Oxford Policy Management

Independent evaluation of the Demand-Driven Impact Evaluations for Decisions (3DE) Pilot

Final Report

21 July 2015

In association with:

Preface

This is the final report for the independent evaluation of the Demand-Driven Impact Evaluations for Decisions (3DE) pilot, which has been produced by Oxford Policy Management (OPM) in association with the Overseas Development Institute (ODI). The report provides findings from testing the Theory of Change (ToC) of the programme and by doing so presents the findings of the evaluation on how the 3DE pilot operated and whether it met its intended objectives. The implications of the findings for scaling up of the pilot and its underlying ToC are also discussed. This report was prepared by Professor Sophie Witter, Andrew Kardan, Molly Scott, Lucie Moore, Denis Wood and Louise Shaxson. The report also benefited from the peer review inputs of Stephen Jones from OPM and the management group at DFID.

This evaluation is being carried out by OPM in association with ODI. The evaluation is led by Professor Sophie Witter and the project manager is Andrew Kardan. The remaining team members are Molly Scott, Lucie Moore, Louise Shaxson and Denis Wood. For further information contact Andrew Kardan at Andrew.kardan@opml.co.uk.

The contact point for the client is David Rinnert at D-Rinnert@dfid.gov.uk. The client reference number for the project is 40091028.

Oxford Policy Management Limited
Registered in England: 3122495

6 St Aldates Courtyard
38 St Aldates
Oxford OX1 1BN
United Kingdom

Tel +44 (0) 1865 207 300
Fax +44 (0) 1865 207 301
Email admin@opml.co.uk
Website www.opml.co.uk

Executive summary

Background: The programme

The 3DE model was designed by the Clinton Health Access Initiative (CHAI) and IDinsight, and was based on the recognition that: a) Ministry of Health (MoH) officials often lack evidence on the most effective and efficient ways in which to deliver known clinical interventions and services; and b) where evidence is generated, it is often not relevant to the operational needs of MoH officials or done within a period that meets decision-making timeframes. The 3DE pilot model was designed to facilitate a more demand-driven approach to the evaluation of health interventions by identifying relevant, suitable and priority evaluation question from the ministries, conducting these rigorously but rapidly and in an affordable manner, catalysing the response to its findings and sharing the lessons learned from this process more widely, so as to influence future evaluation processes.

Under this pilot 3DE was expected to involve eight (later revised to five) impact evaluations that influence managerial decisions in six instances (later revised to four). The pilot had a budget of £2 million.

Purpose of the evaluation

The overall aim of this evaluation was to refine and test a more elaborate ToC, based on the existing design and activities of the programme, and look at the quality of the evaluations. In doing so, the evaluation aimed to understand what outputs and outcomes were achieved by the programme and how the model can be further refined and improved in future. Using evidence gathered through this process an assessment was made as to whether 3DE, as a pilot model, has been successful in supporting and increasing evidence-based policy-making and in building the capacity and changing the behaviour of Ministry staff in terms of them demanding and using evidence.

The overall lessons from the evaluation are expected to inform the Department for International Development's (DFID) future roll-out of this or related initiatives aimed at supporting supply and/or demand for evidence uptake. The main users of the evaluation are DFID (specifically the Evaluation Department, the Research and Evidence Division, and evaluation advisers), CHAI (3DE Management – who have prior knowledge of the 3DE model) and other 3DE partners, in particular the relevant individuals within the Zambian and Ugandan governments who were directly engaged with the model.

Methodological approach

The evaluation takes a theory-based approach, starting from the extended ToC and then seeking to establish, for each of the main domains: (1) what happened in practice (what activities were undertaken by the programme and what were the responses of Ministry and other stakeholders); (2) why what happened took place (particularly the role of the 3DE intervention but also any other relevant factors); and (3) with what results. These findings can then be compared with what was planned in the original programme documents and what the ToC outlines. This analysis will test and refine the ToC for future evaluation work, including the next phase of 3DE, and by so doing answer evaluative questions about the 3DE programme itself. The evaluation incorporates a number of analytical approaches, integrated into the ToC framework, including quality assessment of the evaluations themselves, a rigorous theory-based approach to test and validate the ToC and its underlying assumptions, and a Political Economy Analysis (PEA), all supported through the

conducting of 46 key informant interviews (KIIs) and reviews of over 170 documents related to the 3DE programme and beyond.

Some important limitations are noted, including the relatively short timeframe of the programme, which does not allow for the assessment of health outcome changes or the necessary maturation of catalysation activities. As a result, this evaluation focuses on processes more than outcomes, The Ugandan experience has also received more limited analysis compared to the Zambian one, given that no evaluation has yet been completed in Uganda.

Main findings

Question sourcing

The question-sourcing process was intensive in terms of effort and took longer than expected in both countries. In Zambia a broader entry point was established, while in Uganda 3DE initially worked exclusively with the malaria programme. Ultimately the former approach appears to have been more successful as the targeted four evaluations are being delivered in Zambia, while in Uganda a large number of 'false starts' occurred, and ultimately only one question was sourced, after the switch to working with the HIV programme. There were a number of factors involved, but one issue which is clear is that meeting the different criteria for evaluation questions is demanding.

3DE worked closely with Ministry partners (the MoHs in both countries, but also the Ministry of Community Development, Mother and Child Health (MCDMCH) in Zambia) to source questions but partners found it harder to engage in the prioritisation of questions that involved more technical issues. Where questions were not suitable for impact evaluations, there is no evidence of 3DE connecting its partners with other research organisations.

Evaluation design, conduct and reporting

Evaluation quality was assessed against the relevancy of questions, the appropriateness of design, and the quality of conduct and reporting. The research questions posed by the evaluations are all shown to address relevant healthcare challenges in Zambia and Uganda, and in at least one case the evaluation was timed to meet an important opportunity (a large scale bed-net distribution). The rationale for the particular interventions evaluated in each study is not always well described in the evaluation reports (including a description of underlying challenges and how the intervention mechanism is expected to address them). The overall quality of the design of the 3DE evaluations was assessed as variable, with some weaknesses stemming from the constraints placed on the evaluations in terms of timeframes and budgets. One aspect of the evaluation design that was consistently strong for all evaluations was the choice of primary outcome given the available study period. The evaluations all focused on measures that could be plausibly expected to change over a period of months if the intervention was effective. Although all evaluations did make an appropriate choice of primary outcome, there were some issues with the indicators used to track these outcomes.

The evaluations were also well designed to make efficient use of the available budget and were aligned to a large degree with current practices in health facilities. The overriding concern with the design of 3DE evaluations is that the findings were not easily generalisable to other contexts (i.e. there was a problem with findings' low external validity). Many of the evaluations were only able to cover a limited geographic area and a small sample. There is also a concern that some of the 3DE evaluations may not have delivered sufficient internal validity (particularly the Decongestion study) despite their randomised design and given the small size of the treatment groups. The time and budget constraints also affected the implementation of the interventions themselves, which in some cases may have been too ambitious for a short evaluation period.

In view of some of the concerns outlined above it is not clear that the choice of a Randomised Control Trial (RCT) always made the best use of the available budget. In some cases a simple operational pilot or process study may have provided sufficient evidence around the implementation of interventions to help guide future programming decisions. This is particularly the case for interventions that sought to reinforce existing practices rather than providing new and previously untested solutions (such as the Early Infant Diagnosis (EID) Simple Intervention and the Decongestion intervention).

The evaluations appear to have collected good data using appropriate techniques. Where the data-collection processes are reported on it seems that the processes were good. Sample sizes were an issue for some of the evaluations. Quantitative findings were for the most part presented well. However, there were also some important weaknesses in this respect. The discussion section in each technical report generally reflected the quantitative findings well but there were some ways in which the description and interpretation of findings could have been improved. The explanation and interpretation of results could also be further developed, with the overall findings better situated within a broader discussion of the context and likely mechanisms involved.

Dissemination and activities to catalyse implementation

3DE generally has a good awareness of entry points and key stakeholders and disseminated well to key stakeholders, largely at the programme level. However, in order to provide rapid feedback, presentations preceded finalisation of reports, which has some risks. Thus, the ensuing 'policy decisions' (for the three completed evaluations) took the form of advisory notes. There has been limited scope for 'catalysation' work (3DE providing supporting models, costing and plans for scale-up) and uptake has been limited to date.

3DE did not have a specific capacity-building plan beyond working closely through the stages of the programme with MoH/MCDMCH partners. Interviews indicate that individuals who worked closely with 3DE did benefit in terms of capacity development. More broadly, there is an expression of latent demand for evidence, although not necessarily for evaluations specifically. Both ministries (MoH and MCDMCH) lack a wider strategic approach to evidence and research, and there is no indication that this has changed as a result of 3DE.

Key stakeholders did not always have a clear understanding of the findings and questions about the external validity of results for other areas of the country and in 'normal' health system conditions were raised. Ownership of findings was partial. Limited staff time, a lack of capacity in terms of research staff in key partner agencies, and a lack of incentives for evidence use were some of the factors behind this.

Explanatory factors

A number of issues are highlighted related to context and internal factors that have contributed to the programme outcomes. Among the contextual factors, the lack of an effective strategic prioritisation of evidence-based decision-making is highlighted as a constraint, along with unclear roles in Zambia (linked to the split of the MoH into two ministries in 2012), the fragmented supply of research, and its continued dependence on external funding. Internal factors include positive ones, such as a strong starting base for CHAI, which was well embedded in the MoH, as well as negative, such as the partnership breakdown in 2014 and an initial understaffing of the 3DE Uganda programme.

Conclusions

The overall evaluation question was whether the 3DE model has been successful in its stated goal of supporting and increasing evidence-based policy-making, building capacity and changing the behaviour of Ministry staff in regard to demanding and using evidence. The answer, based on the evidence available to the evaluation team, and given the current stage of the programme, is that there has been very limited contribution to changing evidence-based policy-making, capacity and behaviour in both countries. The main reasons behind this limited impact are judged to be two-fold:

1. This goal was inherently over-ambitious for a three-year pilot. The overall goal, particularly in terms of building capacity and changing behaviour, requires a longer timeframe; and

2. The programme had a number of aspirations that were not all compatible with one another.

3DE aimed to be demand-led, focused on robust impact evaluations, rapid/responsive and affordable, as well as catalysing action. A number of tensions or trade-offs exist within and between these aspirations. The overall lesson from the pilot, according to the evaluation team, is that even a very professional partnership cannot deliver on all of these in contexts like Zambia and Uganda, which are relatively typical or even amenable to the use of evidence in decision-making in the health sector. None of this implies that these trade-offs were badly managed by 3DE, but there needs to be reflection on which are most important and how to set realistic priorities for the next phase of the programme. It is also important to clarify what 'demand-led' really means. In the evaluation team's view, the 3DE model is responsive to demand but until there is a much higher level of evaluative thinking and capacity within the MoH/MCDMCH, what 3DE provides is still effectively a supply-side activity.

