficiently powered)*	
o To provide evidence after endline analysis *(e.g. to explore reasons for quantitative findings or assess the applicability of quantitative evidence to other contexts)*	
C.6 Categorise the type of mixed methods study (studies can meet multiple criteria):	
1 Sequential explanatory design (QUANT-->qual)	
The quantitative component is followed by the qualitative. The purpose of sequential explanatory design is to explain quantitative results using qualitative findings. The quantitative results guide the selection of qualitative data sources and data collection, and the qualitative findings contribute to the interpretation of quantitative results. E.g. Using qualitative enquiry to assist in explaining and interpreting the findings of a quantitative study post quantitative analysis.	
2 Sequential exploratory design (qual-->QUANT)	

The qualitative component is followed by the quantitative. The purpose of sequential exploratory design is to explore, develop and test an instrument (or taxonomy), or a conceptual framework (or theoretical model). The qualitative findings inform the quantitative data collection. E.g following up formative qualitative work on the evaluation (not the intervention) with quantitative analysis to validate or generalize the findings.	
3 Convergent designs (qual+QUANT or QUANT+qual)	
The qualitative and quantitative components are concomitant and integrated. The purpose of convergent designs is to examine the same phenomenon by interpreting qualitative and quantitative results (bringing data analysis together at the interpretation stage), or by integrating qualitative and quantitative datasets (e.g., data on same cases), or by transforming data (e.g., quantization of qualitative data), or to support a qualitative study with a quantitative sub-study (measures), or to better understand a specific issue of a quantitative study using a qualitative sub-study. E.g confirm, cross-validate, or corroborate findings by bringing together an impact evaluation and a qualitative component.	
Overall score integration	*Put overall integration score here*
Weighted quantitative score (Overall quant score *10)	
Weighted qualitative score (Overall qual score *13)	
Total score out of 260 (weighted quantitative score + weighted qualitative score)	

Appendix II. Data for review of mixed methods impact evaluations.

