Centre of Excellence for Development Impact and Learning

March 2022

Design Paper 3

POInT Research Design Paper

Calum Davey, Matthew Juden and POInT team: Elizabeth Allan, Audrey Prost, Macartan Humphreys, Alan Jacobs, Lily Medina, Henry Mwambi Medina, Henry Mwambi

LIDC London International Development Centre

UKaid from the British people

Oxford Policy Management

About CEDIL

The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by the UK government through UK Aid. The mission of the Centre is to test innovative methodologies in evaluation and evidence synthesis and promote evidence-informed development. CEDIL-supported projects fall into three programmes of work: evaluating complex interventions, enhancing evidence transferability, and increasing evidence use.

CEDIL design paper

CEDIL design papers present the protocols of studies supported by the CEDIL programme. They describe the objectives, the methodology and the implementation plan of impact evaluations, evidence syntheses, and methodological papers supported by CEDIL.

About this design paper

This design paper was submitted to CEDIL by the "Putting the theory of change to work: Process-Outcome Integration with Theory" Project L. 292 team.

Please direct any comments or queries to the corresponding author, Calum Davey at calum.davey@lshtm.ac.uk.

Suggested citation: Davey, C., Juden, M., Allan, E., Prost, A., Humphreys, M., Jacobs, A., Medina, L., Mwambi, H. (2022) 'POInT Research Design Paper', CEDIL Design Paper 3. Centre of Excellence for Development Impact and Learning (CEDIL), London and Oxford.

This project was funded by the Centre of Excellence for Development Impact and Learning (CEDIL), supported by UK aid from the UK Government. The views expressed in this research project paper do not necessarily reflect the UK Government's official policies or CEDIL.

POInT Research Design Paper

Calum Davey, Matthew Juden
and POInT team: Elizabeth Allan, Audrey Prost, Macartan Humphreys, Alan Jacobs, Lily Medina, Henry Mwambi

2021-02-26

Table of contents

Table of contents **1**

Policy relevance **3**

Innovation **5**

Methodological exemplar 7

Step 1: theory of change to DAG and priors 8

Figure 1: UPAVAN theory of change 9

Figure 2: UPAVAN arm one candidate DAG 10

Figure 3: UPAVAN arm one candidate DAG with confounding 12

Step 2: identify and link the data with the DAG 13

Step 3: updating the model 14

Step 4: querying the model and reporting 14

Technical design **15**

Case-studies two, three, and four 15

Disability-inclusive graduation (DIG) 15

Intervention 15

Evaluation 17

STAR+ 18

Intervention 18

Evaluation 19

SHARPE 20

Developing middle-range theories / DAGs from theories of change 20

Outcomes and how they will be measured 21

Primary outcome 22
Mediators and moderators 23
Analysis method and establishing causal inference 24
Threats to validity 27
Insufficient power 27
Confounding 28
Figure 3: UPAVAN arm one candidate DAG with confounding 29
Temporality and availability of mediator and moderator variables 30
Advantages of POInT 30
Reinforcing the formulation of high-impact evaluation questions 30
Enabling pre-analysis for a wider set of inferences 31
Using all available data and pre-existing theory to increase confidence in causal claims 31
Exploring heterogeneous effects 32
Targeting transportability 32
Embodying and facilitating interdisciplinarity 33
Ethical approvals and data sharing 34
Reporting 34
Discussion and next steps 35
References 37

Policy relevance

This project will be augmenting the evaluation of four interventions, working alongside the evaluators of these interventions to apply a methodological innovation to improve the inferences made in each case. In brief, the UPAVAN trial is a completed cluster-randomised trial in rural India of an intervention that aimed to improve dietary diversity and weight gain among young children and their mothers; the *Disability inclusive graduation* (DIG) trial is an ongoing cluster-randomised trial of a disability-inclusive version of a well-studied intervention designed to help extremely poor people 'graduate' out of the poverty trap that holds them; the STAR+ trial is starting in 2022 and will evaluate a disability-inclusive version of an apprenticeship and training scheme in Bangladesh; and finally the SHARPE evaluation is a market-systems based intervention focused on access to digital financial services in host and refugee populations in Ethiopia. These are all complex interventions being delivered in various contexts, all funded by the FCDO.

The goal of our project is to develop and implement a flexible and accessible method for combining data from the impact evaluation and data from the process evaluation to make better inferences about the theory of change and the transportability of the results as predictions for other contexts.

The POInT project will make an important contribution to the literature and to evaluation practice. The project aims to unify three strands of modern best-practice evaluation research: theories of change to describe complex interventions (De Silva *et al.*, 2014; Breuer *et al.*, 2015; Davies, 2018), mixed-methods process and impact evaluations (Aarons *et al.*, 2012; Cartwright, n.d.; Creswell *et al.*, 2011; Jimenez *et al.*, 2018), and the use of middle-range theory as the basis for transporting findings to other contexts (Leviton, 2017; Giddens, 1984; Bareinboim & Pearl, 2013). We will do this by developing a procedure to implement an existing method that builds on formal causal inference methods and a Bayesian framework to accommodate various types of data (e.g. qualitative and quantitative) on various aspects of the underlying theory.

The contributions that this project will make fit into three broad literatures. One is the literature on mixing methods, using qualitative and quantitative data together to answer questions. Here there has been a deep discussion of how qualitative and quantitative findings can communicate with one another, what order they should be done in, and what are the epistemological limits set by the two approaches (Creswell *et al.*, 2011). The method that we will be using takes a Bayesian approach to these

epistemological problems, asserting that the various data can be combined through conversion into beliefs and uncertainty about those beliefs.

The second literature to which we will contribute is the analysis of mediation, and the formal causal inference theory that underpins it. Our approach is consistent with modern standards of causal inference (Daniel *et al.*, 2015; Steen *et al.*, 2017; VanderWeele, 2015). It relies on the principle of comparing counterfactuals and describing underlying theory through directed acyclic graphs (DAGs) (Greenland, Pearl & Robins, 1999). The strength of the formal causal inference approach is that it has internal consistency and is grounded by theoretical proofs. However, while the concept of counterfactual thinking is applied consistently in impact evaluation, this application is limited. A more thorough implementation would embed formal causal models into the theory of change. Part of our POInT project will be to adapt theories of change into formal causal models and to demonstrate the strengths and weaknesses of such an approach.

The third literature to which we will contribute is transporting the results of interventions to other places. This is the least developed literature, and previous work on the CEDIL Inception report highlighted that there are a number of deficiencies in the approaches that are available when attempting to learn from evaluations of complex interventions about what to do elsewhere (Davey *et al.*, 2018). A promising approach has emerged recently from the causal-inference framework, essentially defining what can and what cannot be transported from one context to another using the same rules that are at the heart of the causal inference more generally (Bareinboim & Pearl, 2013). A challenge for this approach is that the nodes in the model should have similar meaning between locations, when in reality the actual named elements of an intervention and context will likely differ. Therefore, a degree of abstraction is required so that the DAG can have any generality, which is in keeping with the concept of a middle-range theory.

Although perhaps appearing technical and even esoteric, we aim for the POInT approach to be practical and be used instrumentally by evaluators and users of evidence. This will be facilitated through the application in a number of settings. These are complex and potentially difficult contexts in which to work and to affect change. The settings in India and Uganda are rural with high levels of extreme poverty and reliance on low-intensity agriculture. The project in Bangladesh will cover a number of urban and rural contexts, targeting the ultra poor and also your people without jobs and not in education who have a disability. In Ethiopia, the intervention will be working in an area supporting a large refugee population. Uptake and use of the method will be facilitated by the increasing awareness of the methodological gap

that we are addressing, and the capacity for our method to also address evidence gaps. We hope through application to precisely describe the problem with the ways that various data are used to make inferences about theories of change and in contrast to show how we can go some way to doing more with the various tools we are already using, e.g. process evaluations, theories of change, and mediation analysis.

The focus of the method is to help answer the key question facing policy makers and forward-looking intervention designers: how can we best learn from an evaluation about what to do next time. This is a practical question. Our methodological innovations will target multiple parts of the process of answering this question. For example, we will demonstrate how qualitative and quantitative information can be combined to make optimally-informed inferences about the effects that occurred and the underlying processes that drove those effects. On the way, we will also develop a protocol for specifying the theory of change in terms of a strict causal model with corresponding prior beliefs about the model structure. Making explicit the decisions taken to develop such a causal model will strengthen intervention design and put numbers to the hunches that designers and policy makers have about the likely effects. By combining the prior beliefs about the individual elements of the model (i.e. the arrows between the boxes and how they relate) we can use the same method that we will use to combine the data to query the overall prior belief about the model itself. This allows us to 'stress test' the capacity of the intervention to have the effects that are hoped for when the theory is based on realistic and specific beliefs. While this process is informally done when building and appraising the theories of change behind interventions, ours will be the first approach that can transparently, repeatably, and consistently test theories of change before the evaluation is conducted. We will use the same model to make predictions about what data will be most informative about the effects, underlying causal processes, and transportability of the findings to new contexts.

Innovation

This section of the paper discusses the methodological innovation introduced by the study.