Recommendations

We provide 10 overall recommendations, mostly relating to design and aimed at DFID:

Agree on the focus and design accordingly. In the next phase, it will be important to agree on the core objectives of the programme, and tailor it accordingly. Different objectives – such as capacity building, improving the supply of evidence, improving service delivery, and generating demand for evidence – imply different models.

Tailor to the context. Clearly not all countries will have the same evidence needs and so a starting point for programming should be an understanding of the local institutional and market context, to understand what the gaps are and what existing institutions or networks could be strengthened.

Invest more in evaluative thinking and capacity. Capacity building was an intended indirect benefit in the pilot phase but should receive more priority to ensure a lasting legacy. The legacy of the programme should be increased evaluative thinking and capacity within MoH and MCDMCH to scope, oversee, quality assure and use evaluations.

Embed in local institutions. Whatever the focus chosen, the programme should be embedded in local institutions, with support provided externally as needed but with the key staff who are commissioning, providing, coordinating or brokering based within the Ministry or local research networks and organisations. This would also allow more flexibility about seizing policy 'windows', rather than having to identify them within the constraints of a short-term programme.

Change the performance targets. In the 3DE programme, contributing to a policy decision was a key performance target. While this kept minds focused on the need to ensure take-up of research, there is also a potential conflict of interest between being a supplier of research and helping

ministries to analyse and use evidence in a neutral way. More specifically, if contribution to a policy decision is used as a target, then it should be broadened to include implementation.

Enlarge the toolkit. We question the privileging of impact evaluations as a higher form of knowledge. They have their own limitations, particularly in terms of generalisability, and often fail to provide good insights into the 'how, why and in what contexts' questions. Ministries rightly look for a range of information, including on equity, sustainability, etc. of interventions. Demand-generation or evidence-supply programmes should focus on supporting and providing appropriate tools for different questions.

Timeliness, not rapidity, should be the goal. Evidence should fit with policy needs, but rapidity has costs and is not always required or appropriate to the question. Timeframes should follow on from the question for which the MoH needs an answer – not dictate the question. In some cases, having a longer time period would generate more useful and valuable information for the MoH than one with artificially constrained fieldwork periods.

Monitor value for money (VfM). Information on expenditure in the 3DE programme was not reported for the different stages of the programme, with the result that the cost-efficiency of different stages could not be assessed. In the next phase, this information should be systematically reported.

Ensure quality assurance at all relevant stages. In the pilot programme, the peer review of products appears to have been at the stage of developing protocols, while at the report-writing stage there was no quality assurance process that the evaluation team is aware of. Peer reviewing of final products is important to ensure that findings are robust and accurately presented.

Take a broad approach and ensure adequate support. The differential success in Uganda and Zambia – both environments judged initially receptive to an evidence-based approach – suggest some practical lessons for the next phase, including the wisdom of taking a broad approach to ministerial needs (rather than being locked in to relationships with specific programmes) and also of ensuring adequate staffing to drive forward what has been an intensive process, if a similar approach is adopted.

Table of contents

List of figures, tables and boxes

List of abbreviations

3DE	Demand-Driven Evaluation for Decisions
AAT	Applied Analytics Team
ANC	Antenatal Care
ART	Antiretroviral Therapy
BCC	Behaviour Change Communication
CHAI	Clinton Health Access Initiative
CHW	Community Health Worker
CP	Cooperating Partner
DBS	Dried Blood Spot
DPME	Department of Planning, Monitoring and Evaluation (South Africa)
EID	Early Infant Diagnosis
EPI	Expanded Programme of Immunisation
FCD	Family Clinic Day
GEFA	Global Evaluation Framework Agreement
IRB	Institutional review board
ITN	Insecticide-Treated Net
KII	Key Informant Interview
M&E	Monitoring and Evaluation
MCDMCH	Ministry of Community Development, Mother and Child Health
MDSR	Maternal Deaths Surveillance and Response
MoH	Ministry of Health
NHRAC	National Health Research Advisory Committee
ODI	Overseas Development Institute
OPM	Oxford Policy Management
PEA	Political Economy Analysis
PI	Principal Investigator
PMO	Provincial Medical Officer
PS	Permanent Secretary

RCT	Randomised Control Trial
SAGs	Sector Advisory Group Meetings
SEQAS	Specialist Evaluation and Quality Assurance Services
ToC	Theory of Change
ToR	Terms of Reference
TWG	Technical Working Group
VfM	Value for Money
WHO	World Health Organization

1. Introduction

1.1 The 3DE programme

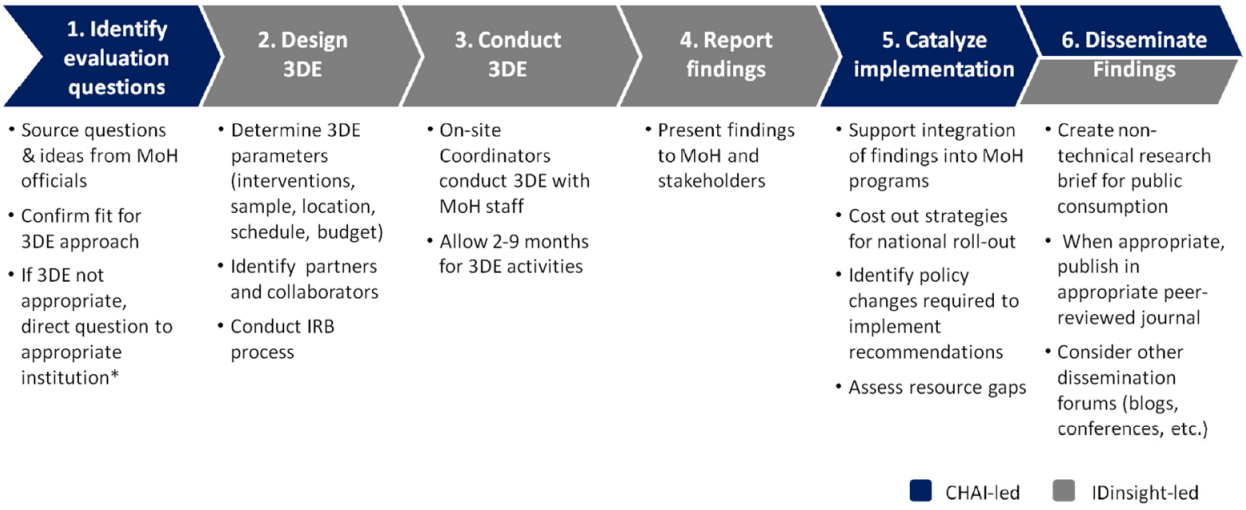
The 3DE model was designed by CHAI and IDinsight based on the recognition that: a) MoH officials often lack evidence on the most effective and efficient ways in which to deliver known clinical interventions and services; and b) where evidence is generated, it is often not relevant to the operational needs of MoH officials or done within a period that meets decision-making timeframes (CHAI 2012).

With this in mind the 3DE pilot model was designed with the aim of undertaking a more demand-driven approach to evaluations of health interventions, which may result in better use and uptake by policy-makers and a change in how MoHs think about using evaluations and general evidence in policy and programme formulation. The model aimed to achieve this by implementing the following objectives:

- Identifying priority evaluation questions from the MoH, which are relevant, appropriate for programme managers, and suitable under the 3DE model;
- Conducting rigorous impact evaluations that address the identified evaluation questions in a timely and affordable manner;
- Catalysing country-level action in response to evaluation results and findings; and
- Sharing evaluation learning more widely and refining future evaluation processes.

The main stages of the 3DE model and related activities[1] are summarised in Figure 1.[2] It is this pilot and its related activities that are subject to this evaluation.

Figure 1 3DE evaluation model



Source: CHAI/IDinsight proposal

[1] See DFID Business Case for more details of the activities.
[2] IDinsight was involved during the first three evaluations undertaken by the pilot. For the last two evaluations the roles allocated to IDinsight were taken over by CHAI.

The 3DE model was piloted in Zambia and Uganda by CHAI and IDinsight with a total budget of £2 million. It targeted the MoH and subsequently MCDMCH in Zambia and the MoH in Uganda, initially through the National Malaria Programme and later on the AIDS control programme.

The pilot was originally expected to produce eight impact evaluations of health interventions, generated by demand from MoHs, and for the evaluations to influence managerial decisions in six instances. The evaluations were expected to be completed and presented to policy-makers within nine months of the commencement of the evaluations. The logframe of the pilot was subsequently revised, with the pilot then expected to complete a minimum of five impact evaluations and for them to influence a managerial decision in four instances.

1.2 Purpose of this evaluation

The overall aim of this evaluation is to assess whether the 3DE model has been successful in supporting and increasing evidence-based policy-making, building capacity and changing the behaviour of Ministry staff in terms of demanding and using evidence. The evaluation primarily aims to do this by refining and testing a more elaborate ToC, based on the existing design and activities of the programme. In doing so, the evaluation aims to understand what outputs and outcomes were achieved by the programme and how the model can be further refined and improved in the future. The overall lessons from the evaluation are expected to inform DFID's future design of this pilot or other initiatives aimed at supporting supply and/or demand for evidence uptake.

The main users of the evaluation are viewed as DFID (specifically the Evaluation Department, the Research and Evidence Division, and evaluation advisers) and CHAI (3DE Management), who have good knowledge of the 3DE pilot, and this report has been produced with these people in mind. Given that one of the main objectives of the evaluation was the refinement, testing and elaboration of the ToC, the evaluation findings are structured and presented along the main stages of the pilot model and across the ToC as laid out in our inception report. Answers to the evaluation questions listed in our evaluation framework are drawn out in the concluding sections of the report.

The report will also be useful to the other 3DE partners, in particular to the relevant staff in the targeted ministries within the Zambian and Ugandan governments who were directly involved with the 3DE pilot and who are knowledgeable about the model. The process of developing the evaluation approach has involved consultation with DFID and initial talks with CHAI. The evaluation is timed to fit with the end of the 3DE programme, which ran from 2012 to 2015. Consultations on the draft were held with DFID and CHAI and comments incorporated into this final report. A summary brief on the report will be produced in due course to inform DFID's operations for the second phase of the pilot.

As laid out in the technical proposal and noted above, the questions in the Terms of Reference (ToR) will be addressed (section 4.1), with the exception of the question relating to impact on global awareness and the VfM component, which it has been agreed are not feasible to address in a robust fashion in the time and with the resources available. Other questions have been further refined following comments on the initial drafts of the inception report. The limitations of this evaluation are articulated in section 2.4.

1.3 Structure of the report

The remainder of this report is structured as follows:

- **Section 2** describes the approach, methodology and conduct of this evaluation, including its limitations.

- **Section 3** presents the main findings of the report across the key stages of the pilot model as presented in Figure 1, and looks at the implications of the findings on the revised ToC. Sections 3.2 to 3.4 look at the key stages of the model. Section 3.5 and section 3.6 look at the ToC at outcomes and impact level, section 3.7 looks the external factors influencing the 3DE model, and section 3.8 looks at the implications of the findings for the refined and elaborated ToC.