Figure 1: Quantitative study design

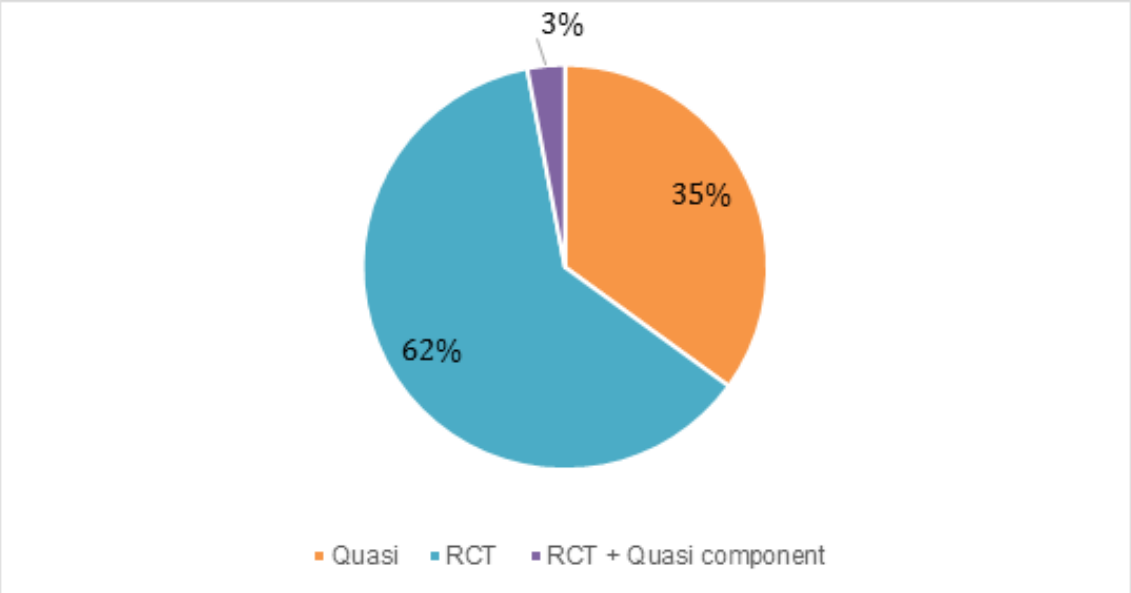


Figure 2: Use of qualitative techniques

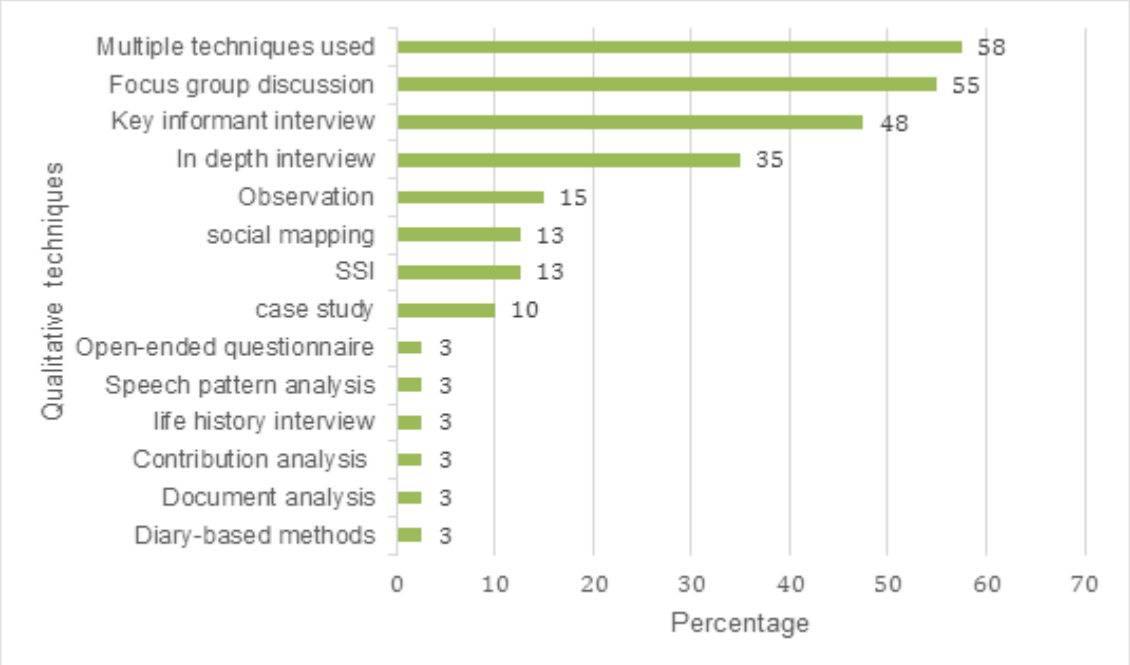


Figure 3: Distribution of studies by sector and country income classification

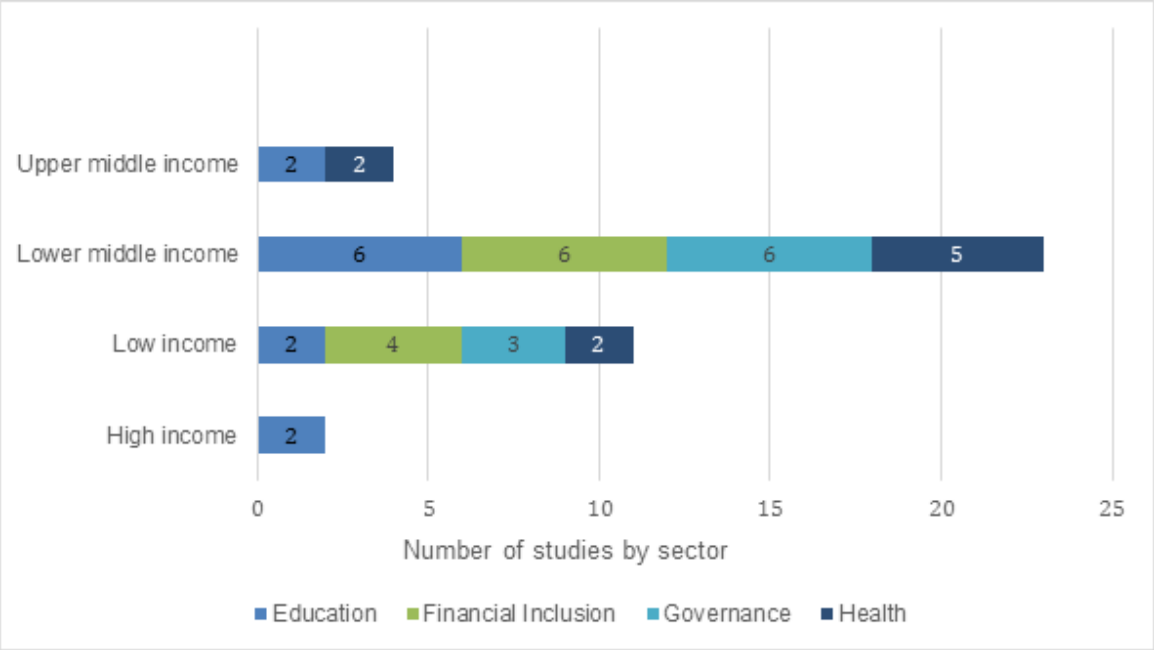


Figure 4: Integration of methods

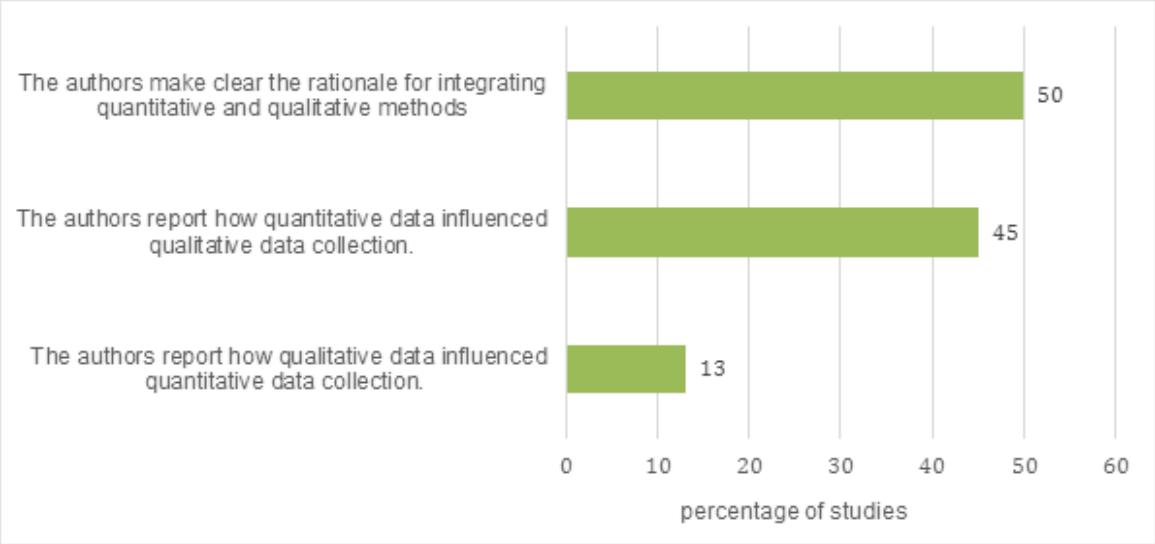


Figure 5: Integration of findings

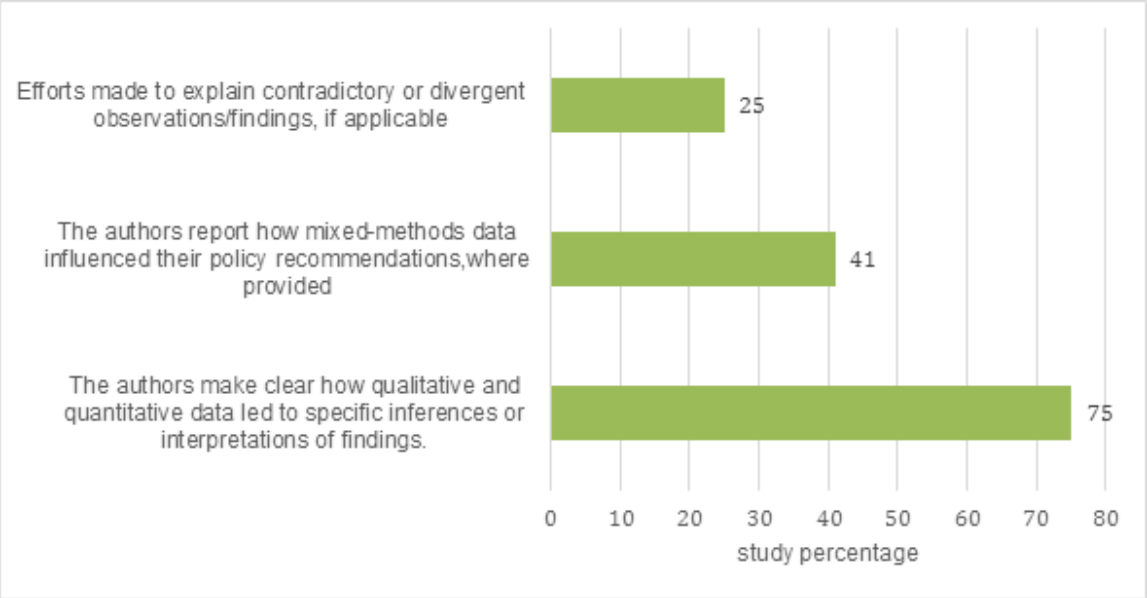


Figure 6: Overall percentage scores by sections A, B and C

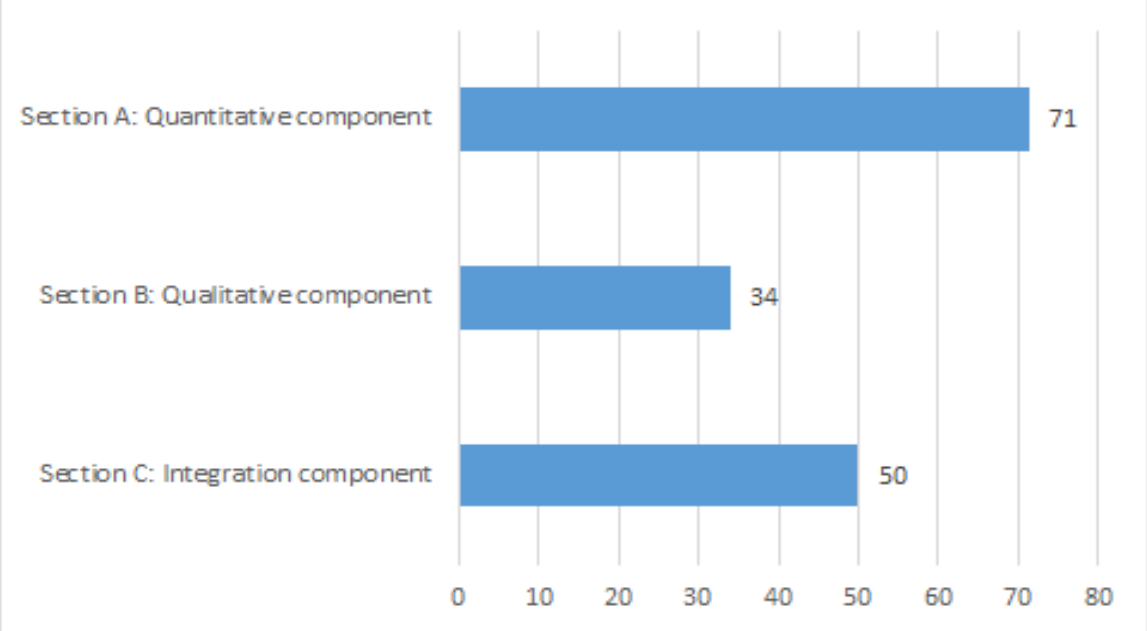


Figure 7: Average scores for part A: Rigour in quantitative methods

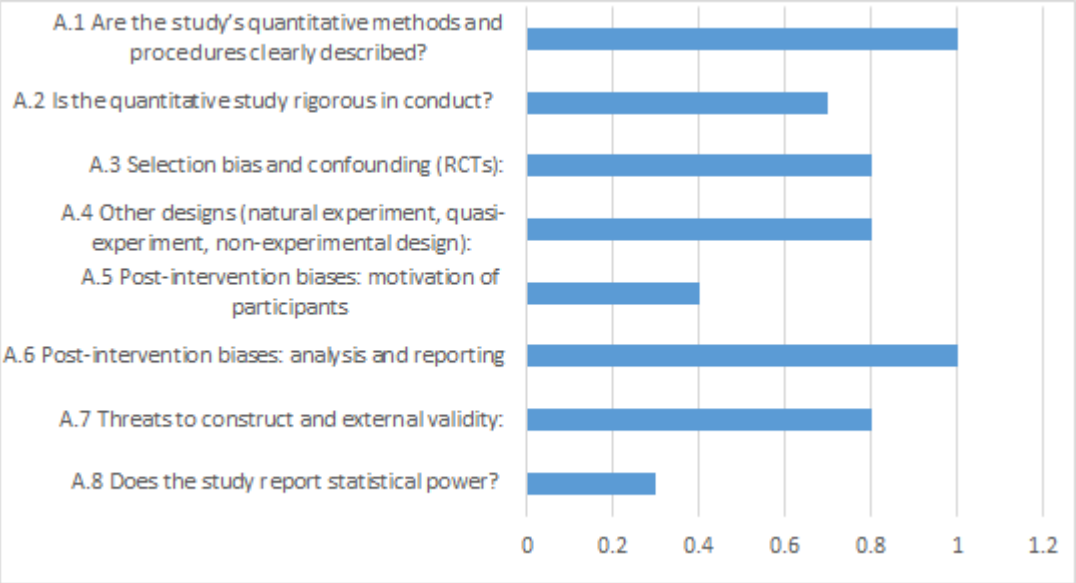


Figure 8: Average scores for part B: Rigor in qualitative methods

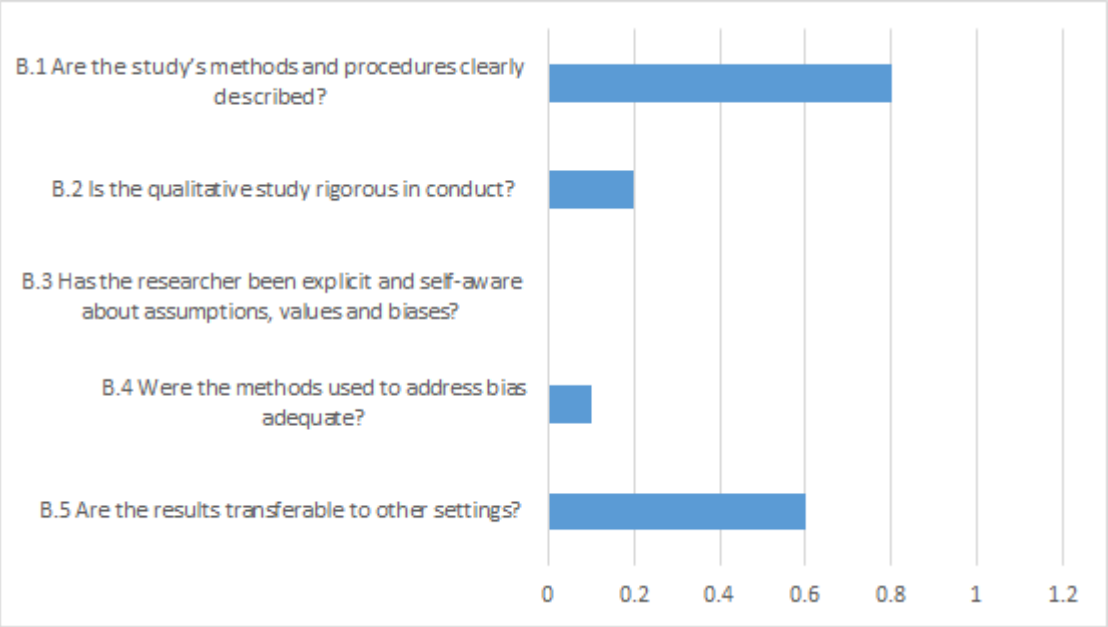


Figure 9: Average scores for part C: The integration of mixed methods

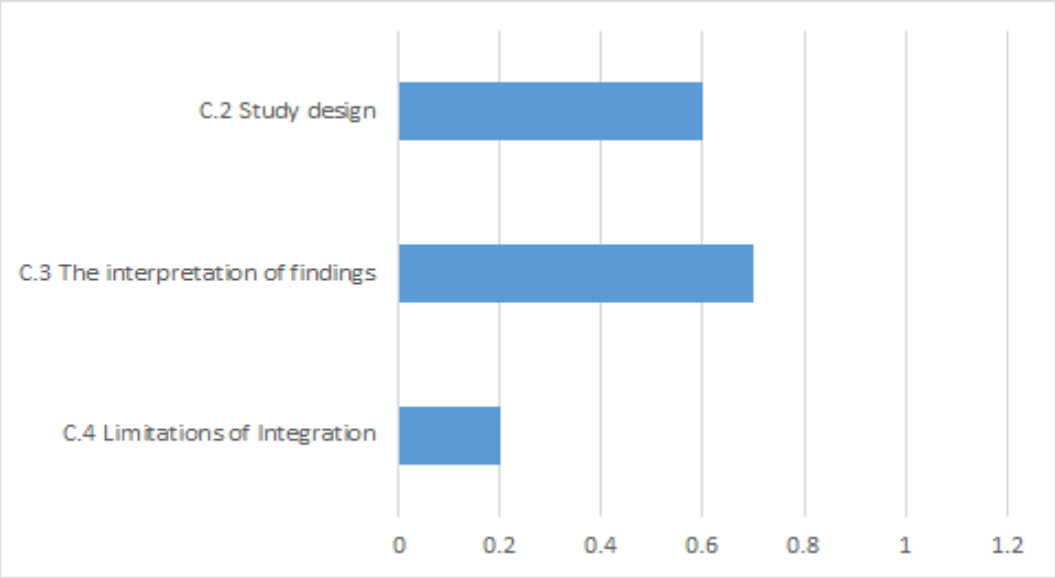
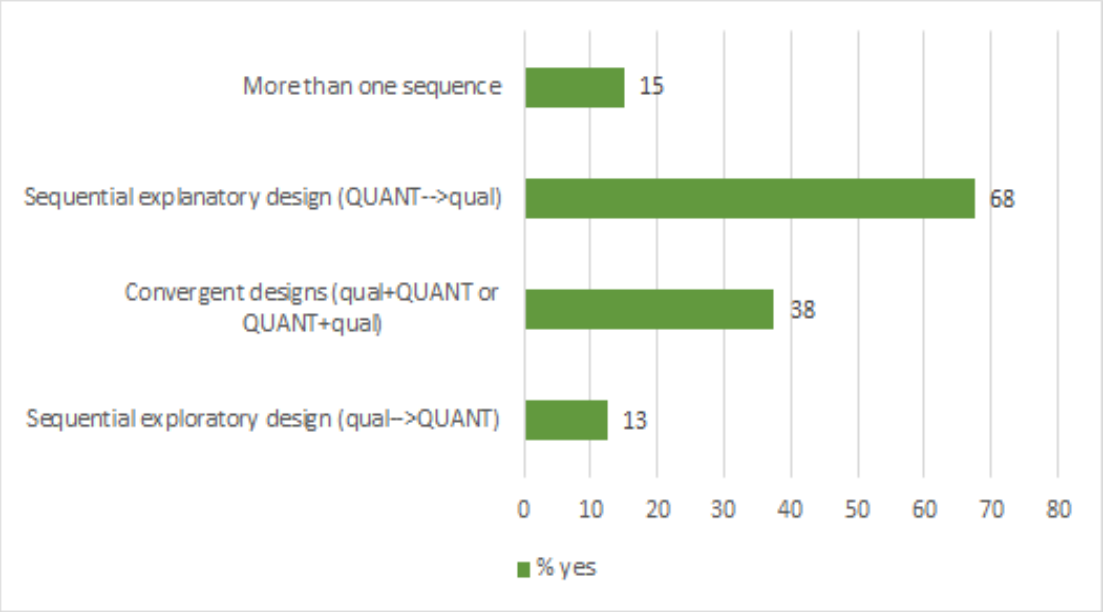


Figure 10: Distribution of studies by the categorisation of mixed methods



Appendix III. Complete tool for mixed methods systematic reviews.

Critical appraisal of mixed methods systematic reviews[11]

Review identifying information (author, year):
Provide a full reference for all articles and reports used to complete this tool (e.g. systematic review technical report, systematic review protocol, systematic review summary report, journal article):
Date of appraisal:
Name of coder:
Date of the last search for studies in review:

Part 1: Critical appraisal of systematic review conduct and reporting

Section A: *Methods used to identify and include studies*

A.1 Did the report contain an explicit statement that the review methods were established in a study protocol and did the report indicate whether there were any deviations from the protocol? o The authors refer to a published protocol o The authors indicate that a protocol or guide was produced but this is not publicly available o The authors indicate any deviations from protocol explicitly (even if just to say that 'there were no deviations from protocol')	o Yes o Partially o No *Coding guide:* *YES: The authors refer to a published protocol and report any deviations from protocol.* *PARTIALLY: A protocol/plan is mentioned but not formally referenced or available, or deviations from protocol indicated.* *NO: All other.*
Comments (note important limitations or uncertainty):	

[11] We have drawn on the following sources in developing this assessment framework: 1) Supporting the Use of Research Evidence (SURE) Collaboration. SURE checklist for making judgements about how much confidence to place in a systematic review (Lewin et al., 2009); 2) AMSTAR2 (Shea et al., 2016); 3) Langer (2017).