The POInT project will develop a protocol for adapting theories of change into formal causal models. Theories of change will be simplified where necessary to focus on the key research questions. Based on the simplified/focused theory of change we will construct the causal model according to the rules of 'directed acyclic graphs' (DAGs), which are diagrams with boxes and arrows that depict assumptions

about underlying causal processes. However, these models of the assumed causal relations are incomplete without information about the strength and direction of the causal effects, and our confidence in those values. We will therefore work with the implementation teams to discern what their prior beliefs are about the causal processes in the DAGs.

The causal models will be the structure through which various data collected in evaluations can be used to make better-informed inferences. For example, quantitative information on the outcomes will be combined with qualitative data on the potential mediators and moderators. We will combine the data to make these inferences using Bayesian modelling in the statistical software R . A freely downloadable package for R called 'CausalQueries'[1] contains the verified procedures and code to conduct this analysis.

The result will be more precise estimates of effects and learning about the processes and contextual contingencies of interventions' effects.

The POInT project will contribute in some way to all three of CEDIL's programmes of work.

- **(1) Evaluating complex interventions**
    - All of the case studies are complex interventions. The method is designed to work with the theory of change
    - The approach tackles head-on the problem of translating complexity into analytical tractability
    - Helps to understand the epistemological limits of formal analysis of complex interventions
    - Helps develop materials and protocols for more precisely describing the underlying theory and the beliefs
- **(2) Enhancing evidence transferability**
    - Ensuring that findings are transferable
    - More robust inferences about the role of context, whether the markers of context are qualitative or quantitative
    - Combining evidence from different sources (extension)
- **(3) Increasing evidence use**
    - Rigorous description of priors
    - A formal evidence-use cycle through the Bayesian framework

[1] On CRAN at https://cran.r-project.org/web/packages/CausalQueries/index.html

- By requiring formal and specific DAGs and priors at the start, proving an 'in' for evidence into theories of change

Methodological exemplar

As a methodological exemplar and to explain how the methodological innovation will be tested we will describe how the approach will be used in the case of the UPAVAN study (Kadiyala *et al.*, 2018). This vignette is intended to illustrate the mechanics of the method. Although as yet incomplete (the method is being developed throughout the project) the example is also intended to initiate interest among other researchers to adopt and adapt the approach in their own work. The materials and training that we will produce will facilitate uptake.

The UPAVAN trial is the only case study that is already completed. For this reason, some of the steps will be different from the other case studies. In particular, we are not able to ascertain a true set of prior beliefs about the effects of the intervention since the data have been analyzed and reported. We will also be unable to affect data collection. However, by first working retrospectively with a high-quality trial, we will be able to test out other aspects of the method in preparation for the more challenging cases.

The UPAVAN trial assessed the effectiveness of a participatory video-driven agriculture-focused addition to women's self-help groups in Keonjhar district in the state of Odisha in India. In this district 86% of the 1.8 million people work in agriculture. In 2015-16, 30% of women were underweight and 19% of children were wasted (NFHS-4, 2005).

The evaluation design was based around a four-arm cluster-randomised trial (for details see the protocol, Kadiyala *et al.*, 2018). 148 villages were randomly allocated 1:1:1:1 to four arms. The control arm (Arm 0) received the standard assistance available from the government. Treatment arm 1 ('AGRI') received twice monthly videos about locally appropriate agricultural improvements and a follow-up to the women after each video. Treatment arm 2 ('AGRI-NUT') received the same agricultural videos once a month interspersed with nutrition-focused videos also once a month. Women in this arm also received follow-up visits after each video. Treatment arm 3 ('AGRI-NUT+PLA) received the agricultural videos once a month interspersed with the nutrition videos or a participatory-learning-and-action meeting, as well as the same follow up after each video or meeting. In all arms, including control, there was training

of frontline workers. The interventions were delivered over 32 months. All women were eligible to take part in the interventions.

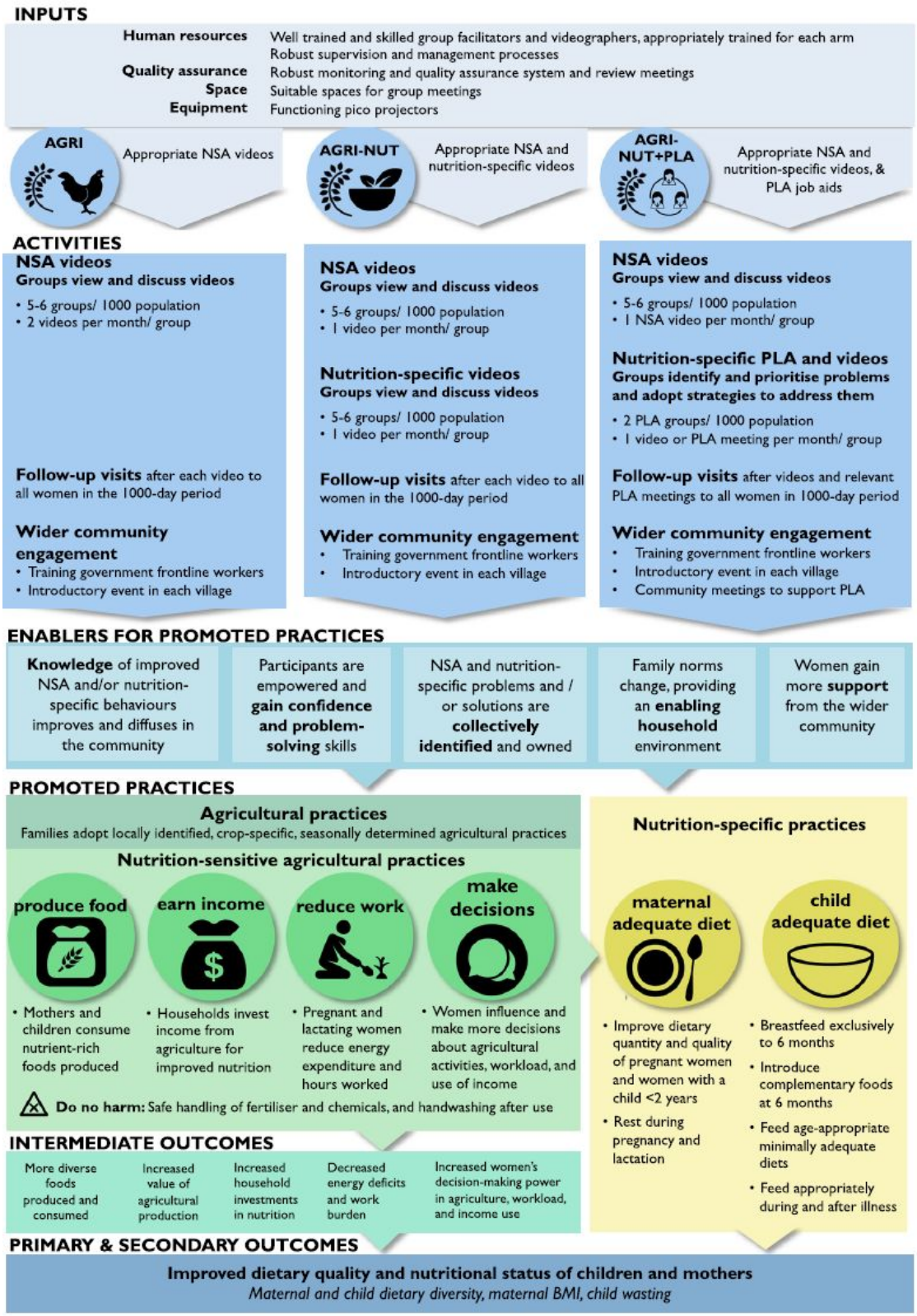
The primary outcomes for the trial were child minimum dietary diversity (the proportion of children aged 6-23 months consuming four or more of seven food groups as recalled from the previous day) and maternal low weight, measured as the BMI among non-pregnant non-postpartum mothers of children aged 0-23. These and the secondary outcomes (maternal minimum dietary diversity, child wasting, and other outcomes along the theory of change) were all measured with repeated cross-sectional surveys of pregnant women and mothers of children less than two years of age. The baseline was in November 2016 to January 2017 and the endline in November 2019 to January 2020. The surveys are linked at the village level, but the individual participants, although overlapping, cannot be reliably linked between the baseline and endline surveys.

A process evaluation assessed the fidelity of the interventions' implementation, clarified the mechanisms behind the intervention effects, and identified contextual factors associated with variation in the outcomes. For this the evaluators used focus groups (17 groups, 181 members in total), monitoring data, 32 family case studies (91 interviews in total), three focus groups with CSPs as well as collating reports and collecting information about the process in the endline survey.