- **Section 4** concludes and provides a set of recommendations for the future implementation of the model.

The report also includes a number of annexes that provide further detail on the wider context and evidence used for this study. **Annex A** provides the original ToR of this evaluation; **Annex B** lists the name of key informants consulted as part of this evaluation; **Annex C** presents our evaluation framework and the list of evaluation questions it aimed to answer through the testing of the ToC of the programme; **Annex D** presents the refined ToC following lessons learned through this evaluation; **Annex E** looks at the details of our assessment of all 3DE impact evaluations; **Annex F** describes the political economy of decision-making within the MoH and its implications on the 3DE model; and finally **Annex G** summarises the review of literature on the role of evidence in shaping policy-making and the factors associated with evidence uptake. It also summarises some of the other initiatives aiming to build capacity for evidence use and uptake.

2. Evaluation design

2.1 The evaluation framework

The evaluation framework is structured around the two main objectives of the evaluation,[3] namely:

- Refine and develop the ToC and test its underlying assumptions and causal mechanisms; and
- Assess the quality of the 3DE pilot products.

In undertaking these two objectives the evaluation aimed to answer a number of subsidiary evaluation questions relating to the relevance, efficiency, effectiveness and impact of the pilot. The evaluation framework detailing these questions and the approach undertaken in answering them are summarised in Annex C.

2.2 Analytical approach

The evaluation takes a theory-based approach, starting from the extended ToC and then seeking to establish, for each of the main domains: (1) what happened in practice (what activities were undertaken by the programme and what were the responses of Ministry and other stakeholders); (2) why what happened took place (particularly the role of the 3DE intervention but also any other relevant factors); and (3) with what results (intended and unintended). These findings can then be compared with what was planned in the original programme documents and what the ToC outlines. This analysis will test and refine the ToC for future evaluation work, including the next phase of 3DE, as well as answering evaluative questions about the 3DE programme itself.

The evaluation incorporates a number of analytical approaches, integrated into the ToC framework, including:

- A theory-based approach that drew on contribution analysis, to provide us with a structured approach to understanding the role of the intervention alongside other factors that may have influenced the processes, outputs and outcomes listed in the ToC;
- Quality assessments of the evaluations themselves, adapting existing Specialist Evaluation and Quality Assurance Services (SEQAS) and Global Evaluation Framework Agreement (GEFA) checklists – these are relevant to answer questions about robustness, which are built into the ToC; and
- PEA, to understand the context in which the 3DE work has taken place and how this influences and explains the processes and results the evaluation will document.

It was not possible to establish a credible counterfactual in this evaluation. However, interviews and document analysis are used to explore qualitatively how the 3DE pilot differed from what existed before, and what stakeholders consider would have happened in its absence.

The evaluation considers the programme as a whole, including consideration of how it was established and the prioritisation of evaluation questions. However, each individual evaluation

[3] A third objective looking at the VfM of the evaluation was dropped during the writing of the inception report with DFID agreement.

conducted by 3DE constitutes a case study within the overall framework, and these can be compared in order to generate a richer understanding of differences and similarities.

As most of the work has been undertaken in Zambia, this has been the focus of the evaluation; nevertheless, a modified set of questions was used in Uganda to learn from the experience there. The evaluation will therefore be able to draw from evidence from two national settings and across five different evaluations.

Mixed methods were used, combining a review of the literature with structured analysis of programme documents, expert opinion and KIIs. A structured comparison of sources allowed for the triangulation of findings in regard to most of the evaluation questions.

2.3 Data collection and analysis

Full details of the data-collection tools and methods are provided in the inception report. In brief, five main sources of evidence were used for the evaluation:

1. Quality assessment of five evaluations

The quality of the 3DE evaluations was assessed against a set of specific questions covering the following dimensions: *Planning and Context*, *Introduction*, *Methods*, *Data*, *Evaluation Conduct* (i.e. data collection, entry and cleaning), and *Analysis and Reporting*. A set of questions was developed drawing from:

- the quality assurance templates used under DFID's SEQAS, in combination with;
- the quality assurance templates OPM developed as part of the quality assurance processes for an OPM-led consortium (ePact) undertaking evaluations under DFID's GEFA; and
- the authors' own knowledge of quality RCTs.

It was applied to all five evaluations conducted by 3DE, although only three have been finalised and so can be examined in relation to all quality assurance questions. Details of the assessment are found in Annex E.

2. KIIs

Forty-six KIIs were conducted, using semi-structured topic guides. The participants were:

- 16 3DE personnel (10 from CHAI and six from IDinsight);
- 13 personnel from government ministries in Zambia;
- Four personnel from government ministries in Uganda;
- Five DFID personnel; and
- Eight personnel from international donor agencies and other institutions in Zambia.

The interviewees were sampled based on a set of criteria provided during the inception phase and applied to an initial list of stakeholders provided by CHAI. For Zambia the list of interviewees was supplemented with additional individuals whose name we came across either through review of programme documents or through initial interviews and in consultation with our local political

economy analyst. In Uganda the list of interviewees was significantly smaller and most were based on stakeholder list provided by CHAI (See 2.4).

The interviews were thematically coded and entered into an analysis spreadsheet, which was structured according to the key nodes in the ToC. By reading down the columns, views from informants and evidence from programme documents could be triangulated on each topic, and a summary of evidence created.

3. Document analysis

Over 170 programme documents of various types were read, with around 60[4] thematically coded and entered into the same spreadsheet used for the KIIs. An important caveat is that much of this evidence was generated by the project and may have therefore been biased toward demonstrating progress and success. Internal evidence is given weight in the summary description of what was done and why. External evidence is given more weighting in the final evaluative judgements.

4. Literature review

A literature review was undertaken to understand the background and wider global context for the programme. This focused on a number of topics, including:

- Evidence on general experience and the efficacy of demand-led evaluations;
- Evidence on evaluation use and uptake by policy-makers;
- Evidence on formal and informal barriers and enablers for conducting and using impact evaluations by policy-makers/government officials;
- Other global initiatives and experiences of demand-led evaluations; and
- Reviews of ongoing and recent impact evaluations in the health sectors of Zambia and Uganda.

The body of literature reviewed included research papers, articles, theoretical discussion papers and synthesis reports drawing together the findings of other work. A summary of findings is found in Annex G.

5. Political Economy Analysis

A Political Economy Analysis of the health sector in Zambia was conducted to better understand the contextual factors influencing the outcomes of the 3DE model with particular focus on resource allocation and decision making and use of evidence in the policy making process. The findings from the PEA study are summarised in Annex F.

2.4 Evaluation limitations

Some important limitations are to be noted, including the following:

- the timeframe of the programme (with only three evaluations having completed their full cycle relatively recently) does not allow for the assessment of health outcome changes or necessary

[4] Many of the documents were received toward the end of the analysis stage, and while they were read and incorporated into our analysis were not codified given the time constraints.

maturation of catalysation activities; the evaluation therefore focussed on process more than outcomes in the end.

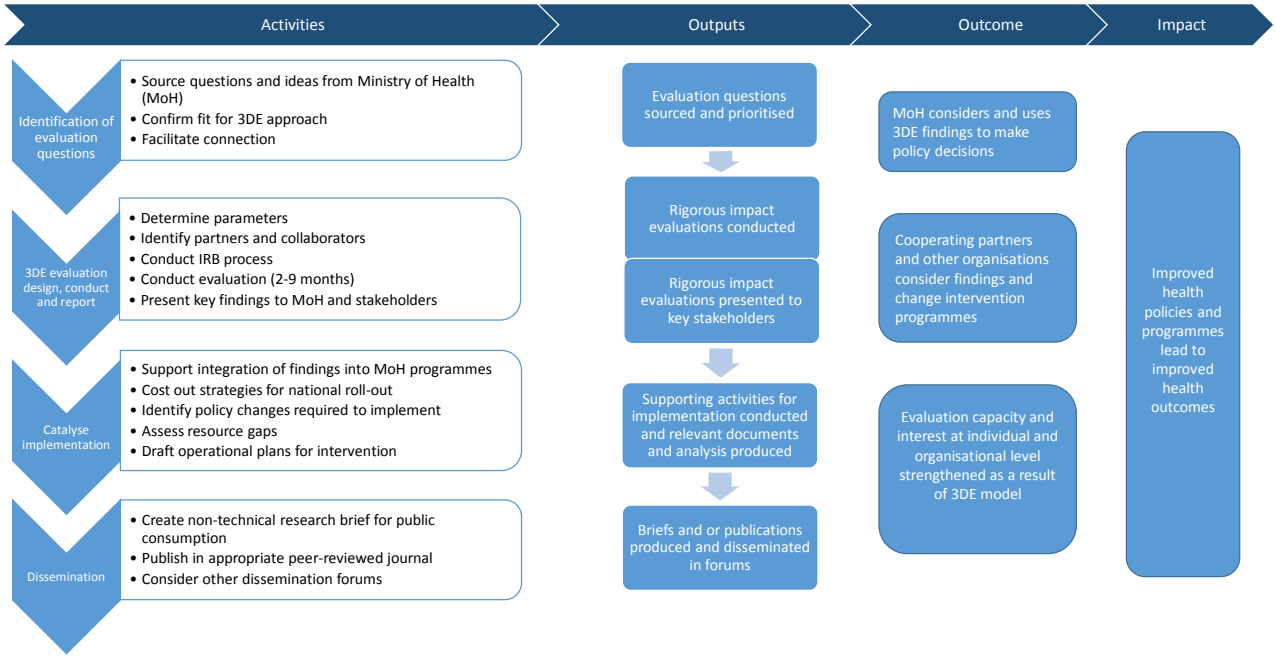
- after discussion with DFID at the inception stage, the VfM element of the ToR was removed, given the absence of credible counterfactuals and the ambitious scope of the evaluation and short timeframe;
- the question in the ToR on global impact was not examined, due to the limited resources of the evaluation and also the lack of sufficient time for global impact to be achieved;
- judgements are based on documents made available to the evaluation team, which may not represent the entire body of 3DE documentation;
- similarly, not all desired key informants were available (though most of the originally targeted people were interviewed); in some cases, questions had to be prioritised according to available time;
- there is always a risk of capture of respondents by the programme; we mitigated this by identifying additional individuals through the review of documents as well through interviews, at least in the case of Zambia; and
- for Uganda, a more limited approach was taken, given that no evaluations have yet been produced there, so the sample size for interviews was smaller.

3. Main findings

3.1 Overview

In the inception report, an elaborated ToC was developed, which mapped the stages of the programme to outputs, outcomes and impact, while identifying a range of underlying assumptions or preconditions for effectiveness at each stage (see Figure 2). We examine the 3DE programme stage by stage in relation to these, summarising the evidence on achievements in Uganda and Zambia, and for individual evaluations where relevant, compared with the expectations of the framework. We conclude the findings by discussing the contribution of 3DE, the explanatory factors behind successes and failures (external and internal to the programme), and the implications for future revision of the ToC.