A.2 Were review questions reported transparently (primary and secondary questions)? Did the authors specify a separate review question for review component: o Question(s) about intervention design and/or implementation processes such as inputs, activities and outputs (e.g. implementation fidelity) o Question(s) about participant or practitioner views such as on targeting, uptake and adherence (e.g. barriers and enablers/facilitators, willingness to pay) o Question(s) about intervention effects on intermediate and final/endpoint outcomes o Other question(s) (specify)	o Yes o Partially o No *Coding guide - check the answers above* *YES: For any aspects of the causal chain answered using systematic evidence, the review should have clear questions.* *NO: Only the question about intervention effects is asked.* *PARTIALLY: Some questions are specified but not for all parts of the causal chain on which the study claims to report systematic evidence.*
Comments (note important limitations or uncertainty)	
A.3 Were the criteria for deciding which studies to include in the review reported? Did the authors specify: o Participants/ settings/ population o Intervention(s) o Outcome(s) o Study types included (for examples see A.4) o Other (specify)	o Yes o Partially o No *Coding guide - check the answers above* *YES: PIOS should all be mentioned* *NO: Any of P, I, and O or S are not mentioned* *PARTIALLY: PIO are mentioned but S is not.*
Comments (note important limitations or uncertainty)	

A.4 Were study design criteria reported for review sub-components (primary and secondary questions)? Did the authors specify study design criteria for answering different questions along causal chain: o Types of studies for answering questions about intervention design and/or implementation processes, such as inputs, activities and outputs (e.g. process evaluation) o Types of studies for answering questions about participant or practitioner views such as on targeting, uptake and adherence (e.g. ethnographic studies, rapid appraisal methods e.g. participatory rural appraisal) o Types of studies for answering questions of intervention effects on intermediate and final/endpoint outcomes (e.g. RCTs, non-randomised counterfactual-based studies) o Types of studies for answering other questions (specify)	o Yes o Partially o No *Coding guide - check the answers above* *YES: For any aspects of the causal chain answered using systematic evidence, the review should have clear study design criteria and boxes ticked accordingly* *NO: No boxes are ticked* *PARTIALLY: Study types for some questions are specified but not others.*
Comments (note important limitations or uncertainty)	
A.4 Was the search for evidence reasonably comprehensive? Were the following done: o Language bias avoided (no stated restriction of inclusion based on language e.g. only English language studies are included) o No restriction of inclusion based on publication status o Relevant databases searched (Minimum criteria: All reviews should search at least one source of grey literature such as Google; for health: Medline/ Pubmed + Cochrane Library; for social sciences IDEAS + at least one database of general	o Yes o Partially o No o Can't tell *Coding guide - check the answers above:* *YES: All five should be yes* *PARTIALLY: Relevant databases and reference lists are both reported* *NO: Any other*

social science literature and one subject specific database) o Reference lists in included articles checked o Authors/experts contacted	
Comments (note important limitations or uncertainty)	
A.5 Does the review cover an appropriate time period? *Is the search period comprehensive enough that relevant literature is unlikely to be omitted?*	o Yes o Can't tell (only use if no information about time period for search) o No oUnsure *Coding guide:* *YES: Generally this means searching the literature at least back to 1990, but some reviews are able to argue convincingly why searches are more restricted (e.g. updates, evidence from related reviews)* *NO: The search does not go at least back to 1990 and the review does not argue reasonably why restrictions are made* *CAN'T TELL: No information about time period for search* *Note: With reference to the above – there may be important reasons for adopting different dates for the search, e.g. depending on the intervention. If you think there are limitations with the timeframe adopted for the search which have not been noted and justified by the authors, you should code this item as a NO and specify your reason for doing so in the comment box below. Older reviews should not be downgraded, but the fact that the search was conducted some time ago should be noted in the quality assessment. Always report the time period for the search in the comment box, as well as the date of latest search in the summary box on page 1.*

Comments (note search period, any justification provided for the search period, or uncertainty)	
A.6 Was bias in the selection of articles avoided? Did the authors specify: o Independent screening of full text by at least two reviewers o List of included quantitative studies provided o List of included qualitative studies provided, or indication of studies from which qualitative evidence collected (if relevant) o List of quantitative studies excluded at full text provided o List of qualitative studies excluded at full text provided (only relevant if separate searches undertaken for qualitative evidence)	o Yes o Partially o No *Coding guide:* *YES: All four should be yes for reviews incorporating quantitative and qualitative evidence systematically For reviews that draw on qualitative evidence non-systematically, only three should be yes. For reviews published in journals with wordcount restrictions, supplementary files should be available on the journal website and/or the missing information contained in a report available online.* *PARTIALLY: Independent screening and list of included studies provided are both reported* *NO: All other. If list of included studies is provided, but the authors do not report whether or not the screening has been done by two reviewers, the review is downgraded to NO.*
Comments (note important limitations or uncertainty):	
A.7 Were methods used to code studies in order to avoid incorporating dependent findings into any single analysis? Did the authors specify: o Methods to address dependency in findings at the between study level (e.g. multiple publications of the same report, or on the same dataset)? o Methods to address dependency in findings within individual studies (e.g. multiple outcomes reported, or different specifications reported in analysis or multiple follow-ups)	o Yes o Partially o No *Coding guide:* *YES: both are ticked* *PARTIALLY: one box ticked* *NO: no boxes ticked.*
Comments (note important limitations or uncertainty):	

A.8 Overall – how much confidence do you have in the methods used to identify, include and critically appraise studies? *Summary assessment score A relates to the 7 questions above.* *High confidence applicable when the summary answer to the questions in section A are all assessed as 'yes'* *Low confidence applicable when any of the following are assessed as 'NO' above: not reporting study protocol (A1), not reporting explicit selection criteria (A2), not conducting reasonably comprehensive search (A4), not avoiding bias in selection of articles (A6), and not reporting methods to address dependency in findings (A7)* *Medium confidence applicable for any other*	o Low confidence (limitations are important enough that the results of the review are not reliable) o Medium confidence (limitations are important enough that it would be worthwhile to search for another systematic review and to interpret the results of this review cautiously, if a better review cannot be found) o High confidence (only minor limitations)
Comments (note important limitations).	

Section B: ***Methods used to analyse the findings in the quantitative analysis***

B.1 Were the characteristics and results of the included studies reliably reported? Was there: o Independent data extraction by at least two reviewers o A table or summary of the characteristics of the participants, interventions and outcomes for the included studies o A table or summary of the results of all the included studies	o Yes o No o Partially o Not applicable (e.g. no included studies) *Coding guide:* *YES: All three should be yes* *PARTIALLY: Criteria one and three are yes, but some information is lacking on second criteria, or independent data extraction performed for only a sub-sample of studies.* *No: None of these are reported. If the review does not report whether data was independently extracted by 2 reviewers (possibly a reporting error), downgrade to NO.* *NOT APPLICABLE: if no studies/no data*
Comments (note important limitations or uncertainty)	

B.2 Did the authors use appropriate criteria to analyse the risk of bias in the studies that are included? o The criteria used for assessing risk of bias were reported o A table or summary of the assessment of each included study for each criterion was reported o Criteria were used that focus on the quality of study implementation/ risk of bias (and not other aspects of the studies, such as precision, construct validity or external validity). "Sensible" is defined as a recognised quality appraisal tool/ checklist, or similar tool which critically appraises methodological implementation of included studies (not just design characteristics), such as control for confounding, selection bias, bias in outcomes data measurement and bias in reporting.	o Yes o Partially o No o Not applicable *Coding guide:* *YES: All three should be yes* *PARTIALLY: The first and third criteria should be reported. If the authors report the criteria for assessing risk of bias and report a summary of this assessment for each criterion, but the criteria may be only partially sensible (e.g. do not address all possible risks of bias, but do address some), we downgrade to PARTIALLY.* *NO: Any other*
Comments (note important limitations or uncertainty)	
B.3 Are the methods used by the review authors to analyse the findings of the included studies clear, including methods for calculating effect sizes if applicable?	o Yes o Partially o No o Not applicable (e.g. no studies or no data) *Coding guide:* *YES: Methods used clearly reported. If it is clear that the authors use narrative synthesis, they don't need to say this explicitly.* *PARTIALLY: Some reporting on methods but lack of clarity* *NO: Nothing reported on methods* *NOT APPLICABLE: if no studies/no data*
Comments (note important limitations or uncertainty)	