Step 1: theory of change to DAG and priors

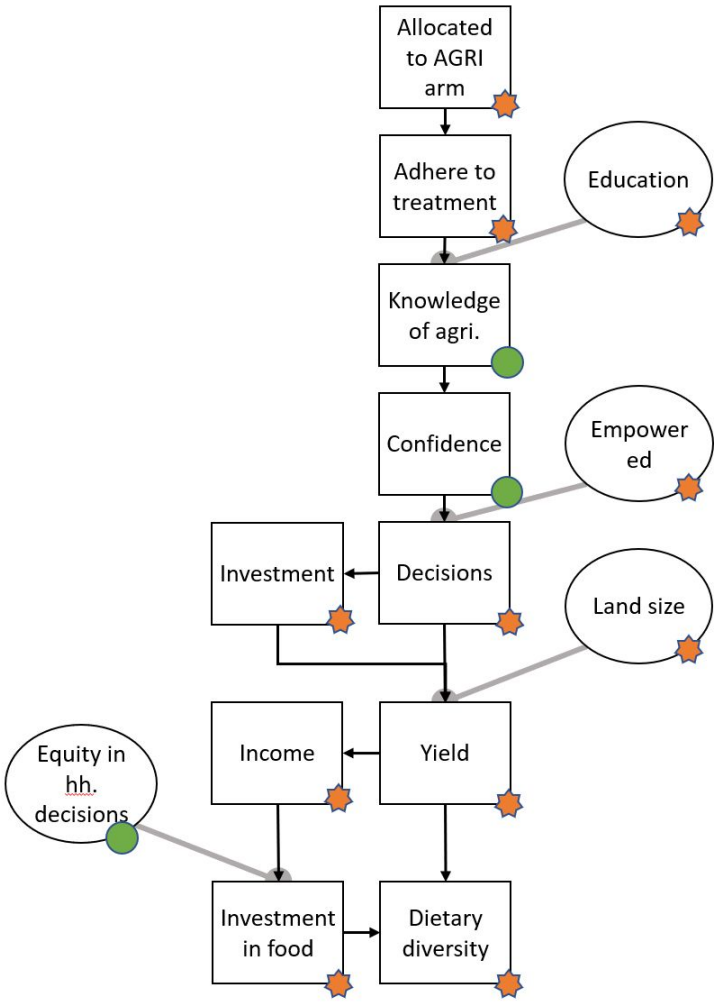
The first step in the process is to convert the theory of change for the intervention into a formal DAG with priors about the underlying causal processes. Fortunately, two members of our team were involved in the original evaluation and are very knowledgeable about the intervention design. The intervention theory-of-change diagram is shown below in **Figure 1**. We are initially going to focus on treatment arm one – the agricultural intervention only. This is the simplest case and will be the basis for more complexity later.

Figure 1: UPAVAN theory of change



The first milestone in applying our method is to generate a DAG that represents information from the theory of change as a step-by-step progression from inputs to the final outcome via any mediators and under the influence of any key moderators. **Figure 2** shows one of our attempts at doing this.

Figure 2: UPAVAN arm one candidate DAG



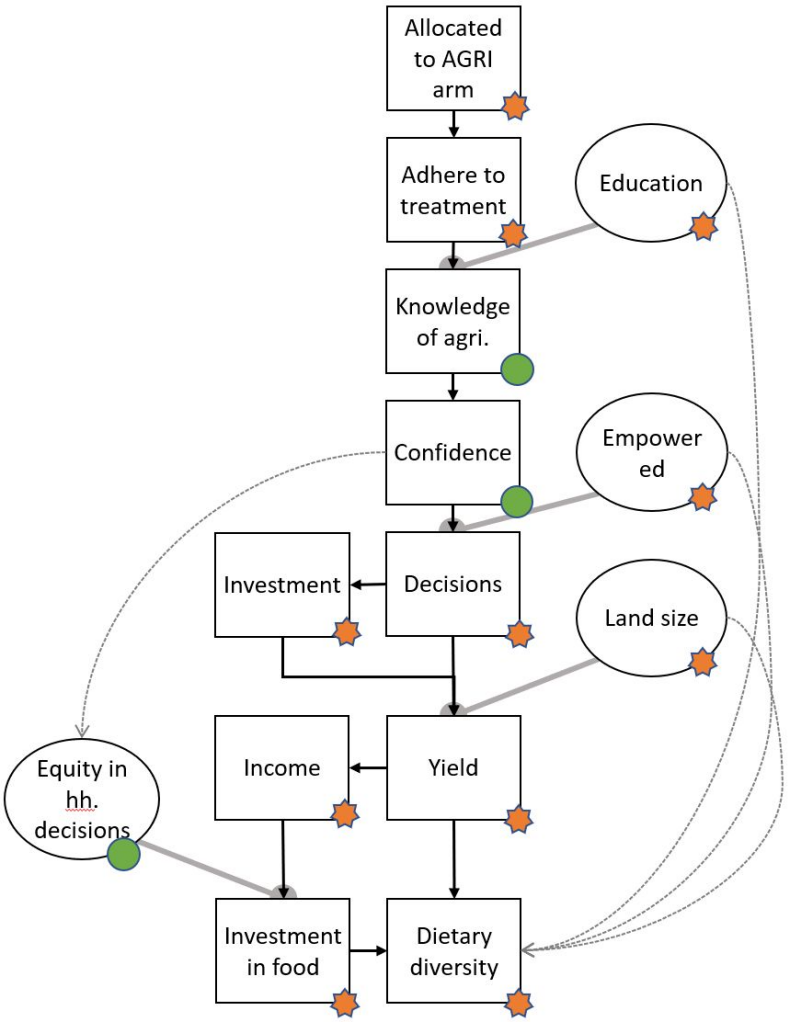
The square text boxes on this DAG, or 'nodes,' begin at the top with allocation to treatment and end at the bottom with the primary outcome of maternal minimum dietary diversity. The nodes between are the mediators that sit on the causal chain from allocation to treatment to the primary outcome. Circular text boxes represent moderators that act to intensify or frustrate the activation of a mediating node on the causal chain. We have represented the data type available for each node with a star or a circle inserted in one corner of the text box. Stars denote a node for which quantitative data is available from the baseline surveys or management information and circles denote the availability of qualitative data

from focus groups or interviews. We generally avoid the term 'mechanism' in our internal documentation because it is somewhat unclear at what level the term 'mechanism' should be used, and this can cause confusion. However, the creation of a DAG involves thinking about and representing causal processes that could helpfully be thought of as the action of mechanisms. Each arrow linking nodes on our causal chain can be thought of as representing the action of something like an Elster mechanism (Elster, 1998) or what Cartwright et al. call a 'tendency principle' (Cartwright *et al.*, 2020), which captures a highly general feature of human behaviour that is nonetheless not universal. For example, the causal link between A) confidence [in changing agricultural practices] and B) [the fact of making more agricultural] decisions is underpinned by the assumption of the action of an Elster mechanism like 'people who are more confident in their decision-making ability will be more likely to assert their opinions and to be responsible for collective decisions.' Similarly, the moderating effect of C) [a woman's level of] empowerment on that causal interaction can be thought of as representing an further Elster mechanism or tendency principle that more empowered women are more likely to act on a feeling of increased confidence and asset themselves in collective processes.

The DAG in **Figure 2**, or a similar, more refined version, will be shared with the UPAVAN implementation team and presented to them with an explanation of what the symbols in the DAG are supposed to represent. We anticipate that they will have suggestions for changes to the main structure of the DAG, which we will be able to incorporate in order to improve it, or which we may reject if we believe that the suggestions add too much complexity for analysis of the DAG to be tractable or are otherwise problematic. This process will be a sort of negotiation and also a way of exploring and refining all stakeholders' thinking about causation for the intervention. For case studies two, three and four, this process will occur before data collection and may be able to guide evaluation teams to collect informative data which they otherwise would not have thought to collect. It is possible that this process will lead to multiple DAGs if there are different, incompatible models for the way causation works in the intervention that different stakeholders support. In this case, the identification of that confusion, if it was not already known, will be a useful finding. Our data analysis will later allow us to test the fit of the data to each of the competing DAGs, possibly allowing us to say which is better supported by the data.

As further discussed below in the subsection on the risk of confounding, it will be necessary to work with the implementation team to further complicate the DAG by adding arrows between nodes representing any routes for confounding that the implementation team believe are possible. **Figure 3** complicates **Figure 2** to show confounding using dashed arrows.

Figure 3: UPAVAN arm one candidate DAG with confounding



The next stage in applying our method will be to gather from the UPAVAN implementing team their prior beliefs about the direction and strength of the causal relationships depicted in the DAG. The DAG itself contains the strongest prior assumption: that of no causal relation depicted by a lack of an arrow between the elements of the diagram. Less strong priors such as restrictions may be used (e.g. that the causal effect can only be positive), and continuous priors that may be 'flat' (i.e. uninformative) or concentrated at particular values. We will discuss these with the implementing team. The implementing team is incentivised to take a positive prior position on the interventions' effects and we will ask for empirical evidence from the past or elsewhere, where possible, to support these prior beliefs. Where empirical evidence is not available, we will document carefully the balance of expert opinion. Our own, and the original evaluators' own, beliefs will also be incorporated. These priors will be aggregated and

fitted to the model unless there is a marked divergence in priors between groups of stakeholders; in such cases it may make sense to maintain a separation between these different viewpoints and to fit two or more different sets of priors to the DAG to represent these different ways of thinking about intervention causation.