Figure 2 Programme ToC



The ToC of the programme stipulates that findings from evaluations that are based on questions raised by MoHs, are conducted in close collaboration with them and that produce timely well-presented results are more likely to be adopted into policy. Improved evidence uptake is expected to result in the implementation of better health policies and programmes that will ultimately improve health outcomes for the population. Supporting the Ministry in following through with evaluations and assisting them in implementing its recommendations are seen, in addition to the relevance and timeliness of the evaluations, as an important feature of this pilot.

3.2 Identification of evaluation questions

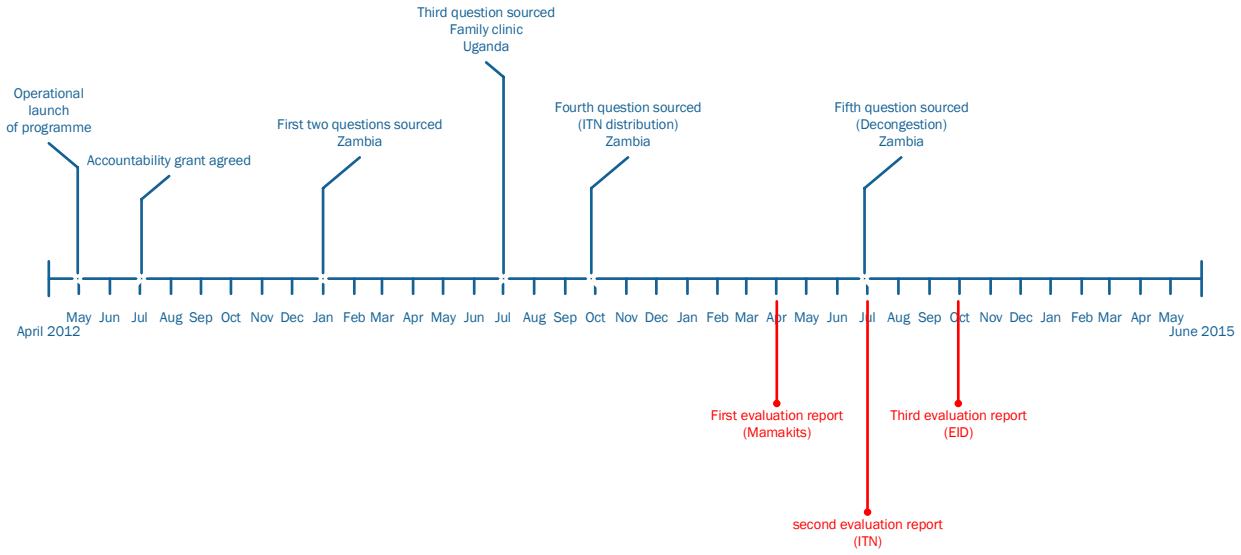
3.2.1 The ToC

Under the 3DE model the first major activity is the sourcing of evaluation questions from the relevant ministries and other national agencies responsible for health policy and implementation (in Zambia this included the MCDMCH, which since 2012 has been tasked with managing the implementation of district health services).

3.2.2 What happened

Question sourcing was a time-consuming and intensive process. In Zambia, IDinsight and CHAI staff worked full time on sourcing questions from July 2012 when the Accountable Grant with DFID was signed and the programme was first formally introduced to the MoH in Zambia to January 2013, when the first two questions were sourced (seven months). In Uganda, it took 13 months, and not until July 2013 was a question finally agreed (the Family Clinic Day – FCD) that could be taken to fruition. In Zambia, questions were agreed in months 7 and 15. A fourth question was not identified until the 24th month (Figure 3). The question-sourcing process was very intensive and thorough, with 3DE staff holding numerous meetings with Ministry officials, programme managers and Cooperating Partners (CP) to come up with impact evaluation questions.

Figure 3 Timeline of 3DE pilot



Entry processes differed across the two countries. In Uganda, the National Malaria Control Programme was the initial focus of the question-sourcing process, and later the AIDS control programme. By comparison, in Zambia the Director of Public Health and Research[5] was approached and an overall list of questions provided. This allowed 3DE to address questions from different programme areas and ultimately allowed more evaluations to be conducted.

The MoH was thoroughly engaged on questions but the involvement of the MoH/MCDMCH in regard to prioritisation was more challenging. In Uganda there was no central list of questions used to initiate question-prioritisation discussions, but many meetings were held in both countries to engage stakeholders with the 3DE concept. In Zambia, a list of priority research questions had been produced with the support of the World Health Organization (WHO) in 2011, and this was used as a starting point for discussions. However, the Ministry's questions were perceived as a 'list of problems' rather than a list of research questions, and the process of transforming them into viable impact evaluation questions involved activities with which MoH staff could not easily engage (literature reviews, consulting experts, assessing required sample sizes, modelling potential impacts, etc.). Consequently, this was something of a 'black box' for non-3DE stakeholders. They were happy about the questions that were taken forward but were not always

[5] This directorate was later renamed to Disease, Surveillance, Control and Research.

clear about the process or why certain questions had been dropped or amended from their original formulation. The final questions addressed did not relate to the original MoH question list. Only one overlapped with the questions prioritised by the MoH during the 3DE process (see Figure 4).

> *Prioritisation was done by CHAI but involving consultations with programme managers. We wanted to be demand-driven but ended up having to drive the process more. There weren't 'off the shelf' ideas in the Ministry. (KII, 3DE, Uganda)*

In Uganda, the 3DE programme officer applied a set of criteria to shortlisted questions in an intuitive way. The criteria were sensible but the process did not involve any MoH participation, either in terms of the development of the criteria or their application.

There were a number of 'false starts'. A number of questions were developed, and taken through an intensive process of development in some cases, before being dropped. The reasons reveal some of the challenges faced by the programme in finding questions that met all of the criteria. Common reasons for questions not being taken forward (see Table 1 and

Table 2) were: the question was not suitable for an impact evaluation; more background research or situation analysis was needed first; it related to a low perceived Ministry priority; and the evaluation or policy change was already happening/had happened and the intervention was changing.

Figure 4 Evolution of question sourcing in Zambia

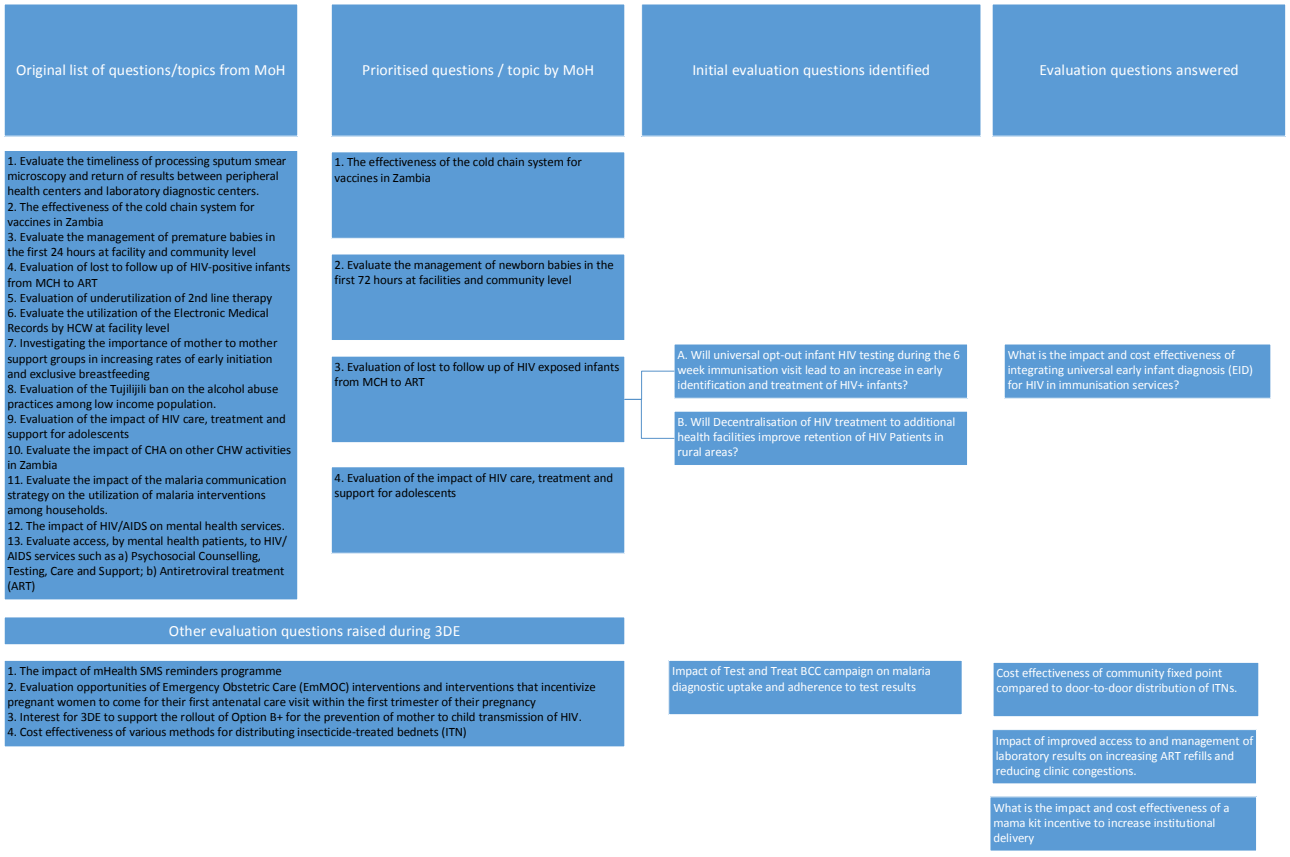


Table 1 Summary of questions sourced but not pursued in Uganda

Question sourced	Reasons why question not pursued
Impact of the Test and Treat Behaviour Change Communication (BCC) campaign, targeted at patients on uptake and adherence to malaria diagnosis and treatment.	Evaluation of a programme to deliver BCC targeted at patients through radio ruled out since power calculations indicated that there were not enough radio stations available to answer the question through a rigorous impact evaluation.
Impact of training of facility staff on malaria treatment outcomes.	Interpersonal communication training intervention for health facility staff considered weak; large impacts not anticipated. Implementers changed interventions in such a way as to make them less suitable for impact evaluation. A more general training programme that was considered was not taken forward as it was perceived to have low external validity to other disease areas, limiting the usefulness of the evaluation in terms of informing MoH decisions. Perception of low appetite from some donors to fund training programmes.
Impact of different time points (six or 15 weeks, five months) for transitioning HIV positive mothers' antiretroviral therapy (ART) from antenatal care (ANC) clinic to ART clinic on patient retention and drug adherence.	Protocols had been developed and design process for the evaluation was underway. MoH policy change (to implement a mother and baby care point, as below) meant that the evaluation question was no longer relevant. Policy change was more substantial than the intervention that 3DE had been proposing to test.
Evaluation to introduce an HIV mother–baby care point at maternal and child health clinics where HIV positive pregnant and breastfeeding mothers and children could receive ART services before being transitioned to the ART clinic at 18 months.	MoH moved ahead with policy change, having done own assessment.
Using village health teams to deliver prophylaxis malaria medicine to pregnant women in villages.	DFID reluctance to mix up funding protocols, as CHAI already had DFID funding to implement the programme separately from 3DE.
Evaluation opportunity relating to cryptococcal meningitis.	A first set of consultations revealed that there was no implementation interest or capacity among partners.
Evaluation of SMS technology intervention to remind care givers to complete vaccination schedules for infants.	Perception of country director that the SMS/ text platform would not be sustainable.