B.4 Did the review describe the extent of heterogeneity? o Did the review discuss the extent to which there were important differences in the results of the included studies? o If a meta-analysis was done, was the I^2 and other appropriate statistic reported (Tau^2)?	o Yes o Partially o No o Not applicable (e.g. no studies or no data) *Coding guide:* *YES: First category should be yes, and second category should be yes if applicable* *PARTIALLY: The first category is yes* *NO: Any other* *NOT APPLICABLE: if no studies/no data*
Comments (note important limitations or uncertainty)	
B.5 Were the findings of the relevant studies combined (or not combined) appropriately relative to the primary question the review addresses and the available data? How was the data analysis done? o Descriptive only o Vote counting based on direction of effect and/or statistical significance o Description of range of effect sizes o Statistical meta-analysis or meta-regression of effect sizes o Other: specify o Not applicable (e.g. no studies or no data) How were the studies weighted in the analysis? o Equal weights (e.g. this is what is done when vote counting is used) o By quality or study design (this is rarely done)	o Yes o Partially o No o Not applicable (e.g. no studies or no data) o Can't tell *Coding guide:* *YES: If analysis based on effect size data (e.g. statistical meta-analysis), appropriate weights and unit of analysis errors addressed (if appropriate).* *PARTIALLY: If appropriate table, graph or meta-analysis and appropriate weights are used, but unit of analysis errors not addressed (and should have been).* *NO: If vote counting is used where quantitative analyses would have been possible.* *NOT APPLICABLE: if no studies/no data* *CAN'T TELL: if unsure (note reasons in comments below)*

o Inverse variance (this is what is typically done in a meta-analysis) o Number of participants (sample size) o Other: specify o Not clear o Not applicable (e.g. no studies or no data) Did the review address unit of analysis errors? o Yes - took clustering of participants within the study into account in the analysis (e.g. used intra-cluster correlation coefficient to calculate effect size standard errors) o No, but acknowledged problem of unit of analysis errors o No mention of issue o Not applicable - no clustered trials or studies included	
Comments (note important limitations or uncertainty)	
B.6 Does the review report evidence appropriately? o The review makes clear which evidence is subject to low risk of bias in assessing causality (attribution of outcomes to intervention), and which is likely to be biased, and does so appropriately o Where studies of differing risk of bias are included, results are reported and analysed separately by risk of bias status	o Yes o No o Partially o Not applicable *Coding guide:* *YES: Both criteria should be fulfilled (where applicable)* *NO: Criteria not fulfilled* *PARTIALLY: Only one criteria fulfilled, or when there is limited reporting of quality appraisal (the latter applies only when inclusion criteria for study design are appropriate)* *NOT APPLICABLE: No included studies*

	Note on reporting evidence and risk of bias: For reviews of effects of 'large n' interventions, experimental and quasi-experimental designs should be included (if available). For reviews of effects of 'small n' interventions, designs appropriate to attribute changes to the intervention should be included (e.g. pre-post with assessment of confounders)
Please specify included study designs and any other comments (note important limitations or uncertainty):	
B.7 Did the review examine the extent to which specific factors might explain differences in the results of the included studies? o Were factors that the review authors considered as likely explanatory factors clearly described pre-hoc (in the protocol) and reported as being identified in protocol? o Were any factors that the review authors considered as likely explanatory factors that were identified post-hoc (after the protocol) clearly indicated as such? o Was a sensible method used to explore the extent to which key factors explained heterogeneity? o Descriptive/textual o Graphical o Meta-analysis by moderators and sub-groups o Meta-regression o Other	o Yes o Partially o No o Not applicable *Coding guide:* *YES: Explanatory factors clearly described and appropriate methods used to explore heterogeneity* *PARTIALLY: Explanatory factors described but for meta-analyses, sub-group analysis or meta-regression not reported (when they should have been)* *NO: No description or analysis of likely explanatory factors* *NOT APPLICABLE: e.g. too few studies, no important differences in the results of the included studies, or the included studies were so dissimilar that it would not make sense to explore heterogeneity of the results*
Comments (note important limitations or uncertainty):	

B.8 Overall - how much confidence do you have in the methods used to analyse the findings relative to the primary question addressed in the review? *Summary assessment score B relates to the 8 questions in this section, regarding the analysis.* *High confidence applicable when all the answers to the questions in section B are assessed as 'yes'.* *Low confidence applicable when any of the following are assessed as 'NO' above: critical characteristics of the included studies not reported (B1), not describing the extent of heterogeneity (B4), combining results inappropriately (B5), reporting evidence inappropriately (B6).* *Medium confidence applicable for any other: i.e. the "Partial" option is used for any of the 6 preceding questions or questions and/or B.2 and/ or B.3 and/ or B.7 are assessed as 'no'.*	o Low confidence (limitations are important enough that the results of the review are not reliable) o Medium confidence (limitations are important enough that it would be worthwhile to search for another systematic review and to interpret the results of this review cautiously, if a better review cannot be found) o High confidence (only minor limitations)
Use comments to specify if relevant, to flag uncertainty or need for discussion	

Section C: *Methods used to include and analyse qualitative evidence*

C.1 Was the qualitative evidence obtained systematically? Did the authors: o Search systematically for all qualitative evidence to answer the review question? o Search systematically for qualitative evidence relevant to the particular contexts in which evidence of effects is available? o Conduct targeted searches for specific types of evidence (e.g. project and implementation documents)? o Only include qualitative evidence contained in included quantitative study reports?	o Yes o Partially o No *Coding guide - check the answer above* *YES: Systematic searches for qualitative evidence were made* *NO: Only evidence from quantitative reports included* *PARTIALLY: Systematic or targeted searches for qualitative evidence relevant to the contexts in which evidence of effects is available.* *NOT APPLICABLE: No qualitative studies/no data used.*

Comments (note important limitations or uncertainty)	
C.2 Were the characteristics and results of the included studies reliably reported? Was there: o Independent data extraction by at least two reviewers o A table or summary of the characteristics of the participants, interventions and outcomes for the included studies o A table or summary of the results of all the included studies	o Yes o No o Partially o Not applicable (e.g. no included studies) *Coding guide:* *YES: All three should be yes* *PARTIALLY: Criteria 1 and 3 are yes, but some information is lacking on second criteria, or partial data extraction done for a sub-set of included studies.* *No: None of these are reported. If the review does not report whether data was independently extracted by two reviewers (possibly a reporting error), code as NO.* *NOT APPLICABLE: if no studies/no data*
Comments (note important limitations or uncertainty)	
C.3 Does the review incorporate qualitative evidence appropriately? Is qualitative evidence incorporated: o To provide evidence on the lower reaches of the causal chain (e.g. on project design, targeting, implementation, participant views) or factual contextual information o To provide evidence suggestive of changes in outcomes which are not measurable quantitatively (e.g. adverse outcomes, outcomes that are not measurable quantitatively)? o To provide evidence suggestive of changes in intermediate and/or final/endpoint outcomes (e.g. evidence for particular sub-groups of participants where sample size insufficient for quantitative evidence to be used)? o To provide evidence on intervention	o Yes o No o Partially o Not applicable *Coding guide:* *YES: Criteria one or two only. Only to assess lower reaches of causal chain, provide contextual information, or provide evidence suggestive of outcomes.* *PARTIALLY: Criteria one or two are yes, and three if reported but it is clear that this evidence is associational not causal.* *No: Criteria four is reported.* *NOT APPLICABLE: if no studies/no data*

outcomes (e.g. impacts on intended intermediate and final/endpoint outcomes for all participant groups)?	
Comments (note important limitations or uncertainty)	
C.4 Did the authors use appropriate criteria to critically appraise included studies? o The criteria used for critical appraisal were reported. o A table or summary of the assessment of each included study for each criterion was reported o Criteria were used that focus on quality of implementation of the study. This means a recognised quality appraisal tool/ checklist (e.g. SURE qualitative evidence checklist, Thomas and Harden, 2008) or a similar tool which critically appraises methodological implementation of included studies (not just design characteristics) – such as approach to sampling, analysis, reporting, and reaching conclusions.	o Yes o Partially o No *Coding guide:* *YES: All three should be yes* *PARTIALLY: The first and third criteria should be reported. If the authors report the criteria for assessing risk of bias and report a summary of this assessment for each criterion, but the criteria may be only partially sensible (e.g. do not address all possible risks of bias, but do address some), we downgrade to PARTIALLY.* *NO: Any other*
Comments (note important limitations or uncertainty)	
C.5 Are the methods used by the review authors to report and analyse the findings of the included studies clear? How were the data reported and analysed? (tick all that apply) o Content analysis of qualitative evidence o Presentation of quotes o Presentation of analysis by themes (e.g. thematic synthesis) o Interpretative synthesis of findings across studies (e.g. meta-ethnography) o Other method (specify)	o Yes o Partially o No o Not applicable (e.g. no studies or no data) Coding guide: YES: Methods used clearly reported. PARTIALLY: Some reporting on methods but lack of clarity NO: Nothing reported on methods NOT APPLICABLE: if no studies/no data

o Not applicable (e.g. no studies or no data)	
Comments (note important limitations or uncertainty)	
C.6 Does the review report evidence appropriately? o The review makes clear which evidence is of high credibility, and which is likely to be of low credibility, and does so appropriately o Where studies of differing quality are included, results are reported and analysed separately by quality status	o Yes o No o Partially o Not applicable *Coding guide:* *YES: Both criteria should be fulfilled (where applicable)* *NO: Criteria not fulfilled* *PARTIALLY: Only one criteria fulfilled, or when there is limited reporting of quality appraisal (the latter applies only when inclusion criteria for study design are appropriate)* *NOT APPLICABLE: No included studies*
Comments (note important limitations or uncertainty):	
C.7 Overall - how much confidence do you have in the methods used to analyse the findings relative to the primary question addressed in the review? *Summary assessment score C relates to the 3 questions in this section, regarding the analysis.* *High confidence applicable when all the answers to the questions in section C are assessed as 'yes'.* *Low confidence applicable when any of the following are assessed as 'NO': C.2, C.3, C.4, C.5 or C.6.* *Medium confidence applicable for any other: i.e. the "Partial" option is used for any of the preceding questions or question C.1 is marked as "No".*	o Low confidence (limitations are important enough that the results of the review are not reliable) o Medium confidence (limitations are important enough that it would be worthwhile to search for another systematic review and to interpret the results of this review cautiously, if a better review cannot be found) o High confidence (only minor limitations)
Use comments to specify if relevant, to flag uncertainty or need for discussion	