Our collaboration with the UPAVAN evaluation is a special case because, as noted, the evaluation is completed. The UPAVAN study will therefore have a different position from which the 'prior' beliefs about the intervention are established. The 'prior' beliefs in the UPAVAN case are in fact informed by the trial analysis and process evaluation. This is a challenge in that it could create confusion: should the results be taken into account during this process or not? While we will do our best to document where the trial results are informing beliefs, we will not be able to rule-out the influence of the experience of conducting the intervention and evaluation on the current beliefs held about the intervention. Therefore, we will take the prior position to be prior to conducting our specific analysis. This is a different prior position than we will take for the other case studies, where the priors will be developed *prior* to conducting the evaluation. In the case of the disability-inclusive poverty-graduation intervention in Uganda, the intervention will have started when we develop the DAG and priors, but the analysis will not. For the STAR+ and SHARPE studies the priors will be developed before the main interventions are delivered, but during or after pilot studies that will test some of the assumptions (priors) in the model.

Step 2: identify and link the data with the DAG

The data available to update the model comes from the baseline, endline, and the qualitative and quantitative sources used in the process evaluation. Together, the data are sparse since a minority of the clusters, let alone the families themselves, have data on the process measures.

Our first approach at the analysis will be done at the cluster level. At the cluster level, the data will take the form shown in the table below, with many rows with missing data. Since at present `CausalQueries` can only accommodate binary data, all data will need to be translated into binary using cut-offs.

Cluster ID	Baseline measures	Process measure 1 (FG)	Process measure 2 (CS)	Endline measures
1	0	0		1
2	0	1		1
3	1	1	0	0
4	1		1	0
5	0			1
...etc.	...etc.	...etc.	...etc.	...etc.

For our second approach, we will allow greater variation in the cluster-level outcomes by modelling the data at the individual level (which will still need to be binary). The cluster ID will be used to link process data (summarised at the cluster level) with the data at the individual level.

Step 3: updating the model

The model will be updated by the data using the R package `CausalQueries`. Where necessary, we will update different DAGs, and possibly DAGs with different priors. This will result in a model with new beliefs about the underlying causal processes that were informed by the priors and the observed data across the causal model.

Step 4: querying the model and reporting

We will query the updated model to discover what it implies about the direction, magnitude, and certainty of various causal processes under different conditions. For example, we will query the model for average treatment effect, and compare this to what was observed in the trial analysis (i.e. without referring to the process data). This will help illustrate the value of the process data for informing inferences, rendering the ATE more precisely than the original evaluation which did not use process data to inform this calculation. We will consider effects under different potential moderators. By analysing the effects of moderators linked to intervention implementation or to features of the context, we can learn more about the transportability of the findings to different contexts on different assumptions

about the fidelity of implementation and the characteristics of recipients and their environment in some target context.

Technical design

We will apply the method to four case studies. These are:

1. UPAVAN trial in India (described above)
2. DIG (disability inclusive graduation) trial in Uganda
3. STAR+ in Bangladesh
4. SHARPE in Ethiopia

Each will pose its own challenges, and we describe our expectations in this section. First we will describe each intervention in more detail. We will then briefly describe the evaluations.

Case-studies two, three, and four

Disability-inclusive graduation (DIG)

Intervention

The disability-inclusive graduation (DIG) programme has a series of activities that together aim to sustainably improve the livelihood of the very poor. This adaptation for people with disabilities was co-designed over a nine-month period by BRAC, Humanity and Inclusion (HI), and National Union of Women with Disabilities of Uganda (NUWODU). They consulted people with disabilities, government officials and civil society, members of targeted communities and participants from BRAC's existing poverty graduation project in Uganda.

The programme activities include four 'pillars', or components, which are intended to be mutually complementary:

Livelihood support. Transfer of an income-generating asset technical training, and individual-level support. Assets will be chosen on the basis of local market opportunities and the capacity of the recipients. This component should lead to improved enterprise management skills, asset accumulation/diversification and increased income.

Social protection. Participants will be supported with a six-month consumption stipend, healthcare subsidy and rehabilitation, physiotherapy, and psychosocial support. This should allow beneficiaries to focus on asset management. There will also be activities to support beneficiaries to overcome access barriers to government and NGO social entitlements (disability-specific and general) and support services (health, education, social protection) in order to support long-term social protection. Social protection should result in increased household food intake, dietary diversity and improved health, increased access to health services, social safety nets and support mechanisms.

Financial Inclusion. Participants will receive inclusive financial literacy training, offered membership of Village Savings and Loans Associations (VSLA), and receive on-going coaching to increase financial management skills and encourage savings behaviour. Financial inclusion should result in improved financial management skills and increased savings, developing participants' ability and confidence to access financial services, cope with shocks and invest in productive assets.

Social Empowerment. Participants will receive coaching to provide individual counselling and life-skills, individual empowerment plans (supported by HI), the formation of inclusive Village Poverty Reduction Committees (VPRCs). These are local governance structures with government/community leadership representation.

There will also be advocacy/social cohesion activities including sensitising village and national leaders on disability inclusion through NUWODU's District Women's Associations. Civil society actors (BRAC UG, HI, NUWODU), in addition to NUDIPU (the overall umbrella organization for disability in Uganda), and government agencies (local government, District Development Committees (DDCs) and the Ministry of Gender, Labour and Social Development (MGLSD), will be engaged through specific training for disability-inclusive graduation, inclusion training, organisational disability-inclusion action plans, specific organisation development training for NUWODU & Disabled People's Organizations (DPOs) and district/ national advocacy/ learning workshops. Social empowerment should result in better social integration within households and communities and improve participant confidence and aspirations.

The DIG programme aims to address barriers to the participation of people with disabilities by conducting a barrier-and-facilitator analysis and adapting project tools, processes and approaches to ensure meaningful inclusion. The team will be providing life skills training, coaching, and emotional support to build confidence through bi-monthly household visits of project field staff (caseworkers). Attitudinal barriers among project participants, project staff, BRAC Uganda staff and key external stakeholders will be addressed through sensitivity training provided by NUWODU, local DPOs and HI.

The project aims to shift norms and behaviours at community level by bringing together local DPO leaders, local and religious leaders in VPRCs to advocate for the empowerment of people with disabilities. The DIG programme team aims to anchor disability inclusion at the organisational level by conducting a disability-inclusion self-assessment of BRAC Uganda, led by NUWODU and HI, to develop a disability inclusion plan and monitor its implementation, including changes at programme, management, staffing, policies and human resources level.

Evaluation

The evaluation of the DIG programme is being led by a team in the International Centre for Evidence for Disability at LSHTM in collaboration with the Independent Evaluation and Research Cell at BRAC and the MRC Unit in Uganda.

The impact evaluation will estimate the effect of the DIG Programme on multidimensional poverty among people with disabilities and their families in Uganda. The primary outcome is per-capita monthly expenditure, at the household level.

The impact evaluation design is a cluster randomised controlled trial in four districts in Northern Uganda. Clusters are villages with at least 10 eligible households (or groupings of smaller villages to include 10 or more eligible households). Stratified by district and BRAC branch office, 160 villages were randomly allocated 1:1 to receive the DIG programme, and half to receive the control condition (business as usual). These villages contained 3,656 participating households, approximately 25% of whom are headed by a person with a disability.

Power calculations before the trials started estimated that the comparison between the arms will be sufficiently precise for the 95% confidence interval to exclude a null if the difference between the arms at endline is greater than 0.1 standard-deviation of per-capita consumption at baseline. A recent summary of multiple evaluations of the graduation programme found that the effect on per capita consumption was a difference of 0.1 standard deviations of the baseline (Banerjee *et al.*, 2015). Therefore, the evaluation is considered to be sufficiently powered.

Baseline data were collected in November-December 2020, and the endline will be two years later (October-November 2022), which is six months after the 18-month programme is complete. Household consumption measures included consumption of 37 food items (e.g. tea, jackfruit), 15 goods or services (e.g. electricity, mobile air time), 23 infrequent expenses (e.g. shoes, doctor fees), costs, time and

production on crops and livestock, other household income-generating activities and household loans and savings. Items are included on food security, household assistance, household health, fertility and mortality. The questionnaire included the Extended Washington Group Set of questions on disability, as well as items on personal assistance and assistive products/services, participation, attitudes to disability, violence, environmental factors, and social networks.

A process evaluation of the overall programme will be undertaken to understand the pathways to impact and consider which changes may be needed to improve programme impact. The process evaluation will involve field workers and researchers with disabilities to ensure high-quality qualitative data are collected from the disabled participants. As an organisation, BRAC collects a large amount of high-quality monitoring data that will be used to describe the delivery of different components of the intervention.

STAR+

Intervention

The STAR programme is a skills-development scheme that BRAC has been running in Bangladesh since 2012. It has reached approximately 30,00 youth, including 2,100 with disabilities (7% of total graduates). The aim of the programme is to target very poor young people who are out of work and have no hope of returning to education. On-the-job training is provided through apprenticeship under a local Master Crafts Person (MCP). Additional classroom training includes theoretical training on specific trades, and soft skills training such as financial literacy, market assessment, and basic English. Once training is completed, the programme officers will attempt to link participants with potential employers.

STAR+ ('STAR plus') is an adaptation of the original STAR programme that will target and support young people aged 14-24 with disabilities. Although some young people with disabilities have been supported by STAR, these were, according to the programme team, usually people with less severe impairments or with mobility impairments only. STAR+ will be an inclusive programme that will aim to support people with a broad range of impairments, including sensory and learning difficulties. STAR+ is being delivered by BRAC in collaboration with HI and Sense international.