Source: interviews and documents; not comprehensive

Table 2 Summary of questions sourced but not pursued in Zambia

Question sourced	Reasons why question not pursued
Impact of SMS reminders (sent to patients or community health workers (CHWs)) on post-natal care attendance rates at six days and six weeks.	It was found, after about six months, that the intervention had already been scaled up in several parts of the country. There were also questions around the scalability of the component of the intervention that included SMS reminders sent directly to mothers. Ethical concerns were raised around evaluating the programme using an RCT, since this would involve withholding services from some women. Yet it was felt that a non-randomised design would not yield robust findings.
Evaluation of the malaria communication strategy.	Difficult to randomise the mass communication methods used in the strategy within the time and resources available under 3DE.
Effectiveness of the cold chain system for vaccines in Zambia.	This evaluation opportunity was among the early list of prioritised questions in Zambia from September 2012. Difficult to identify a specific evaluation question since little information was found to exist about the current performance of the cold chain. 3DE felt that a situation analysis of the problems in the cold chain would be required before an impact evaluation could be considered. Also foresaw difficulties around assessing measles outbreaks as the key outcome, and attributing changes in this to failures in the cold chain.
Evaluate the management of newborn babies in the first 72 hours at facility and community level.	This evaluation opportunity was among the early list of prioritised questions in Zambia from September 2012. UNICEF and WHO had already released guidelines indicating which interventions related to newborn management are effective, so additional impact evaluation evidence not expected to be a high priority for the MoH. For other questions in this topic area, clinical and observational research was felt to be more appropriate

Question sourced	Reasons why question not pursued
Evaluate the loss to follow-up of HIV exposed infants from Mother and Child Health to ART.	This evaluation opportunity was among the early list of prioritised questions in Zambia from September 2012. Three questions explored: 1. community engagement in HIV treatment 2. Will universal opt-out infant HIV testing during the 6 week immunisation visit lead to an increase in early identification and treatment of HIV+ infants? 3. Will Decentralisation of HIV treatment to additional health facilities improve retention of HIV Patients in rural areas? For the first question a concept notes were drafted for the evaluation questions in October 2012, but later conversations with the National ART Coordinator in July 2013 indicated that the question was not a high priority for him. Concern about the scale-up potential of the programme and retention of community volunteers. Question 2 was taken up and question 3 was not taken forward. The last question over time morphed into the question on decongestion in the urban areas.
Evaluation of underutilisation of second line therapy.	Initial review of the area suggested that the question would not be appropriate for an impact evaluation. More research needed to understand why second line therapies are being underutilised before a suitable intervention could be developed.
Evaluate the impact of HIV care, treatment and support for adolescents.	This evaluation opportunity was among the early list of prioritised questions in Zambia from September 2012. However impact evaluation evidence was felt to be less appropriate than other kinds of evidence, for example information to understand the scale of the problem (which is currently lacking).Concern over scalability of interventions. Fragmented intervention landscape, and no intervention to evaluate identified.
Evaluate the utilisation of Electronic Medical Records by health workers at facility level.	Not a prioritised question. The system under consideration was found to be supported by many partners with complex political interests. Was felt that an evaluation would be better undertaken by a partner with existing involvement in the system.
Investigating the importance of mother-to-mother support groups in increasing rates of early initiation and exclusive breastfeeding.	Not a prioritised question. Not clear whether this was a priority area for the government. Intervention not felt to be high impact or to address the likely barriers to exclusive breastfeeding. Concerns about the scalability of the intervention, since effectiveness of support groups may be dependent on specific characteristics of support workers and clinic settings.

Question sourced	Reasons why question not pursued
Evaluation of the Tujilijili ban on the alcohol abuse practices among low-income populations.	Not a prioritised question. Considered to be difficult to evaluate since the evaluation would have to be retrospective (the ban has already occurred). Therefore an experimental impact evaluation would not be possible. The policy around this question has already been implemented, so there was felt to be low potential for impact evaluation evidence to contribute to a further policy shift. Alcohol abuse not considered a high-priority area for the government. Lack of information about the scale of the problem and specific target populations.
Evaluate the timeliness of processing sputum smear microscopy and return of results between peripheral health centres and laboratory diagnostic centres (TB).	Not a prioritised question. Operational and process-oriented research needed first to uncover more information about the nature of the problem. New technologies in this field are not yet widely used in Zambia, so opportunities for an evaluation are limited. Unclear potential for scale-up in evaluation questions related to TB. Further conversations with stakeholders and partners necessary to determine the potential for response to an evaluation.
Evaluate the impact of Community Health Assistants on other CHW activities in Zambia.	Evaluations were already being carried out in this area. The Community Health Assistants policy has also already been approved and is being rolled out nationally.
The impact of HIV/AIDS on mental health services.	Not a prioritised question. Mental health not felt to be a priority area for government, so potential for new evidence to contribute to a policy shift low. Lack of information about the burden of mental illness in Zambia, especially in combination with HIV. Outcomes of interest for an evaluation not clearly defined. Low implementation capacity; few partners in this area.
Evaluation opportunities in the field of cervical cancer screening and treatment.	No specific evaluation question had been identified at the time the question was suggested. Many other organisations found to be working in this area.

Source: interviews and documents; not comprehensive

Criteria for prioritisation made sense but were demanding. As evidenced in Table 1 and above, 3DE aimed to find questions that were feasible to implement in the timescale, had potential for wide impact, could be addressed through a rigorous experimental or quasi-experimental counterfactual-based impact evaluation, where there was an intervention being rolled out that could be used for the study, and where the chances of catalysing change in the future were strong. It is clear that these criteria are, in combination, very demanding and sometimes in tension with one another, and this in part explains the difficulty of identifying successful questions in Uganda and reaching the targeted four evaluations in Zambia.

The first criterion applied to any proposed question was whether an impact evaluation using experimental or quasi-experimental counterfactual-based designs could be applied or not. This was in line with the original purpose of the pilot:[6]

> *Whenever managerial and operational constraints allow, the 3D Evaluation team will develop a randomized evaluation protocol, as this will produce the most robust findings that are likely to stand up to external scrutiny. However, there may be instances where randomized approaches are not possible and it would require outsized expenditure, time or operational disruption, or randomization is deemed unfeasible for ethical or other concerns. In this situation, other quasi-experimental options will be explored, including regression discontinuity designs, instrumental variables, statistical matching, differences-in-differences, and time-series analyses.*

Where original topics were not suited to the specific evaluation designs the pilot had in mind the questions were either discarded or alternative topics or questions, relevant to the broader issues raised by the Ministry, were explored:

> *In some cases, the original topics suggested by the government did not appear to be appropriate for 3DE, but through consultations with the government and other partners, 3DE shifted to focus on other similar topics. For example, in Zambia the government was initially interested in evaluating the malaria communication strategy. Although this was determined not to be appropriate for the 3DE model because the mass communication methods used in the strategy were difficult to randomise and rigorously evaluate within the time and resource limitations of the programme, 3DE determined that there may be other high-priority questions related to malaria. In consultations with government officials working on malaria, additional questions were raised regarding the impact of different approaches to ITN [insecticide-treated net] distribution on ITN usage, and an evaluation related to this secondary topic was being explored in September 2013.*[7]

> *Yes – there are many outstanding priorities and areas of research that need to be better understood but haven't been picked up before. Whether the most appropriate questions were ultimately selected is a different issue. There are some questions that have been identified repeatedly and addressed in different ways, but don't end up being tackled (such as reasons for high absenteeism of health facility staff/teachers). If you only end up evaluating things that are evaluable by your precise scientific definition, or targeting issues that are well known, then the exercise doesn't address real priorities. (KII, Uganda)*

No connections were facilitated. While it was envisaged that where 3DE could not respond to a research need it would help the MoH to connect with other groups, this did not happen in practice. 3DE staff reported that the other questions were not researchable. Programme staff did not believe there were questions that lent themselves to be covered by other organisations; however, it is not clear whether this was because of their focus on impact evaluation questions and there not being enough impact evaluation questions or whether there were not enough research questions more

[6] CHAI–IDinsight 3DE proposal (July 2012).
[7] CHAI (2013), 'Interim Annual Report', October.

generally. However, in one case, a systematic review and meta-analysis was done for the MoH in Uganda on how to improve rates of Intermittent Preventive Treatment of Malaria in Pregnancy. This was considered to provide adequate information for a decision to be made.

The broader political context has implications for question sourcing. Our PEA analysis (Annex F) has identified a number of factors that were likely to have affected the question sourcing process, in addition to issues identified above. Limited human resources and budgets allocated to research and the culture of handing responsibility upwards makes establishing the demand for evidence from impact evaluations a challenging one, this is further exacerbated by a lack of strategic approach to evidence across the ministry. Under these circumstances the wide consultation process (in the case of Zambia) was appropriate; however the time it necessitates may have been underestimated.

3.2.3 Review of the assumptions

In this section we examine whether the assumptions laid out in the ToC were in fact realised in the 3DE pilot. The assessment of whether the ToC itself was valid, relevant and complete is revisited in section 3.8, when we consider the implications for the ToC of the learning gained through the evaluation process.

Contextual factor – There is sufficient demand in the MoH for evidence, and in particular evaluations

Interviews, document review and analysis of technical working group (TWG) minutes suggest that the MoH/MCHMCH does have a demand for evidence that varies across the different units within and across ministries, but that evaluations are only a small component in the range of evidence required and evidence is not the only factor influencing decisions. Funding cycles, taking advantage of opportunities etc. can over-ride the need for evidence in taking decisions.

The type of evidence often used by the officials related to routine administrative data, fact finding missions and field visits. The evidence that was deemed as most in need was operations research and situation analyses to identify and rectify bottlenecks in delivery or service and research to better understand the behaviour and motivation of end users in regard to non-utilisation, as well as synthesis of evidence from other contexts. The mid-term review of the National Health Strategic Plan (2011–2016) in Zambia emphasises this need and also acknowledges that 'the M&E [monitoring and evaluation] information is used and referenced for justification of selected interventions in the annual work plans'. The report also highlights insufficient use of data at all levels of government and further notes that capacity and funding for research is limited.