Part 2: Integration of evidence and overall assessment

Section D: *Methods used to analyse the causal chain and reach conclusions*

D.1 Does the review use a programme theory? Did the authors present: o A logic model articulating the intervention causal chain from inputs/activities through to outcomes? o Assumptions, either in the logic model itself or discussed in supporting text? o A (middle-range) theory (economic theory, e.g. trade theory; social theory, e.g. diffusion theory; etc.) which informed the logic models and/or from which inferences can be made about mechanisms and contexts under which outcomes might occur?	o Yes o No o Partially o Not applicable *Coding guide:* *YES: Some theory is used, whether an intervention level logic model or causal chain, or formal theory, and underlying assumptions are explicitly described.* *NO: None are reported.* *PARTIALLY: A theory is used but underlying assumptions are not reported.*
Comments (note important limitations or uncertainty)	
D.2 Does the review incorporate qualitative evidence in the design? o Was the logic model or theory articulated at protocol stage? o Is qualitative evidence cited in the development or explanation of the theoretical approach/logic model? o Is qualitative evidence incorporated systematically (i.e. based on systematic searches), in order to develop the theoretical approach/logic model? o Is qualitative evidence cited in development of relevant questions?	o Yes o No o Partially o Not applicable *Coding guide:* *YES: At least 1 and 2 or 3 are reported.* *NO: None are reported.* *PARTIALLY: 1 or 4 are reported.*
Comments (note important limitations or uncertainty)	

D.3 Did the review conduct analysis of intermediate and endpoint outcomes along causal chain? o Did the review conduct analysis of primary endpoint outcomes? o Did the review conduct analysis of primary intermediate outcomes? o Did the review conduct analysis of secondary endpoint outcomes? o Did the review conduct analysis of secondary intermediate outcomes? Note: Primary outcomes are outcomes that must be reported regardless of any other outcome. Any relevant study including a primary outcome is eligible for inclusion in the review. Secondary outcomes are outcomes that are only reported if primary outcomes are also reported. Any relevant study including a secondary outcome but not a primary outcome is ineligible for inclusion.	o Yes o Partially o No o Not applicable (e.g. no studies or no data) o Can't tell *Coding guide:* *YES: Boxes 1 and 2 are ticked* *PARTIALLY: Boxes 1 and 4 or 2 and 3 are ticked.* *NO: analysis of outcomes along causal chain is not undertaken and only endpoint outcomes are analysed (and outcomes at different stages of the causal chain were excluded).* *NOT APPLICABLE: if no studies/no data* *CAN'T TELL: if unsure (note reasons in comments below)*
Comments (note important limitations or uncertainty)	
D.4 Does the review incorporate qualitative evidence in the analysis? o To answer specific review questions about intervention design (e.g. project portfolio information) o To answer specific review questions about barriers and facilitators/enablers of implementation (assumptions or risks in the causal chain/ logic model) o To answer specific review questions about adherence or participant views? o To provide information on context for included quantitative studies (e.g. moderators or implementation fidelity)? o To provide evidence on 'middle-range' causal mechanisms or contextual factors (e.g. policy context, second order changes e.g. general equilibrium effects, sustained	o Yes o No o Partially o Not applicable *Coding guide:* *YES: 1, 2, or 3 plus 4 or 5 are reported.* *NO: None are reported.* *PARTIALLY: Any other combination.*

adherence)	
Comments (note important limitations or uncertainty)	
D.5 Does the review incorporate qualitative evidence in other aspects of the analysis? o To assess applicability of synthesised evidence to particular contexts? o To discuss the quantitative findings in light of qualitative evidence, drawing on qualitative evidence collected systematically? o To discuss the quantitative findings in light of qualitative evidence (i.e. in the discussion section only), not drawing on evidence collected systematically?	o Yes o No o Partially o Not applicable *Coding guide:* *YES: 1 or 2 are reported.* *NO: None are reported.* *PARTIALLY: 3 is reported.*
Comments (note important limitations or uncertainty)	
D.6 Does the review integrate the findings from quantitative and qualitative evidence? Is qualitative evidence: o Presented in a separate results section? o Integrated with evidence on effects using qualitative methods (e.g. further iterations of the theory of change) o Integrated with evidence on effects using quantitative methods (e.g. through formal statistical testing of moderators and sub-groups identified in qualitative analysis) o Some other method to integrate the evidence? (specify)	o Yes o No o Partially o Not applicable *Coding guide:* *YES: 1 and 2 or 3 are reported.* *NO: None are reported.* *PARTIALLY: 1 is reported only.*
Comments (note important limitations or uncertainty)	

D.6 Is quantitative and qualitative evidence integrated to form conclusions and implications? o Weight of quantitative evidence assessed using GRADE or other methods o Weight of qualitative evidence assessed using cerQUAL or other methods o Some method is used to integrate findings from quantitative and qualitative approaches (e.g. summary of findings table) to determine conclusions and implications?	o Yes o No o Partially o Not applicable *Coding guide:* *YES: All are reported.* *NO: None are reported.* *PARTIALLY: Only 1, 2 or 3 are reported.*
Comments (note important limitations or uncertainty)	

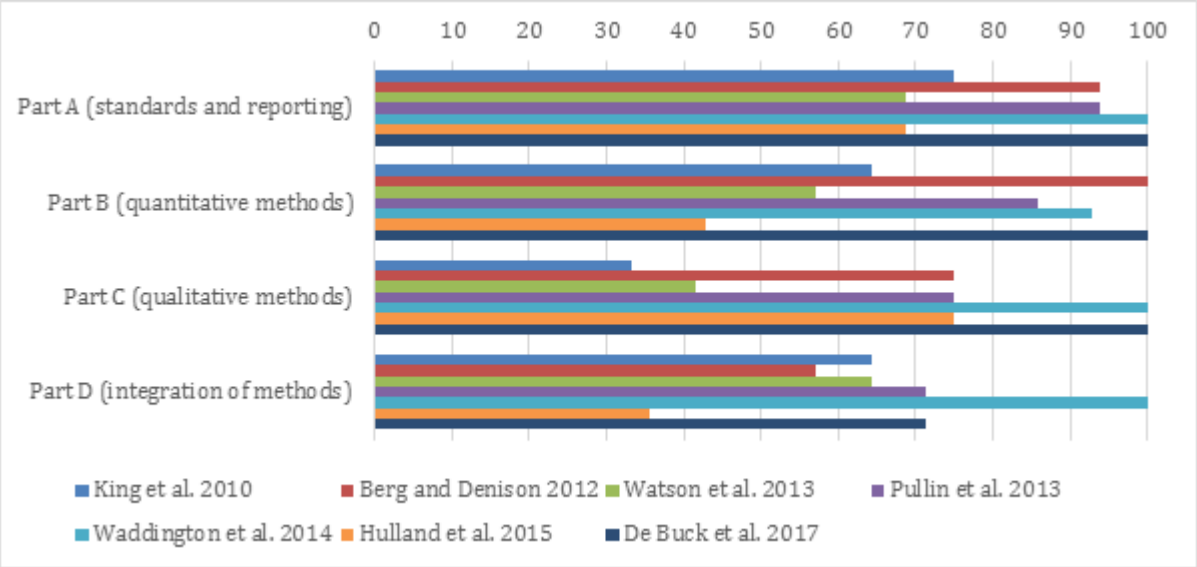
Section E: Overall assessment of the reliability of the review and the incorporation of qualitative evidence

E.1 Are there any other aspects of the review not mentioned before which lead you to question the results?	o Additional methodological concerns – only one person reviewing o Robustness o Interpretation o Conflicts of interest (of the review authors or for included studies) o Other o No other quality issues identified
E.2 Are there any mitigating factors which should be taken into account in determining the reviews reliability?	o Limitations acknowledged o No strong policy conclusions drawn (including in abstract/ summary) o Any other factors
Use comments to specify if relevant, to flag uncertainty or need for discussion	
E.3 Based on the assessments in sections A, B and C of the methods, describe the confidence in the review findings.	
E.4 Categorise the type of mixed methods study (studies can meet multiple criteria): o Sequential explanatory design (QUANT-->qual)	The purpose of *sequential explanatory design* is to explain quantitative results using qualitative findings. E.g., the quantitative results guide the selection of qualitative data sources and data collection, and the qualitative findings

The quantitative component is followed by the qualitative. o Sequential exploratory design (qual-->QUANT) The qualitative component is followed by the quantitative. o Convergent designs where the qualitative and quantitative components are concomitant and integrated. These encompass triangulation (qual+QUANT) or embedded designs (QUANT+qual).	contribute to the interpretation of quantitative results. E.g. Systematic review of effects in which the quantitative analysis is followed by qualitative evidence synthesis to explore applicability of findings in particular contexts. The purpose of *sequential exploratory design* is to explore, develop and test an instrument (or taxonomy), or a conceptual framework (or theoretical model). The qualitative findings inform the quantitative data collection. E.g. Systematic review of effects informed by theory of change developed drawing literature including qualitative evidence at protocol stage. The purpose of *convergent designs triangulation design* is to examine the same phenomenon by interpreting qualitative and quantitative results (bringing data analysis together at the interpretation stage), or by integrating qualitative and quantitative datasets (e.g., data on same cases), or by transforming data (e.g., quantization of qualitative data). E.g. SR of effects using methods to integrate findings from quantitative and qualitative synthesis, e.g. iterative theory of change analysis, meta-regression analysis drawing on moderators identified from qualitative evidenec. The purpose of *embedded/convergent design* is to support a qualitative study with a quantitative sub-study (measures), or to better understand a specific issue of a quantitative study using a qualitative sub-study. E.g. SR of effects with sub-component examining participant views or implementation processes (barriers and facilitators/enabler analysis).
Comments (explain why or note uncertainty)	

Appendix IV. Data for review of mixed methods systematic reviews.