The intervention will be piloted this year (2021) with around 300 young people before being finalised for a trial with 700 (and 700 control) in 2022. The intervention components will include:

- Participant young people with disabilities receive technical (apprenticeship and classroom training) and soft-skills training over 6 months
- Training materials will be adapted to be inclusive, and all staff and MCPs will be trained
- Participants who complete the training will be placed into jobs
- Participants will receive assessment of their impairments and be provided with assistive technologies and rehabilitation support
- Sensitisation events will be conducted with the families, communities, and market actors (potential employers, clients/consumers)
- Workplaces will be assessed for accessibility and reasonable accommodation will be made

Evaluation

The evaluation of STAR+ is being led by the ICED in collaboration with BRAC University, Dhaka.

The aim is to identify the effect of the modified STAR programme on employment among young people with disabilities. Primary outcome measures will be employment status and average hourly earnings.

In the second quarter of 2022, 700 young people will be recruited into the intervention. These 700 will be drawn at random from an over-sample of 1,400 young people, with the 700 not selected acting as the controls. Since many aspects of the intervention are at the cluster-level, the random selection will be done at the cluster-level.

For the power calculations the evaluators assumed that the proportion of participants in the treatment arm who were in employment at the end of follow up would be like the proportion observed in the evaluation of the original (non-inclusive) STAR evaluation (Das, 2021). For the control arm, the assumed proportion in employment would be lower than observed with the original STAR evaluation, consistent with the additional barriers facing the target population for this trial. With 700 participants in each arm, or 17 clusters per arm, there would be more than 80% power with a high ICC of 0.2 if the control arm employment level is 45%. For low ICC of 0.1, the control arm could have an employment level of 53% for adequate power. The calculations are approximate, but they imply that the trial will have adequate power for the main effects.

There will be an integrated process evaluation, looking at the implementation, the role of the context, and be part of exploring the underlying mechanisms. Again, BRAC has a strong monitoring and evaluation framework that will provide information on the delivery of the intervention.

SHARPE

SHARPE is a humanitarian project using a market systems approach to improve livelihoods among both refugees and host populations in three locations in Ethiopia. The project will work in the Jijiga, Dollo Ado, and Gambella regions, where each region has a large refugee camp.

SHARPE will collect data on markets in the region and develop a response. Their early research suggests that a market in digital finance could be created and would be useful for local people. They hope that successful implementation of a digital finance platform will provide infrastructure for other development and humanitarian support programmes, in particular the World Food Programme delivery of cash transfers for food.

The programme is being evaluated by the International Food Policy Research Institute (IFPRI). They will work with SHARPE to evaluate the overall response to the market systems need, and also conduct rapid RCTs to inform key decisions about the rollout. We will partner with IFRPI to improve the inferences from these rapid RCTs. This will trial POInT in an operational context and test its utility for answering questions that can guide rollout.

Developing middle-range theories / DAGs from theories of change

As shown in the UPAVAN vignette above, the first step in the application of the method is to work with the implementation and evaluation teams to translate the theory of change into a DAG, and elucidate the priors. We will attempt different methods for achieving this, carefully documenting what works best for whom under what circumstances, and summarising these findings in a protocol for future research use. This stage of our project has been affected by the COVID pandemic since we will be unable to hold face-to-face workshops. Instead, virtual meetings will be held and we will make best use of a generally increased level of competence with online tools, such as using virtual whiteboards and shared documents.

Examples of the methods that we will use include:

- Iterative conversion of the theory of change into a DAG. We will propose possible DAGs and use criticism from the implementation and evaluation teams to guide a series of improvements

- Brief, basic training in the language of DAGs. Short presentations, exercises, and examples will be used to explain the basics of DAGs to the teams. These will be recorded and made available for others to use.
- Elucidating priors, requiring the following steps:
  - Conveying the importance and role of priors in Bayesian analysis. This will be achieved through accessible documentation, presentations, illustrations, and examples.
  - Starting with the strongest assumptions. We will start by acknowledging the lack of (or removing) arrows (edge) between nodes in the DAG where the prior for a causal effect is zero. The lack of an edge is a prior belief that is set when drawing the DAG. Recognising this should help make the identification of priors more intuitive.
  - Next, identifying edges where restrictive priors can be set. These are the next strongest priors, such as that the intervention can only increase the level of the intermediate outcome. Some of these will be intuitive, and restrictions do not initially require specification of a distribution.
  - Next, cycle through each edge and discuss and agree the prior distributions.
  - Finally, for each node with more than one incoming edge, discuss and agree priors on the interactions.

The consolidated priors across the DAG will allow for certain inferences to be made about the theory even before any data is collected. It will allow us to determine what the expected effect of the intervention is, and its distribution. We will be able to inform the evaluators of the probative value of different data sources, potentially improving the efficiency of the studies for which our collaboration is at an early enough stage to affect data collection.

Outcomes and how they will be measured

Once we have a DAG or DAGs with one or more sets of consolidated priors about the strength and direction of causal relationships between nodes in the DAG, the next stage in our analysis of each interventions will be to confront that model with qualitative and quantitative data generated by the process and impact evaluations.

Primary outcome

In order to make the application of our novel method manageable for the team and easily comprehensible for readers of our work, we intend to limit our analyses to the causal relationship between intervention components, contextual factors, and a single primary outcome for each intervention. Assessing the effect of the intervention on multiple outcomes would require a proliferation of DAGs that might breach the limits of the capacity of our team, and might make our research outputs difficult to understand, our method more difficult to adapt and adopt. We intend to start simple, with an application of our method to the simplest arm in the UPAVAN evaluation, in which recipients were exposed only to nutrition-sensitive agriculture videos, and not in addition to nutrition-specific videos or participatory learning and action meetings. This limited application of the method will provide valuable learning for our team as well as a clear, concise case-study in the method. This case-study will provide the raw material for training materials, and the first methods briefs and workshops seeking to influence fellow researchers.

Our applications to case-studies two, three and four will necessarily be more complicated. As we describe above, the DIG intervention includes four components which are intended to be mutually reinforcing. Similarly, STAR+ involves multiple intervention components. The market systems approach of the SHARPE intervention may result in an intervention that is highly complicated, with different recipients being exposed to different and sometimes overlapping aspects of an intervention designed to operate at multiple levels.[2] As a result of the proliferation of intervention components, our DAGs for subsequent case studies will be more complicated than for the UPAVAN study, posing technical and capacity challenges that we are confident will test our team and be overcome. This will be the case even if we restrict our analysis to one primary outcome of interest, so we consider that a necessary limitation of scope. In all cases, primary outcome data will be generated quantitatively at the individual level using survey instruments. Importantly, this will allow us to use baseline data to disaggregate outcomes by recipient characteristics such as gender and disability status in order to identify heterogeneous effects.

[2] Currently, the SHARPE timeline suggests that our collaboration with SHARPE will finish last. This is a happy circumstance, meaning that working from UPAVAN through DIG and STAR+ to SHARPE will represent an increasing challenge to the application of our method, allowing us to progressively take on harder problems and to document our approach to these challenges in sequence.

Mediators and moderators

Each causal model will have a number of mediators that are expected to explain (and from a more ontologically Realist perspective, bring about) the effects of the intervention on the primary outcome. The mediators will be measured using three sources: monitoring data, qualitative data, and the endline survey. None of the trials have or are planning to have a midline survey of all participants. Qualitative data collection will be by means of in-depth interviews, semi-structured interviews, life-histories or other individual qualitative data generation instruments as well as focus group discussions. In the first instance, we will attempt to work with the coded interpretations of the qualitative data produced by process evaluation teams. However, we will need to re-analyse this data for our purposes. Process evaluation teams may interrogate the data in a narrative manner, using their impressions of rich, qualitative information to inform nuanced interpretations of e.g. the way in which recipients felt about the intervention. Our purposes will be different and our tools limited in the complexity of data they are able to interpret. We will seek to analyse the extent to which the presence or absence of some mediator or moderator conditions the presence or absence of other nodes on the causal chain. In order to interrogate data using CausalQueries we will need to work with a simplification of the data that can be expressed as binary or possibly as categorical ordinal data. This means that we will need to reinterpret qualitative data by 'binning' them into one of two or possibly more bins. This may take the form of translating the data to '1' if one of the data coders' tags is present or '0' if it is not, or it may require re-coding in order to achieve this transformation to binary (or ordinal) data. This process is necessarily reductive. It means that a large amount of the information in the qualitative data is lost. However, the status quo alternative is that 100% of the data is lost for answering some sorts of research question, and that when the data is preserved for answering other questions, it cannot be rigorously combined with quantitative data.

Some of the moderators represented in our DAGs are likely to be measures of compliance with, take up of or exposure to intervention components. For example, the UPAVAN trial monitoring data allows us to measure whether women saw one or more of the videos that they should have seen if they were assigned to treatment and adhered to it. Other mediators might be intermediate outcomes that are observed through qualitative process data such as the increased feeling of confidence in their ability to change agricultural practices that participants in the UPAVAN trial were expected to express.

We will discuss with the primary evaluators of the interventions if they can add items to their endline questionnaire or alter their procedures for collecting qualitative data if we find from the prior model that particular information would be important for interpreting the final model.