Demand for impact evaluation evidence is generally lower, although the health sector is seen by some as more advanced in this respect, with staff having been exposed to clinical trials during training. Key informants in both Uganda and Zambia suggested there was growing pressure to provide evidence to back recommendations for change or new programmes. However, while the Zambia National Health Strategy 2011–2016 highlights the need for evidence-based decision-making, this is presented as work in progress. Ministry staff need more support in developing their skills to hone questions into research questions, as well as funding to pursue them. Much evidence-generation is still commissioned and used by donors.

At the district level in Zambia, districts are encouraged to include research questions in their annual plan, but it is not clear that they are doing this or that this is prioritised when funds are limited:

> *One of the challenges we have in responding to evidence is that we have a limited budget. Currently there is no budget specifically allocated to research – it doesn't sit anywhere.*

> *There is some disconnection in planning: we present our budget to the department of planning and the department of planning presents it to the Ministry of Finance, so we are dependent on how it is presented by the planning department and the MoF's reaction. Since I've been here we have not received the full amount we have requested, and I don't know whether this is because research is not a priority or not. It is confusing because we don't know if the money is there or not. We have had a number of governing declarations, such as 2% of budget of Ministry should go to research, but I don't know if that has ever happened. (KII, MoH, Zambia)*

In both country contexts, key informants feel that there is need to develop an evidence-based culture and stimulate demand for evidence, alongside capacity to meet the need and use the evidence.

Assumption – There is interest from the Ministry to engage with the sourcing of questions

There was engagement from specific individuals within the Ministry in developing the original question lists. However, the Ministry was less involved in transforming the list of question topics identified into impact evaluation questions and it is difficult to assess the level of interest in this process. Engagement took the form of intermittent feedback from and discussion with the 3DE team.

The one question where 3DE appeared to be responding to MoH interest more than leading the selection of the question in Zambia was the Decongestion study – which may however have been less suited to impact evaluation methods (see below). In Uganda, the FCD topic seems to have had active engagement and leadership from its principal investigators (PIs), although this does not equate to wider engagement within the Ministry as a whole.

Assumption – The Ministry has a number of questions it needs answers to (ideally situated in a wider research/evaluation policy)

The Ministry in Uganda, according to informants, lacked a coordinated approach to research, with individual programmes developing their own M&E priorities. The 3DE programme was therefore unable to engage with 'the Ministry agenda' as a whole, and there was no initial list of priority questions.

In Zambia, a list of prioritised questions existed but they required considerable manipulation. In addition, it is not clear how up to date they were, as they had been drawn up in 2011, may not have been comprehensive, and research priorities can change quickly. Moreover these were produced with support from CPs.

Assumption – There is agreement on the prioritised list of questions

There was no evidence from interviews that agreement was reached with the MoH on a prioritised question list – rather, an iterative process occurred by which individual questions were adopted, adapted and discarded or developed over time by the 3DE team, in consultation with individuals within the MoH. This assumption may not, however, be a critical one, as a rolling process may be as effective, if not more so, as a static list agreed at one point in time.

Assumptions – Good relations are established with appropriate sections within MoH; appropriate officials or units are identified and are part of question sourcing and prioritisation

Through its previous work and engagement with the MoH in Zambia, CHAI had already formed strong working relationships with a number of departments prior to the start of 3DE, although the Applied Analytics Team (AAT) leading 3DE was less established in its local working relationships. Given the split into MoH/MCDMCH, new relationships with teams in the latter had to be forged. CHAI was largely successful in doing this, but whether all sections with relevant interests were reached is harder to say. Planning officials were not always aware of the 3DE programme, and it is clear that links were focused on programme-level staff. With the exception of the Mama Kits, the evaluation questions aligned closely with areas of traditional CHAI expertise, and it is possible that sub-sectors with important evaluation questions were neglected. Given the limited number to be conducted, this pragmatic approach may well have been justified.

In Uganda, strong relationships were forged but with specific units (the National Malaria Control Programme and the AIDS Control Programme).

3.3 Evaluation design, conduct and reporting

3.3.1 The ToC

The key assumptions laid out in the ToC for evaluation design and conduct were that a rigorous evaluation must be designed appropriately, conducted and analysed with close adherence to protocols and guidelines that ensure its quality, but at the same time done rapidly to meet the demands of the Ministry. Additionally, the ethical review board must approve the design protocol in a timely manner, and programme implementers must be willing to roll out the programme in accordance with the evaluation design.

3.3.2 What happened

In this section we focus on the findings from our assessment of the quality of the 3DE evaluations. The other activities under this component are discussed when reviewing assumptions in section 3.3.2. This section is organised into three sub-headings to capture different dimensions of evaluation quality.

1. Given the problem that the evaluation was intended to address, were the research questions appropriate?
2. Given the research question, was the design of the evaluations robust?
3. Given their design, were the evaluations carried out well and how good was the quality of reporting?

At the time of writing, the FCD and Decongestion evaluations were still underway, so we can only comment on the planning and design phases. A full breakdown of our assessment of each evaluation is given in Annex E.

Appropriateness of questions

The research questions posed by the evaluations are all shown to address relevant healthcare challenges in Zambia and Uganda, but the rationales for which particular interventions were evaluated are not always fully justified. A good description of the depth of the problem faced in each case and the need for further research is given in the introductory sections of the Technical Reports and Study Protocols.

The rationale for the particular interventions evaluated in each study is not always so well developed. To fully justify the research questions it is important to provide an indication of why the interventions under test are thought to be sensible strategies to address the problem outlined.

Although intervention logic is outlined in the concept notes accompanying each study, these sections are generally brief and the majority of evaluations did not provide a ToC (although we recognise that this was not necessarily the responsibility of 3DE to articulate). Even in the absence of an explicit ToC, it would still be useful for the evaluation documents to explore in more detail the expected constraints facing the populations in each area that have contributed to healthcare challenges, and how the interventions are expected to alleviate those constraints. Without this rationale it is not always clear that the interventions were well conceived given what is currently known about the situation in each area, and therefore worth evaluating. Further details relating to each evaluation are outlined in Annex E.

Robustness of evaluation design

The overall quality of the design of the 3DE evaluations was variable, with some weaknesses stemming from the constraints placed on the evaluations in terms of timeframes and budgets.

One aspect of the evaluation design that was consistently strong for all evaluations was the choice of primary outcome given the available study period. The evaluations all focused on measures that could be plausibly expected to change over a period of months if the intervention was effective. This implied a focus on outcomes that were directly targeted by the interventions (such as uptake of a service), rather than final welfare outcomes. Many high-level health and welfare outcomes are slow to emerge and would not have been suitable for the rapid 3DE evaluation model. Although all evaluations did make an appropriate choice of primary outcome, however, there were some issues with the indicators used to track these outcomes. These are discussed in relation to individual evaluations in Annex E.

The evaluations were also well designed to make efficient use of the available budget. Since primary data collection is often the biggest cost driver in an evaluation, it was a sensible choice to use data from secondary sources where possible. Where primary data was used it had a clear role to play in collecting information on an outcome indicator not provided by administrative sources or in providing qualitative information. Fieldwork processes were also aligned with current practices in health facilities to a large degree so as to minimise the additional workload for evaluation and health facility staff.

The overriding concern with the design of 3DE evaluations is that the findings were not easily generalisable to other contexts and thus have low levels of external validity. For 3DE evaluations to have strong external validity, the findings of the evaluation should allow policymakers to be confident about:

i.) Whether the interventions would make a positive contribution to key outcomes in other areas across the potential scale-up region.
ii.) Whether and how interventions need to be adapted in order to be most effective in different contexts.

Many of the evaluations were only able to cover a limited geographic area and small sample, making the findings highly specific to that region. This makes it difficult to understand whether similar results would arise if the interventions were scaled up to other areas. This weakness is acknowledged in several of the technical reports. In view of their limited scope, in order for the quantitative findings to usefully inform a policy decision affecting a wider population, they would need to be supplemented with a thorough understanding of the reasons why the observed results arose, and how the study area and sample compare with others in the potential scale-up region. This is necessary to be able to assess the likely effects the interventions in other areas, and if any adjustments to intervention might be necessary to best suit different contexts. Although many of the evaluations did include a qualitative component to explore some of these questions, this was often not sufficiently in-depth or integrated with the quantitative results to reduce external validity concerns.

It is also not clear that some of the evaluations achieved strong internal validity despite their randomised design. The studies would have been internally valid if the intervention and non-intervention groups were sufficiently similar at the start of the evaluation period to be confident that differences in outcomes observed at the end were caused by the effects of the intervention alone. Randomly allocating the intervention helps to ensure this, but may not be sufficient if the number of units of randomisation is small. Less than 30 units in each treatment group is generally considered relatively small for an RCT[8] and for 3DE evaluations the number of units was 20 for EID, 15 for Mama Kits, eight for Decongestion and 23 for FCD. Some evaluations did take additional measures to try and preserve internal validity, such as purposefully selecting a randomisation scheme that delivered balanced groups at baseline along certain dimensions (EID) or pair-matching intervention and non-intervention units (Decongestion). This is helpful to improve confidence in the findings. However, although it is never possible to guarantee that different treatment groups are balanced according to all relevant characteristics (including those that cannot be observed or measured), the risk that these differences arise by chance is larger when the number of units is small. Consequently, there remains a concern that some of the 3DE evaluations may not have delivered sufficient internal validity (particularly the Decongestion study) despite their randomised design.

The constraints to time and budget also affected the implementation of the interventions themselves, which in some cases may have been too ambitious for a short evaluation period, in our view. The FCD, EID and Decongestion evaluations all assessed interventions that involved some reform to health systems in terms of what services were provided and how patients were processed in facilities. This is a relatively more substantial kind of intervention than the delivery of a service that is otherwise separate from the day-to-day running of health facilities. It is likely that some time would be required for these health systems changes to become consolidated and for routine operations to be established. A short evaluation period may not accurately reflect the potential effectiveness of the intervention if the assessment occurs while this adjustment period is still underway. We suspect that this is part of the reason why the EID evaluation did not reveal expected improvements in all outcomes.

In view of some of the concerns outlined above it is not clear that the choice of an RCT always made the best use of the available budget. In some cases, a simple operational pilot or process study may have provided sufficient evidence around the implementation of interventions to help guide future programming decisions. This is particularly the case for interventions that sought to reinforce existing practices rather than providing new and previously untested solutions (such as the EID Simple Intervention and the Decongestion intervention). Further details of the design of individual evaluations are given in Annex E.

Conduct and reporting of evaluations

In this section we consider the quality of the delivery of the evaluations, reporting and interpretation of results. We will not comment on the conduct and reporting of the FCD and Decongestion evaluations since these remained in progress at the time of writing.

Aspects of the quality of the delivery of the evaluation include whether high-quality data were collected following appropriate ethical protocols and whether a sufficiently large sample was included to detect a statistically significant policy-relevant effect size.