Figure 11 Systematic reviews of effects incorporating qualitative evidence – overall percentage scores



Note: 'yes', 'partially' and 'no' responses were coded as 1, 0.5 and 0, respectively. Percentages of total for each section are reported here (part A includes 8 questions, parts B and D include 7 questions and part C includes 6 questions).

Figure 12 Systematic reviews of effects incorporating qualitative evidence – type of mixed methods approach used

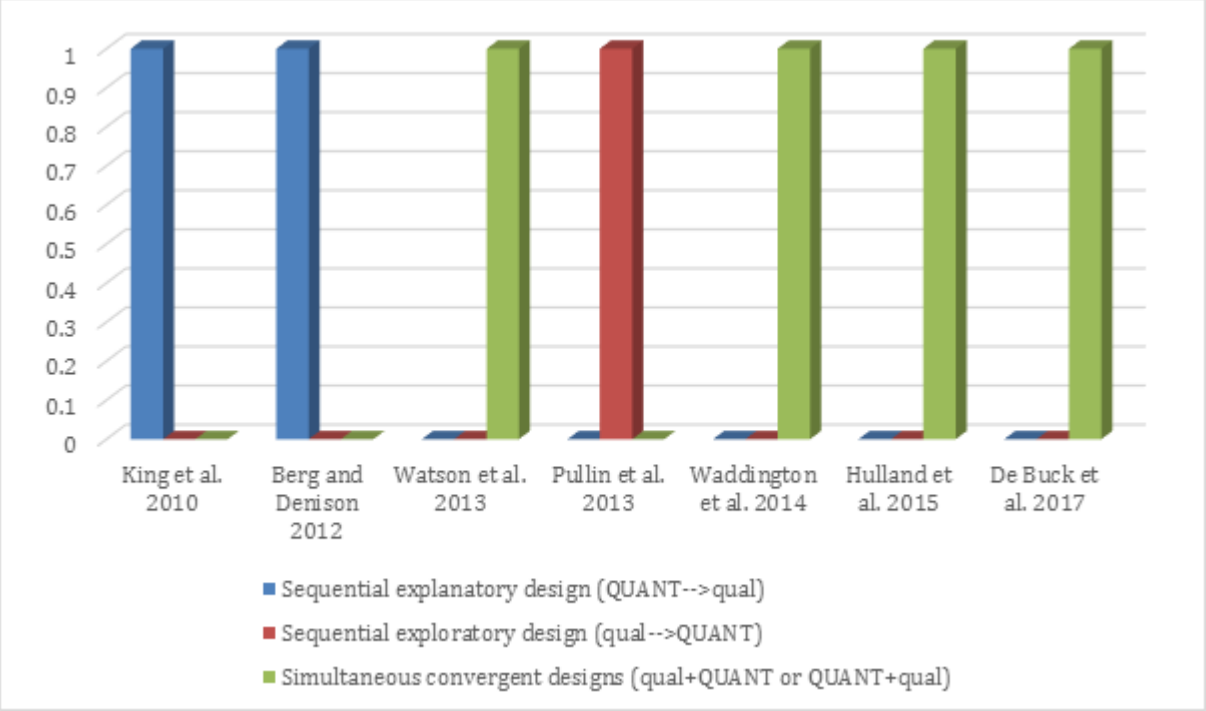


Figure 13 Average scores for part A: rigour of conduct and reporting of SR

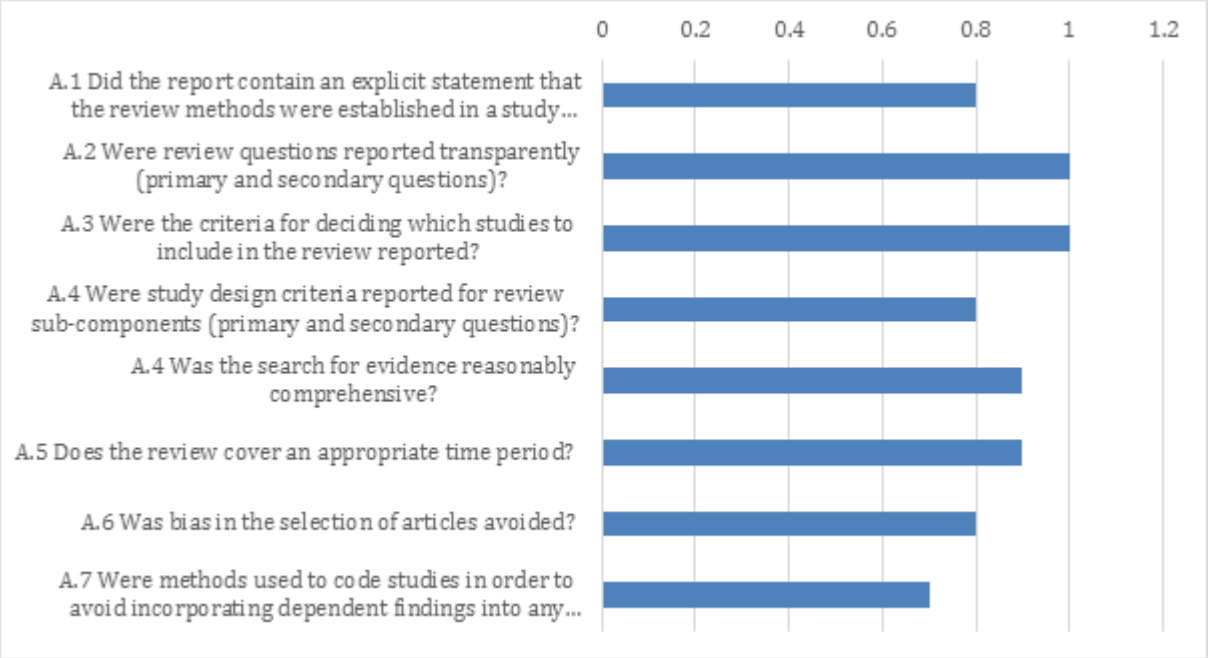


Figure 14 Average scores for part B: rigour of the conduct of quantitative synthesis

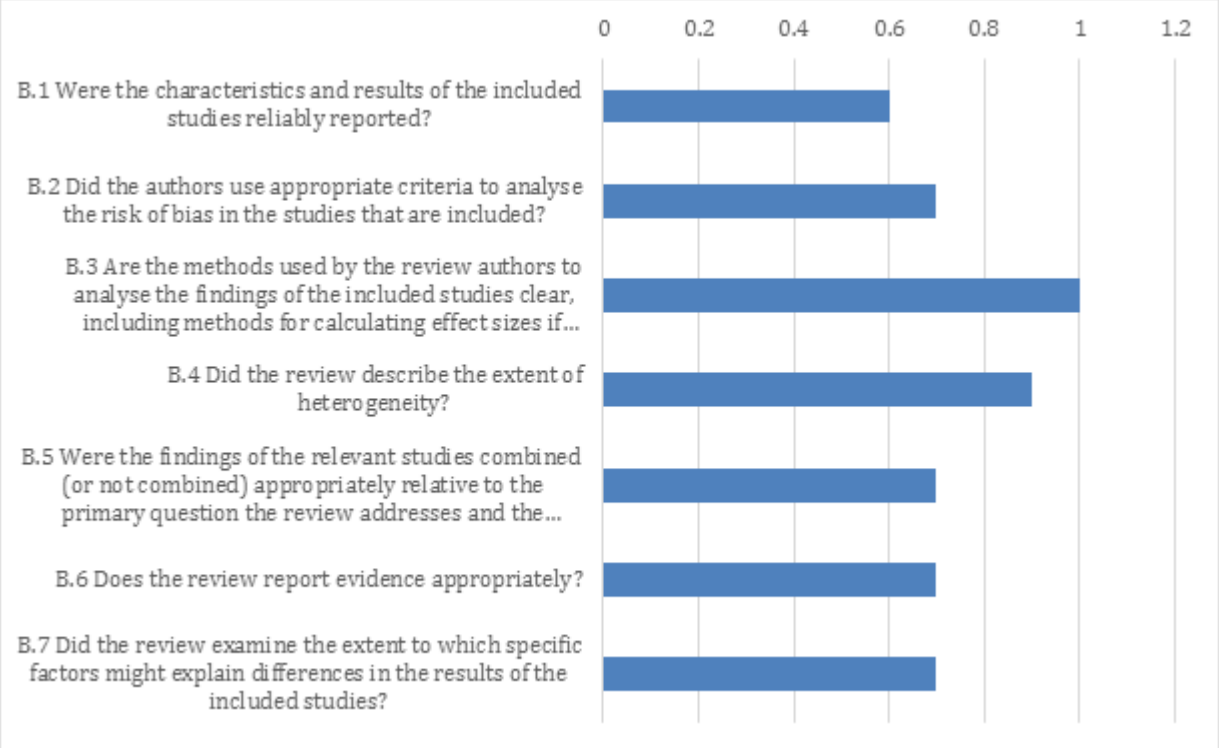


Figure 15 Average scores for part C: rigour of the conduct of qualitative synthesis