Below we discuss the limitation that many of our measures of mediating variables will be measured mostly or exclusively in the intervention arms of the trials. However, it should be noted that conventional methods for analysing mediation rely on complete data, whereas POInT will allow incorporation of sparse data, both limited to the intervention arm only or collected in a minority of people or places, as will be the case for the qualitative data.

As much as possible we will treat moderators as contextual factors that interact with the mediated causal effects of the intervention. These will ideally be measured before or close to the start of the intervention, and we will only use data from time-varying moderators measured later where there is little theoretical reason to believe that the intervention might have altered the moderator.

Analysis method and establishing causal inference

The POInT project includes the building of theory, elicitation of priors, and the querying and interpretation of the results for policy, but at the heart is the CausalQueries package and underlying mathematics that combines the theory and priors with data from different sources. CausalQueries is both an approach and a causal-modelling front-end R package that enables model fitting using Stan, an open-source Bayesian model-fitting software. Since we intend for this method to be used widely, we will document the execution of the modeling fitting carefully, creating video tutorials, manuals, and troubleshooting guides.

The CausalQueries approach is built on the 'Bayesian integration of qualitative and quantitative information' (BIQQ) approach, that was developed by Macartan Humphreys and Alan Jacobs (Humphreys & Jacobs, 2015). Both Macartan and Alan are members of our team, along with Lily Medina who was the main author of the CausalQueries R package.

The BIQQ approach showed that it was possible to combine 'between case' information, such as exposure to an intervention and the final outcome, with 'within case' information, such as qualitative 'clues' that indicated one way or another whether there was a casual process occurring that linked the exposure with the outcome. This was achieved in part by specifying in advance what kinds of inference would be made when observing the presence or absence of the qualitative clues. Some clues would, for

example, be relatively uninformative if the intervention was as likely to be causal when the clue was observed and when it was not (so called 'straw in the wind'), whereas another might be informative if the intervention was very likely to be causal if the clue was observed even if little could be learned from it not being observed (so called 'smoking gun').

The CausalQueries method developed from BIQQ by adding more complete and explicit causal models. Key to understanding the method is to think of the overall effects in the sample, for example of an intervention on an outcome, as the average of case-specific effects. This is of course how all average effects are found, but there can be a temptation to interpret an average treatment effect (ATE), for example, as being the effect experienced by all participants. The BIQQ and CausalQueries approach explicitly recognises this underlying heterogeneity in what is referred to as the 'causal types'. In the case of an intervention effect on the primary outcome, these are: responding positively to the intervention, negatively, contrapositively, or not at all. Understanding the distribution of these causal types is the task of estimating the average effects; understanding the reasons for diversity of causal types is the task of moderator analysis. For example, imagine the simple case of an exposure and an outcome (e.g. X -> Y), where both are binary. We take for granted that X can be zero or one, such as in a randomised trial and is unaffected by anything else (there are no arrows going into X). Y can be affected by X in one of four different ways. If X = 0 then Y can be either zero or one, and if X = 1 then Y can also be zero or one. The distribution of the causal types will determine the average effect of changing X from zero to one for all participants.

The causal model, or DAG, is specified using CausalQueries using the `make_model` command and arrows and labels, such as `make_model('X->Y')`. CausalQueries brings the DAG together with the causal types reasoning above in the way as we would approach a mediation analysis, by decomposing the *overall* causal types across the model into their constituent 'nodal types', i.e. the causal types operating at each node in the DAG. (Note that 'causal types' and 'nodal types' are the same in the simple case X->Y described above because there is only one node with incoming causal effects.) The distribution of overall causal types will be determined by the distribution of the nodal types, and because of this the number of possible causal types increases very quickly with the addition of nodes in the model. Where there are multiple arrows arriving at the same node, for example at Y in X->Y<-M, then the number of nodal types will be 2^{2^k}, where k is the number of arrows arriving at the node. In other words, the approach allows for complete interaction between the causal effects occurring at the nodes.

As we have already mentioned, the DAG contains priors wherever two nodes are *not* connected by an arrow, since this is the strong prior of no causal effect between these nodes. Less strong but nonetheless important priors can be set using restrictions. We might, for example, take the position that a causal effect must be positive if it exists at all (e.g. 'that Y is not decreasing in X'). These restrictions can be set on the model using the `set_restrictions` command, in this case `set_restrictions(model, decreasing('X', 'Y'))`. Restrictions are set at the level of the nodal type and through the structure of the model imply restrictions in the causal types.

We may wish to set numerical priors on the distributions of the nodal types. These are specified as the alpha arguments in a Dirichlet distribution, which is commonly used for the priors in Bayesian statistics. By default, the alphas for all nodal types is one, which implies a 'flat' or uninformative prior. But if we had reason to believe that one set of nodel types should be more common than another, then we can set the alphas to higher values for the types we think are more likely and lower for those we think are less likely. For example, in X->Y we have four nodal types at Y. If we think that there should be a small positive causal effect of X on Y then we could set the distributions of Y = 1 | X = 1 and Y = 0 | X = 0 to higher values than those for Y = 1 | X = 0 and Y = 0 | X = 1. A challenge for our project will be translating prior beliefs held by the teams into priors specified in terms of the distribution of nodal types; this will be most complex with interactions.

Using CausalQueries, and Stan, the causal model and associated priors will be updated with the observed data to produce posterior distributions across all processes in the DAG and the joint distribution of the entire model.

A technical limit of CausalQueries is that it is so far limited to binary data. While this is generally unproblematic for the intervention, it may be limiting to only allow binary outcomes (or dichotomisation of continuous outcomes). The utility of the method may be reduced because information is lost when continuous data are dichotomised (Altman & Royston, 2006). Some policy-making audiences may prefer to see effects on continuous distributions, such as the increase in household per capita consumption, as opposed to effects on the probability of being in one or other group, such as 'ultra poor' and 'poor'. The CausalQueries methodology can in theory be extended to include such outcomes, and in POInT we will extend the package to allow a continuous primary outcome. One way of approaching this problem may be to allow for ever greater numbers of categories, and a positive side-effect would be to extend the code to allow for categorical data with more than two categories.

Threats to validity

Our project will exclusively be using data collected as part of the main evaluations of the case studies. Therefore, there may be limitations that we cannot influence. However, these are all well-resourced and carefully designed evaluations by experienced teams. Three of the four studies are cluster randomised trials, with little risk of selection issues or spillovers. The quality of the qualitative work is also likely to be high, with well-resourced process evaluations being completed by appropriate teams.

For example, the UPAVAN process evaluation team, as well as making use of quantitative data from baseline, endline and management information systems, generated a variety of high-quality qualitative data. They conducted 32 family case studies with pregnant women, mothers of children under two, their husbands and in-laws, totalling 91 semi-structured interviews. Seventeen focus groups discussions were also convened with a total of 181 recipients, as well as three focus group discussions with implementation staff. Selection of recipients was balanced using the same criteria as the stratification of recipients for quantitative data generation. Data were transcribed and translated into English before being tagged in Nvivo using a thematic approach to capture themes related to each component of the theory of change as well as any further emergent themes. This data was then coded; adding, pruning or aggregating nodes via consensus between the four coders. These results were then discussed with the implementation and the wider evaluation teams to expose them to challenge or confirmation.

In addition to any threats to validity of the evaluations with which we are partnering, there are some threats to validity that are specific to our project and which we cover briefly below.

Insufficient power

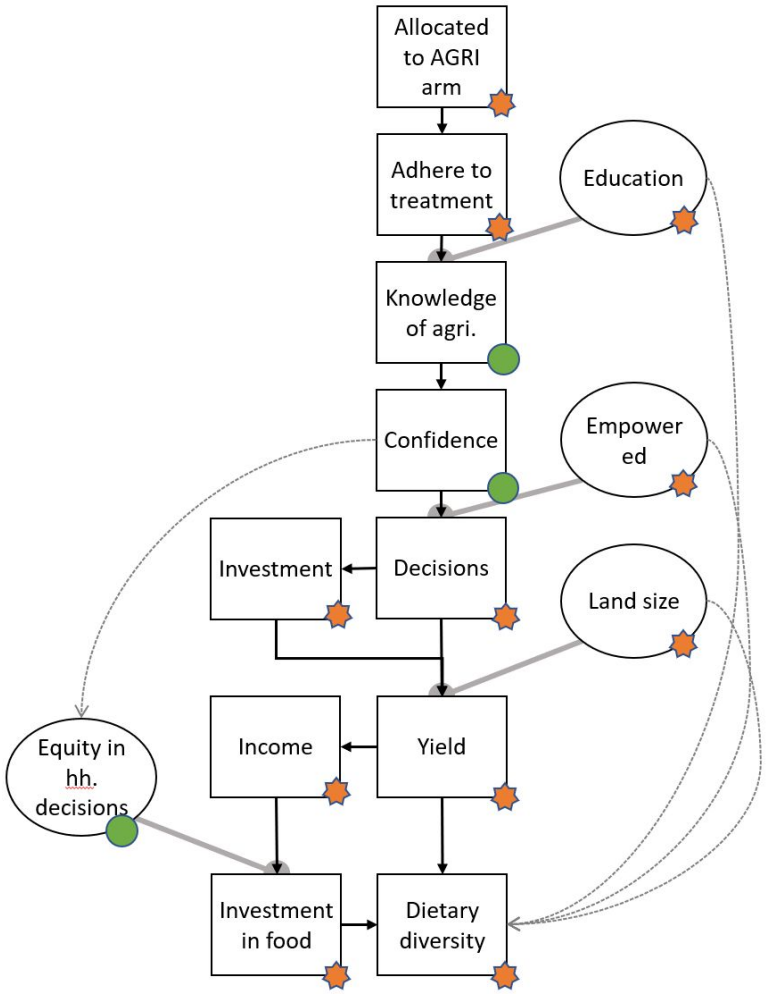
We do not have the ability to influence the sample sizes of the evaluations with which we are partnering. This presents the risk that, although powered to detect small average treatment effects on the primary outcome, trials might not be sufficiently powered to detect small causal effects that our analysis tells us might be informative; for example heterogenous effects for any subgroups of recipients identified on baseline characteristics; or effects of moderators only present for some recipients. However, for three of our partner evaluations we can influence the process data that is collected, and we will have thought in advance about which data will be most informative. Our analysis of the DAG(s) created that capture the action of intervention mechanisms might tell us that there are hoop tests, smoking guns, or even doubly-decisive tests that could confirm some causal hypothesis about the action