Based on the information available, the evaluations appear to have collected good data using appropriate techniques. In recognition of some of the shortcomings of relying on administrative sources a range of validity checks were included to confirm quality and address inconsistencies. In some cases we did not have enough information to assess the quality of data-collection processes and fieldwork. The technical reports did not always document whether

[8] See Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. "Bootstrap-based improvements for inference with clustered errors." *Review of Economics and Statistics* 90.3 (2008): 414–427.

evaluation staff received training in data gathering, whether a pilot was conducted, and how data were physically collected where they were drawn from administrative sources. Where this information is reported it suggests that processes were good.

Sample sizes were an issue for some of the evaluations. It seems that the sample size for the ITN evaluation was smaller than was intended. This also seems to be the case for Mama Kits, but the presentation of sample size calculations is unclear so it is difficult to determine what the intended sample size was.

Quantitative findings were for the most part presented well. However, there were also some important weaknesses in this respect. The Mama Kits and ITN evaluations did not provide enough information to confirm that the randomisation groups had statistically similar characteristics before the intervention roll-out. This is important to demonstrate that the observed changes in outcomes at the end of the evaluation period were due to the intervention alone. The ITN study also did not report any details or findings from the 'recently evaluated door-to-door intervention', which is necessary given that its main conclusions are based on a comparison between this and the community fixed point strategy. Finally, the definition of some key outcome variables was unclear in the Mama Kits evaluation, which made it difficult to understand exactly what analysis was performed.

There were some ways in which the description and interpretation of findings could have been improved. The closing sections of technical reports sometimes gave a slightly misleading impression of what the evidence from the evaluation actually showed. For example, the Mama Kits evaluation discussion suggests that the evidence provided shows that the use of Mama Kits can increase facility delivery rates across Africa. This is an overstatement of the wider conclusions that can be reasonably drawn from the study since no justification is given to support the claim that the evaluation results are indeed generalisable to other rural African settings. There are similar examples in the ITN and EID evaluations where the closing statements and recommendations of the technical report do not fully reflect the quantitative findings.

The explanation and interpretation of results could also be further developed. This problem partly relates to some specific individual results that are not well explained, such as the finding in the EID evaluation that the interventions did not lead to a statistically significant increase in infant Dried Blood Spot (DBS) tests. The repeated stock-outs of DBS test kits would seem to be a highly plausible explanation for this result, but this is not given as a possible reason. Yet a more general issue in this regard is that the overall findings were not always situated within a broader discussion of the context and likely mechanisms. As previously noted, given the small geographical areas covered by the 3DE evaluations a deep understanding of how and why the results arose in that particular area is necessary to assess whether similar outcomes might be expected if the intervention were scaled up elsewhere. This requires outlining how the study area compares with potential scale-up areas in the country, which constraints to the intended outcome are likely to be prevalent in that area, and how the intervention under test addressed them given these particular conditions. Further details in relation to each completed evaluation are provided in Annex E.

A summary of the quality assessment is provided in Table 3 below. Green indicates that it could not have been better, amber that some elements are missing or amiss, while red indicates serious limitations. Annex E provides a more detailed assessment of each evaluation.

Table 3 Summary of quality assessment for each study

Category	Proposed questions	Mama Kits	ITN	EID	Decongestion	FCD
Planning and context	1.1 How relevant are the evaluation questions to the priority questions of the Ministry? (explored as part of validation of the ToC)					
Introduction	2.1 Is the evaluation question(s) written simply and clearly?					
	2.2 Are the evaluation questions suitable given the short duration of the evaluation period?					
	2.3 Is there an adequate description of the intervention to be evaluated (this should include detail on the intervention's target groups, timescale, geographical coverage, anticipated impact, outcomes and outputs, intervention logic and/or ToC)?					
	2.4 Is there a discussion of other programmes or interventions that may also affect impact, outcome and output indicators?					
Method	3.1 Is an RCT the most appropriate method to answer the evaluation question?					
	3.2 Is the unit of randomisation appropriate?					
	3.3 Did the randomisation produce treatment and control groups that were similar at baseline?	Not enough information to assess	Not enough information to assess		N/A	N/A
	3.4 Are issues related to spill-over effects/externalities (e.g. untreated individuals are affected by the treatment) considered and dealt with appropriately?					
	3.5 Are issues related to imperfect compliance (i.e. people in treatment group not being treated, or people in control group being treated) considered and dealt with appropriately?					
	3.6 Are local and national contextual factors that could affect the evaluation considered?					
	3.7 Is the timing of the data collection appropriate given the timing of the intervention?					
	3.8 Can the findings be expected to have reasonable external validity to inform a wider policy or programmatic decision?					
	3.9 Were there any trade-offs in design due to the relatively short timeframe of the evaluation, and if so what were they?					
	3.10 Are there other significant methodological limitations (not mentioned above)?					

Category	Proposed questions	Mama Kits	ITN	EID	Decongestion	FCD
Data	4.1 Were the most suitable data sources selected? If primary data collection was undertaken, were the most suitable data-collection methods selected?					
	4.2 Have the sampling frame and the sampling populations been correctly defined?					
	4.3 Is the sampling procedure rigorous and appropriate? (What is the sample representative of?)					
	4.4 If primary data collection was undertaken, are survey instruments well constructed (clear, robust skip patterns, relevant answer codes) and are they adequately described?				Not enough information to assess	
	4.5 Are secondary data sources adequately described and has their quality been checked to determine the data is reliable?				Not enough information to assess	
	4.6 Were sample sizes adequate?				N/A	N/A
	4.7 Were sample size calculations done well and are they presented?					
	4.8 If primary data collection was undertaken, are any biases from non-response discussed?				N/A	N/A
Data collection	5.1 If primary data collection was undertaken, were data collected in an appropriate and respectful manner, taking into account cultural and ethical dimensions, as determined from the protocols submitted for ethical approval, the field manual and the characteristics of the data collectors?					
	5.2 If primary data collection was undertaken, were the instruments tested and validated (e.g. pre-testing of questionnaires)?					
	5.3 If primary data collection was undertaken, were the instruments translated and back translated?				Not enough information to assess	Not enough information to assess
	5.4 Were field teams trained to gather data before the start of the intervention? If primary data collection was undertaken, were the field teams trained by the same people who made and tested the survey instruments?					
	5.5 Has there been an appropriate level of oversight and data quality assurance in the data collection?					
Data entry and cleaning	6.1 If a survey was undertaken on paper, was the data double entered and were discrepancies between the two entries systematically resolved by checking the hard copies?			N/A		N/A

Category	Proposed questions	Mama Kits	ITN	EID	Decongestion	FCD
	6.2 Was the data cleaning done is a robust, clear and transparent way and does it include both range and consistency checks?	Not enough information to assess	Not enough information to assess	Not enough information to assess	N/A	
Data analysis	7.1 Are primary analysis methods appropriate? If regressions are used, are they correctly specified and are standard errors calculated correctly?					
	7.2 Are the key indicators clearly defined (including how they are calculated), and are they suitable to measure the outcomes of interest?					
	7.3 Have sampling weights been used correctly?	N/A	N/A			Not enough information to assess
	7.4 Are departures from the full sample size (non-response) explained and has any non-random attrition been identified and dealt with correctly?			Not enough information to assess	N/A	N/A
	7.5 Have any differences between treatment and control groups at baseline been accounted for in measures of impact?				N/A	N/A
	7.6 Is the analysis disaggregated to show outcomes and impact on different groups and sexes? Did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?				N/A	N/A
Reporting	8.1 Are quantitative results presented systematically and logically?				N/A	N/A
	8.2 How clear are the links between data, interpretation and conclusions?				N/A	N/A
	8.3 Were negative/discrepant results addressed or ignored?				N/A	N/A
	8.4 Are the final recommendations and conclusions plausible?				N/A	N/A
	8.5 Have alternative explanations been explored and discounted?				N/A	N/A

3.3.3 Review of the assumptions

The first assumption related to the quality and rigour of the evaluations is discussed extensively above.

Assumptions – Appropriate partners identified for implementation of the evaluation; funding and implementing partners are willing to implement the programme in accordance with the evaluation design and protocol

The choice of question was informed in part by the availability of partners rolling out an intervention that could be used to generate relevant impact evaluations, rather than starting from questions and then seeking appropriate partners. The issue of suitable partners for the intervention appeared to be more of a constraint in Uganda, where one malaria question was derailed by disagreements over design and implementation and, ultimately, 3DE developed and rolled out its own intervention.

In Zambia, identifying implementation partners was not highlighted as problematic by any informants or documents. This may reflect the wider casting of the net in terms of programme areas in Zambia. In some cases, such as the Mama Kits and EID evaluations, there was a choice of implementers and they were prioritised according to which areas they were working in and willingness to collaborate, according to programme documents. Where implementers did not have a budget for evaluation, it was a win/win situation for them.

The implementation of the evaluation, as opposed to the intervention, was done with the MoH/MCDMCH in all cases; however, the degree of involvement of the MoH PIs and co-PIs varied according to individual time and interest, as would be expected. For the Uganda FCD study, the PI was clearly very involved and played a leading role. This was less evident in the Zambia studies, where the evaluation activities were led by 3DE but with intermittent feedback and communication with the MoH partners on design and results in particular.

Assumption – Evaluation goes through ethical board and is approved in a timely manner

In Zambia, ethical approval was sought from a private Institutional Review Board (IRB) (ERES Converge) and the MoH prior to commencing data collection. For the EID evaluation, the Zambia Centre for Applied Health Research and Development partner had to submit to Boston University IRB as well. In Uganda, approval came from the Mildmay Uganda Research and Ethics Committee, as well as the Uganda National Council for Science and Technology and district-level officials. Despite these multiple hurdles, ethical review boards were not a major cause of delay, judging from the evaluation timelines (approvals were completed within a month). Either the risk of delays was not realised or was well managed. However, for some studies (i.e. the Decongestion and FCD studies), protocols were resubmitted more than once. This presumably reflects design changes from the 3DE side, rather than changes enforced by ethical review boards.