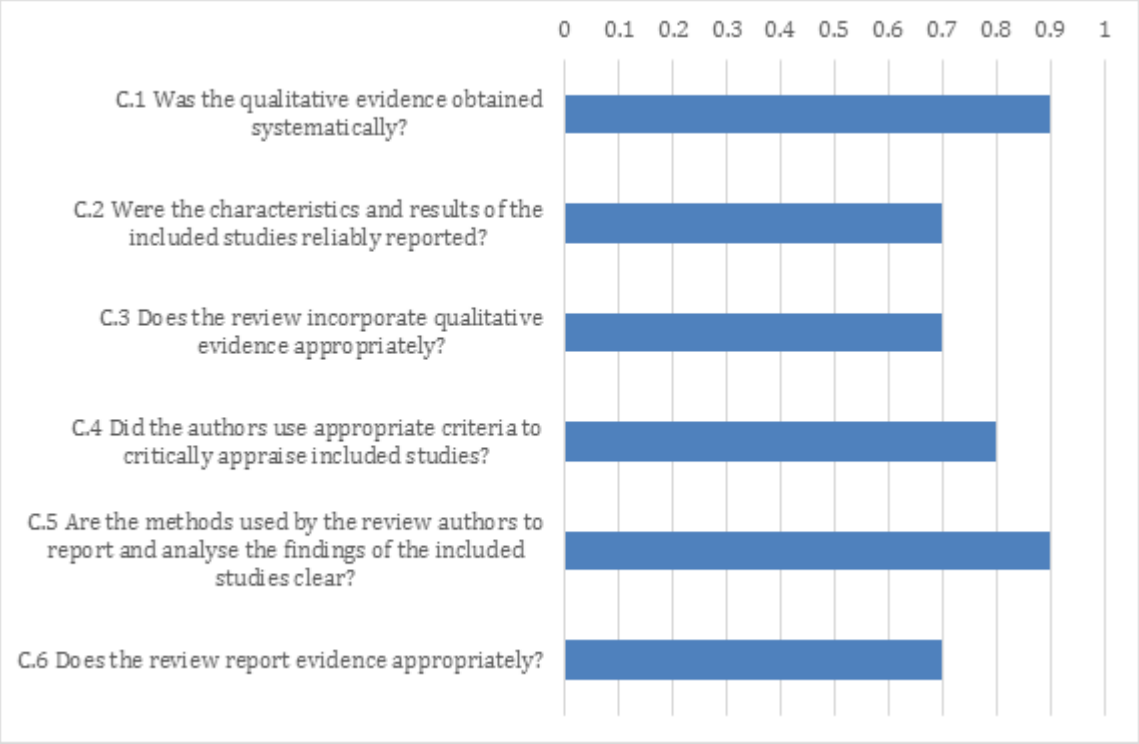
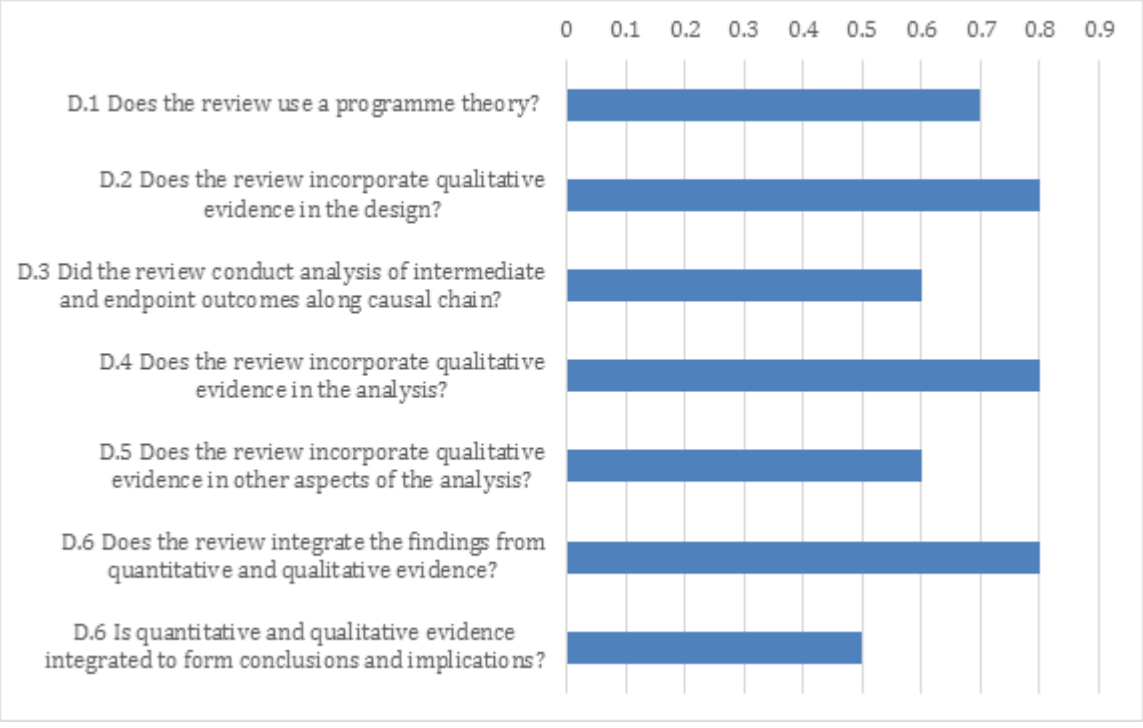


Figure 16 Average scores for part D: integration of quantitative and qualitative evidence



Appendix V. Complete scores for mixed methods impact evaluations.

Rank	Total weighted quant score (A) (out of 130)	Total weighted quant score (B) (out of 130)	Total weighted rigour score (out of 260) (A+B)	Total Integration scores (C) (out of 6)	Funder
1	110	91	201	6.0	DFID
2	120	65	185	6.0	UNICEF
3	115	85	200	5.5	DFID
4	100	91	191	5.0	DFID
5	80	26	106	4.5	WFP
6	115	85	200	4.0	World Bank
7	90	65	155	4.0	IFPRI
8	95	52	147	4.0	DFID
9	105	39	144	4.0	3ie
10	105	39	144	4.0	USAID
11	70	65	135	4.0	DFID
12	90	33	123	4.0	DFID
13	120	65	185	3.5	
14	120	46	166	3.5	
15	105	52	157	3.5	
16	75	39	114	3.5	
17	125	78	203	3.0	
18	85	59	144	3.0	

19	80	52	132	3.0	
20	85	46	131	3.0	
21	75	39	114	3.0	
22	75	33	108	3.0	
23	60	39	99	3.0	
24	95	46	141	2.5	
25	85	33	118	2.5	
26	105	46	151	2.0	
27	105	39	144	2.0	
28	95	33	128	2.0	
29	110	13	123	2.0	
30	70	52	122	2	
31	100	0	100	2.0	
32	85	13	98	2.0	
33	110	26	136	1.5	
34	95	7	102	1.5	
35	90	33	123	1.0	
36	80	33	113	1.0	
37	95	7	102	1.0	
38	110	33	143	0.5	
39	125	0	125	0.5	
40	100	13	113	0.0	

Appendix VI. Complete scores for mixed methods systematic reviews.

Rank	Total weighted score Part A (standards and reporting) (%)	Total weighted score part B (quantitative methods) (%)	Total weighted score part C (qualitative methods) (%)	Total weighted score part D (integration) (%)	Funder
1	100	93	100	100	3ie
2	100	100	100	71	3ie
3	94	86	75	71	3ie
4	94	100	75	57	GEF
5	75	64	33	64	3ie
6	69	57	42	64	DFID
7	69	43	75	36	3ie

Contact us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDIL2018