of moderators in causing the primary outcome of interest. If such strong tests are implied by the DAG, then we do not need to be able to detect small effects to gain powerful information about the theory of change. We may, for example, be able to disprove that some aspect of the intervention is *necessary* for achieving an outcome just by observing one instance of the outcome without the presence of that aspect of the intervention. In cases where our theory tells us that small effects might be expected, we may be unable to detect them. This will nonetheless be an informative methodological finding, telling us that a trial considered to be well-resourced is nonetheless incapable of testing some important aspect of its theory of change.

Confounding

As discussed above, the CausalQueries method recognises that no causal connection between nodes is a very strong assumption. In the UPAVAN DAG in **Figure 2**, above, there are fairly few arrows. We have only drawn arrows representing 'edges' between nodes where the intervention theory of change theorises a strong connection. However, there may be many weak connections between nodes that risk confounding our analysis of the causal effects of mediators and moderators on further mediators and the whole causal chain through to the final outcome. In order to minimise unmeasured confounding, for the purposes of data analysis many more arrows representing possible confounding causal connections between nodes must be added to our DAGs as in **Figure 3**, reproduced below.

Figure 3: UPAVAN arm one candidate DAG with confounding



This sort of DAG minimises unmeasured confounding by accounting for the main sources of confounding that the primary evaluation and implementation teams anticipate. We can also minimise unmeasured confounding of the primary outcome by controlling for a wider set of observed potential confounders that are not included in the DAG. Nevertheless, we cannot exclude the possibility of unmeasured confounding by unobserved confounders of our results. CausalQueries permits us to allow for unmeasured confounding when calculating our results. This is useful in allowing us to quantify the extent to which unmeasured confounding reduces confidence in our results compared to a specification that assumes no unmeasured confounding.

Temporality and availability of mediator and moderator variables

Many of the mediating variables will be collected in the endline surveys. This raises the risk of reverse causation in models where there is the potential for the outcomes to change the mediators through feedback. The use of process data will help address this issue, as we will be able to use data collected throughout the delivery of the intervention rather than rely solely on the endline survey.

Much of the monitoring and qualitative data will be collected on the intervention arms only. Although we will encourage data collection in both arms where our models suggest it would be informative, we expect that this will remain an issue. Since this is a common occurrence, showing that the method can account for situations where there is little or no data on mediators or moderators in the control arm will be a strength of the method. We will, for example, be able to use data on adherence to the intervention to update confidence in the overall causal effect, even if measured in a minority of places with the process evaluation. In contrast, complier average causal effect estimation requires complete data on adherence in the intervention arm (Peugh *et al.*, 2017).

Data quality should be high since the evaluations are all being conducted by world-class multidisciplinary teams. However, both the quantitative and qualitative data are at risk of bias. We will monitor the implementation of data collection to observe whether tools that are used are appropriate for the context and have been adequately field tested (for example, with cognitive interviews). For DIG and STAR+ we have assurances that everyone involved in the research will participate in disability-awareness training, and that researchers with disabilities will be involved in the design and conduct of the trials. Where possible, disabled field workers will be employed for the questionnaires. For qualitative research there will be particular effort made to involve disabled researchers and to intensively trained non-disabled researchers, in collaboration with local disabled-people's organisations.

Advantages of POInT

Reinforcing the formulation of high-impact evaluation questions

As discussed above, our project aims to augment four existing evaluations, helping those evaluation teams to make better use of all of their data to respond to their evaluation questions. Our method will also help those teams to ensure that they are asking the right questions through their evaluation. All of the evaluations with which we are partnering are of interventions designed to address pressing social problems, whose primary research questions are formulated to respond to that social problem.

However, our method adds considerable value. At the point of eliciting priors, we will prompt teams to reflect on their use of the existing evidence base including relevant evidence synthesis products such as narrative and systematic reviews, meta-analyses and gap maps (for example, on disability (White, Saran & Kuper, 2018)), ensuring that research questions are maximally informed by these sources.

Enabling pre-analysis for a wider set of inferences

Given our application of a novel method, and especially in light of the fact that our method is being developed in the process of applying it to our four case studies, our analysis of the four case studies cannot be pre-registered. Further, the analysis that we will contribute to each of the four case study evaluations will fall outside of their pre-registered analysis plans. This means that the results of each application of our method will be published separately from the main evaluation results for each case study and with an acknowledgement of the fact that our analysis were not pre-registered. Nevertheless, it is a great strength of POInT that our method allows for the pre-specification and registration of analyses for a wider range of inferences than is currently the case. Once POInT is a proven method of making inferences which combine qualitative process and quantitative impact data to make arguments for transportable results it will be possible to pre-specify this analysis, rather than leaving the integration of process and impact data to an ad-hoc and informal process, as is currently the case.

Using all available data and pre-existing theory to increase confidence in causal claims

Applying POInT using CausalQueries to our four case studies will allow us to calculate an alternative estimate of the average treatment effect for the primary outcome that is based not only on quantitative data in the baseline and endline surveys, but also on monitoring data and qualitative data from the process evaluation, and on prior theory. The resultant estimate will be an improvement over an estimate generated using frequentist statistics in that it will be based on more information and in that the method of its estimation allows us to distinguish between two types of uncertainty. These are A) uncertainty resulting from limited sampling from a hypothesised underlying 'true' distribution, and B) uncertainty resulting from imperfect coherence with pre-existing theory, or model uncertainty.

Consider a situation in which frequentist statistical estimation of the average treatment effect generates a moderate, positive, statistically significant estimate with an associated confidence interval. This might be accompanied by two possible sets of observations of process and monitoring data. 1) process data on mediators shows changes in these mediators roughly as suggested by the theory of change. For

example, using the UPAVAN case, women who saw videos reported feeling more confident, taking more agricultural decisions, improving their crop diversity, and thereby improved their dietary diversity. Another possibility is that 2) process data on mediators shows changes in those mediators that are not those suggested by the theory of change, despite recipients having the same improved primary outcomes. For example, women who saw videos did not report feeling more confident or taking more decisions, and did not appear to have improved their crop diversity but nevertheless improved their dietary diversity. A frequentist approach to estimation of the average treatment effect will result in the same point estimate and confidence interval in both cases. That confidence interval would reflect only one sort of uncertainty: sampling uncertainty. An informal discussion of the lack of fit with existing theory and the surprising findings from mediation analysis might be included in the paper, or not. Using POInT, the confidence associated with the calculation of the average treatment effect would not merely be an expression of sampling uncertainty but also of model uncertainty, resulting in a different, transparent, quantified level of confidence in the point estimate of the average treatment effect in situation 1) than in situation 2).

Exploring heterogeneous effects

As discussed in the section of this paper dealing with the measurement of outcomes, it will be possible to disaggregate primary outcomes data by gender, disability status and other recipient characteristics in order to describe heterogeneous effects. In addition, we will be able to explore the causation of those heterogeneous effects by analysing the extent to which they are driven by the presence or absence of mediators represented in the DAGs. For example, we will be able to ask questions like ‘was it the case that higher outcomes from STAR+ for people with certain disabilities were driven by their being more likely to be placed in jobs?’ or ‘was it the case that women’s lower attendance at the evening training sessions led to worse outcomes?’

Targeting transportability

By exploring and explaining heterogeneity of effects as described above, POInT provides a formal, transparent, replicable method for specifying the characteristics of contexts to which evaluation results should be expected to be transportable. By formulating the underlying theory as a DAG, we can draw on rigorous methods for transporting causal effects through the language of causal inference (Bareinboim & Pearl, 2013). For some evaluations, this rigorous method for integrating information about the role of context will pierce the ‘black box’ and shed light on the circumstances under which policy-makers can

expect programmes' impacts to replicate. As described above, for other evaluations the application of this method will reveal the fact that researchers are not confident about the contexts to which results will be transportable, or will reveal that the confidence researchers had prior to conducting the evaluation is undermined by the observations made. These too would be valuable findings.

Embodying and facilitating interdisciplinarity

To best implement POInT, we have constituted a deeply interdisciplinary team composed of specialists in both quantitative and qualitative analysis. Further, our team includes the original authors of BIQQ who hail from a political science background as well as development specialists and specialists in impact evaluation. Our team is composed of members from the Global North and South, being co-led by Dr Calum Davey and Professor Elizabeth Allen at LSHTM and with statistical analysis being led by Professor Henry Mwambi and supported by a postdoctoral researcher at the University of KwaZulu-Natal. Our different epistemological, ontological, methodological and geographical backgrounds equip us with the tools to understand and integrate different methods and to establish effective partnerships with different sorts of teams.

POInT facilities interdisciplinary work by providing a protocol for integrating not just quantitative and qualitative data, but also the theoretical insights of different practitioners at the DAG design and prior specification stages. The DAG under construction and the priors fitted to that DAG provide a 'boundary object' (Mollinga, 2008) that researchers from different disciplinary backgrounds can use to ground discussions they might not otherwise have. In the same vein, the move from mutually incomprehensible frequentist statistics on the one hand, and qualitative discussion of generative causes on the other; to Bayesian reasoning under POInT may provide a set of 'boundary concepts' (Mollinga, 2008) that allow researchers from different backgrounds to talk in a common language. We have described above how we hope that the process of creating the DAG and specifying the priors will allow for reflection on the extent to which all the available information from different disciplines has been used to inform the theory, identifying areas of the theory for improvement and targets for data collection that might otherwise have been missed. This may sound utopian, and the reality may be different. It is our prior that POInT will be useful in this way; the extent to which this proves to be the case will be one of the themes of our learning about our method and will be presented in our learning papers.

Ethical approvals and data sharing

For the UPAVAN trial, informed consent was obtained from village leaders for the participation of villages in the trial before randomisation and baseline survey. Data collectors sought informed consent from adult participants in writing or by thumbprint, and, for children, from their primary caregivers. Ethics approval was granted from the Odisha government's Institutional Review Board, Research and Ethics Committee, Department of Health and Family Welfare, Government of Odisha (date approved Sep 03, 2016, Letter No. 141/SHRMU). Ethics approval was granted from the LSHTM Interventions Research Ethics Committee (date approved Oct 10, 2016, Reference No. 11,357).

For the DIG trial, data collectors sought informed consent from all adult participants in writing. Ethical approval was awarded from the London School of Hygiene & Tropical Medicine (date approved Sep 21, 2020, Reference No. 22619), and the Mildmay Uganda board (date approved Jul 7, 2020, Reference No. 0604-2020). A research permit was granted from Uganda National Council for Science and Technology (UNCST).

For STAR+ and SHARPE, ethical approvals are underway.

For all case studies, data will be entered, cleaned, and anonymised by the primary evaluators, under supervision of an experienced data manager. Anonymised data will be shared with the POInT team using a secure-data-transfer protocol. Data will be held on servers at our respective universities.

Reporting

We will report the results of the POInT project using the following media:

- Papers
    - Protocol for the case-study applications
    - Paper 1: application to UPAVAN trial
    - Paper 2: elucidating theoretical priors
    - Paper(s) 3: integrated inference in three case studies
    - Paper 4: allowing for non-binary variables
    - Paper 5: formalising the value of process data in evaluation
- Webinar, *Combining process and impact data for better inference*

- Workshop, *Combining process and impact data for better inference: practical applications*
- Guidance, *Practical application of CausalQueries for evaluation of interventions*
- Online tool, *Defining causal models for evaluation*
- Training workshop, *Using CausalQueries for evaluation of interventions*
- Blog series
    - How can theories of change be made more useful for analysis?
    - Using process data to improve confidence in impact evaluation results
    - Learning about mechanisms by combining impact and process evaluation
    - Designing impact evaluations with inference for elsewhere in mind from the start
    - + summary blogs for each paper
- Videos: Workshop materials, walk-throughs, and lectures uploaded to YouTube channel *CausalQueries*

Our stakeholder engagement plan includes engagement with knowledge brokers to improve uptake.

Discussion and next steps

The POInT project has many strands, each with multiple steps, and will be applied to four case studies. It is an attempt to formalise inferences about the effects of interventions as a stronger basis from which to make predictions elsewhere. The goal is to better leverage theories of change and process evaluations. Achieving this will require innovation in the construction of formal causal models from conventional theories of change; elicitation of prior beliefs; coding of qualitative process data; overcoming the technical challenge of bringing it all together; reporting the findings accessibly, and making the method available for others to use easily.

Our aim is that POInT will increase the reliability, relevance, and use of the evaluation results. The method should provide greater certainty about causal estimates of the programme and underlying theory. Development of DAGs, priors, and explicit modeling of the processes will encourage precision about the priority development questions. Exploring the theory and the moderating effects of contextual variables will assist with transferability of findings to new contexts. Such results should be more likely to lead to policy change. Our ambition is to change research practice itself, to facilitate precision in the description of interventions and estimation of their effects.

The COVID-19 pandemic has affected the timelines of the project. Our team has been limited by delays in signing contracts and recruiting staff due to capacity shortages. The UPAVAN trial is completed and unaffected, and the DIG trial has been delayed by only six months. STAR+ has been delayed by twelve months by COVID-19, and SHARPE is also running behind schedule. These delays may have implications for the timeline of our project.

References

Aarons, G.A., Fettes, D.L., Sommerfeld, D.H. & Palinkas, L.A. (2012) Mixed methods for implementation research: application to evidence-based practice implementation and staff turnover in community-based organizations providing child welfare services. *Child maltreatment*.17 (1) pp.67–79.

Altman, D.G. & Royston, P. (2006) The cost of dichotomising continuous variables. *BMJ*. 332 (7549), 1080.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., et al. (2015) A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*. 348 (6236), 1260799.

Bareinboim, E. & Pearl, J. (2013) A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*. 1 (1), 107–134.

Breuer, E., Lee, L., De Silva, M. & Lund, C. (2015) Using theory of change to design and evaluate public health interventions: a systematic review. *Implementation Science*. 11 (1), 63.

Cartwright, N. (n.d.) Why Mixed Methods Are Necessary for Evaluating Any Policy. *NA*. NA.

Cartwright, N., Charlton, L., Juden, M., Munslow, T., et al. (2020) *Making predictions of programme success more reliable*.

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*.pp.2094–2103.

Daniel, R., De Stavola, B., Cousens, S. & Vansteelandt, S. (2015) Causal mediation analysis with multiple mediators. *Biometrics*. 71 (1), 1–14.

Das, N. (2021) Training the disadvantaged youth and labor market outcomes: Evidence from Bangladesh. *Journal of Development Economics*. 149, 102585.

Davey, C., Hargreaves, J., Hassan, S., Cartwright, N., et al. (2018) *Designing Evaluations to Provide Evidence to Inform Action in New Settings*.

Davies, R. (2018) *Representing Theories Of Change: Technical Challenges With Evaluation Consequences*.

De Silva, M.J., Breuer, E., Lee, L., Asher, L., et al. (2014) Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials*.15 (1) p.267.

Elster, J. (1998) *Social mechanisms: An analytical approach to social theory*. Cambridge University Press.

Giddens, A. (1984) *The constitution of society: Outline of the theory of structuration*. University of California Press.

Greenland, S., Pearl, J. & Robins, J.M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*. 37–48.

Humphreys, M. & Jacobs, A. (2015) Mixing Methods: A Bayesian Approach. *American Political Science Review*. [Online] 109 (4), 653–673. Available from: doi:DOI: 10.1017/S0003055415000453.

Jimenez, E., Waddington, H., Goel, N., Prost, A., et al. (2018) *Mixing and Matching: Using Qualitative Methods to Improve Quantitative Impact Evaluations (IEs) and Systematic Reviews (SRs) of Development Outcomes*.

Kadiyala, S., Prost, A., Harris-Fry, H., O'Hearn, M., et al. (2018) Upscaling Participatory Action and Videos for Agriculture and Nutrition (UPAVAN) trial comparing three variants of a nutrition-sensitive agricultural extension intervention to improve maternal and child nutritional outcomes in rural Odisha, India: study protocol for a cluster randomised controlled trial. *Trials*. [Online] 19 (1), 176. Available from: doi:10.1186/s13063-018-2521-y.

Leviton, L.C. (2017) Generalizing about Public Health Interventions: A Mixed-Methods Approach to

External Validity. *Annual Review of Public Health*. [Online] 38 (1), 371–391. Available from: doi:10.1146/annurev-publhealth-031816-044509.

Mollinga, P.P. (2008) *The rational organisation of dissent: Boundary concepts, boundary objects and boundary settings in the interdisciplinary study of natural resources management*.

NFHS-4 (2005) *National family health survey, India*. International Institute for Population Sciences and Macro International.

Peugh, J.L., Strotman, D., McGrady, M., Rausch, J., et al. (2017) Beyond intent to treat (ITT): A complier average causal effect (CACE) estimation primer. *Journal of school psychology*. 60, 7–24.

Steen, J., Loeys, T., Moerkerke, B. & Vansteelandt, S. (2017) Flexible mediation analysis with multiple mediators. *American journal of epidemiology*. 186 (2), 184–193.

VanderWeele, T. (2015) *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

White, H., Saran, A. & Kuper, H. (2018) *Evidence and gap map of studies assessing the effectiveness of interventions for people with disabilities*.

www.CEDILprogramme.org