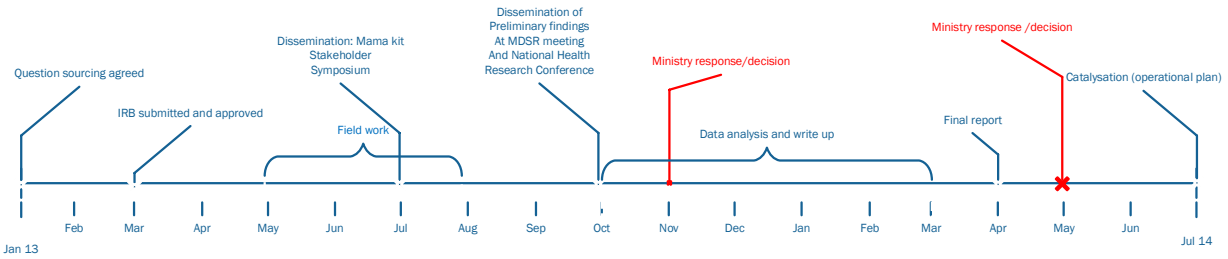
Assumption – The evaluation is conducted rapidly to meet the agreed timelines of the Ministry

Although the 3DE concept hinges on rapidity, it was not always the case that there was a pressing policy ‘window’ that had to be taken advantage of. For the Mama Kits, the evaluation used a UNICEF roll-out opportunistically and had to move quickly before the already procured kits were distributed to targeted districts, but there was no immediate need for information from the MoH perspective – this was a longer-term question that needed to be better understood. The same applies to the EID, Decongestion and FCD studies. The only one where study results were urgently required was the ITN study, due to a massive bed-net distribution planned for spring of 2014. Thus,

it should be noted that timeliness is always important but rapidity may not be a priority in all settings, and restricts the types of questions which can be answered.

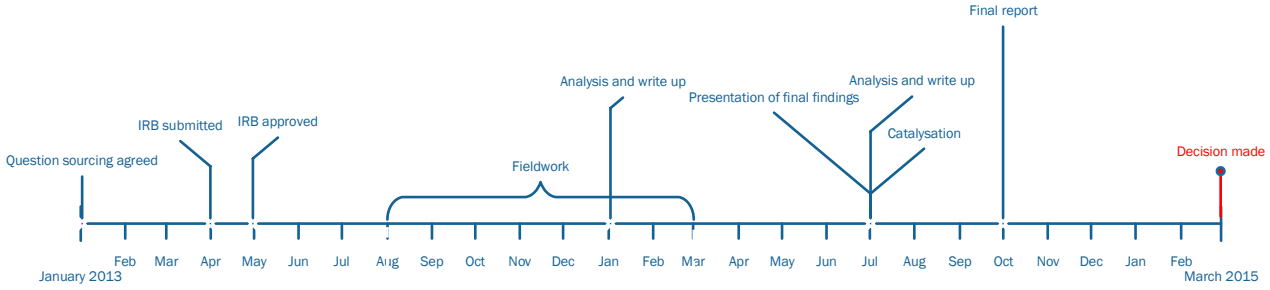
So, although rapid results were part of the core concept for 3DE, with the exception of the ITN study the other evaluations were not rapid in the sense of being completed in six to nine months as planned (see Figure 5 to Figure 9). Leaving aside the sourcing stage, from agreement on a research questions to producing the final report took two years for the FCD evaluation, 18 months for EID, 15 months for the Mama Kits, one year for Decongestion, and nine months for ITN. The consequence of the intensity of the sourcing process added to these study durations meant that catalysation has been squeezed, especially for the later studies.

Figure 5 Timeline for Mama Kits evaluation



Time from question sourcing to preliminary findings (9 months)
Time from question sourcing to final report (15 months)

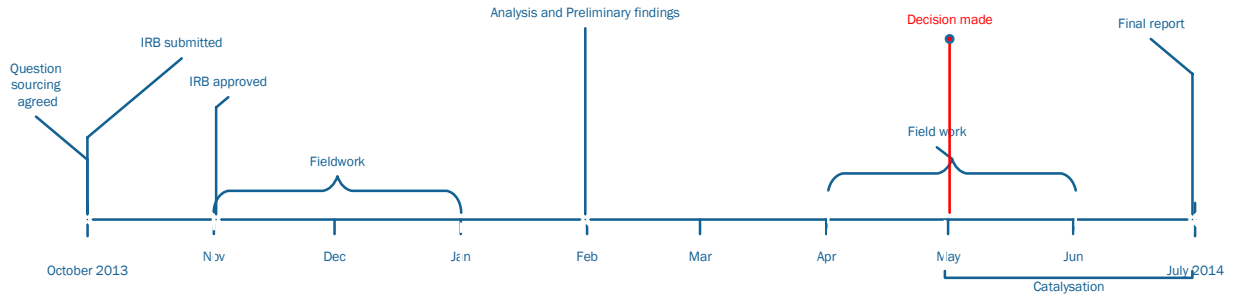
Figure 6 Timeline for EID evaluation



Time from question sourced and agreed to presentation of findings (18 months)

Time from question sourced and agreed to final report (21 months)

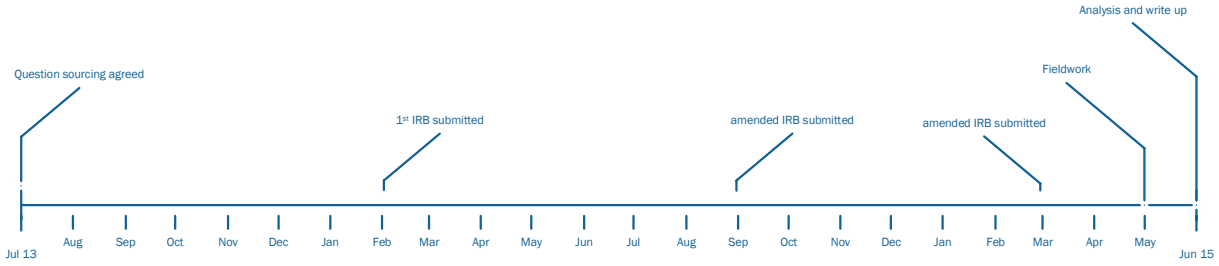
Figure 7 Timeline for ITN evaluation



Time from question sourced and agreed to presentation of findings (4 months)

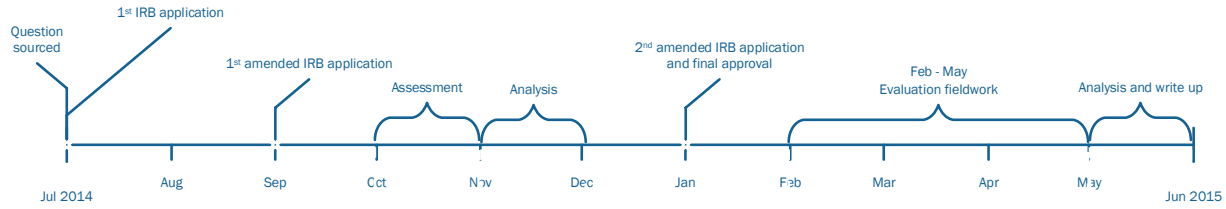
Time from question sourced and agreed to final report (9 months)

Figure 8 Timeline for FCD evaluation



Time from question sourcing to fieldwork (24 months)

Figure 9 Timeline for Decongestion evaluation



Time from question sourced and agreed to analysis and write up (11 months)

3.4 Dissemination and activities to catalyse implementation

3.4.1 The ToC

According to the ToC, communication of the evaluation findings to key stakeholders requires there to be demand and interest in learning about the outcomes of the evaluation, for relevant stakeholders to be identified and informed about the existence of the evaluation, and that the findings are presented in an accessible manner and suitably tailored to different audiences. For the supporting catalysation activities to be appropriately undertaken it is assumed that there are staff within the Ministry who are able to technically engage with the evaluations, for the Ministry to accept the findings of the evaluations and request support for their implementation or accept proposed support. Moreover, these activities are required to be undertaken on time and repeatedly over time (i.e. brokering for a period of time and through different mechanisms) in accordance with the needs and requirements of the Ministry by 3DE staff with the right skill sets and with the engagement of the right stakeholders.

For wider international dissemination, the production of briefs and publications and dissemination in forums is contingent on the need for non-technical briefs and appropriate journals to be identified, for the papers to be submitted and accepted by the journals and for a suitable dissemination forum to be selected and invitations given.

3.4.2 What happened

The evidence on how the three studies that have reached completion were presented, discussed, disseminated, and catalysed are provided in Box 1 to Box 3 below, alongside details of the

documentation of decisions taken and any follow-up actions the evaluation team were able to uncover.

Box 1 Mama Kits evidence uptake

Presentation of findings
Prior to completion of the fieldwork, 3DE raised awareness of the Mama Kits evaluation at a stakeholders meeting on the use of non-monetary incentives to increase facility delivery in Zambia. Subsequently, the preliminary findings were presented at the Maternal Deaths Surveillance and Response (MDSR) Meeting held on 28 October and earlier that month as a poster presentation at the National Health Research Conference. The final report has not been disseminated and was only presented to the three co-investigators of the study at the Ministry in April 2014.

Government response/decisions
A letter sent by the MCDMCH to CHAI on 13 November 2013 (prior to completion of final report) acknowledges the evidence provided and further notes that they have made a decision to 'act upon the findings of the study' and to 'work closely with 3DE team to ensure that, among other things, support is rendered to districts to enable them to be innovative so as to support the inclusion of Mama Kits within their activities'.

Discussions were also held with MCDMCH officials in April 2014, where the sufficiency of the evidence to support scale-up was acknowledged. The officials were of the view that scale-up of Mama Kits should include kits with contents that contain items both for the mother–baby pair but also contain safe delivery items such as surgical gloves, sterilising liquid and chord clamps. Moreover, in the short term they expected the scale-up to require leverage of resources from external partners. The inclusion of Mama Kits as part of the essential commodities provided to health facilities was seen as a medium-term measure.

In May 2014 the MCDMCH circulated a letter to CPs to underline its support for using the kits as a demand creation tool that would encourage facility deliveries. Moreover, the letter encouraged the partners to also use the low-cost Mama Kits to incentivise visits to facilities.

Catalysation
A draft operational plan was developed by 3DE staff in July 2014. This operational plan has not been adopted by the Ministry.

What has happened since?
The EU-funded and UNICEF-managed Millennium Development Goal Initiative has considered the findings of the report and has procured kits of a similar nature but of a higher quality, costing $11. This programme distributed 34,000 kits in 2014 and 20,000 in 2015 in 11 districts across the two provinces of Lusaka and Copperbelt. Some of the targeted districts are peri-urban, where provision of non-monetary incentives may be less effective in increasing institutional delivery. The programme will not distribute any more kits in 2016 or beyond.

The government has so far not committed any budget to purchase Mama Kits and the districts are unlikely to have purchased any kits given their lack of resources. Moreover, some implementing partners have made a decision not to continue with Mama Kits and to focus on other interventions believed – in their view – to be more effective in increasing facility delivery. There is no evidence to suggest other partners have procured any low-cost Mama Kits in accordance with the Ministry's letter.

Overall, the team concluded that:

3DE generally has a good awareness of entry points and key stakeholders and disseminated to key stakeholders for influencing decisions, largely at the programme level. In general, preliminary findings were presented to the TWGs, which are the main forum for sharing technical and operational information relating to specific programme areas in Zambia. The findings from the Mama Kits evaluation were not presented in any TWGs (because no relevant TWG exists), but the final findings for the Mama Kits evaluation were presented at a meeting of stakeholders for the Saving Mothers, Giving Life Phase II scale-up launch in May 2014. The ITN evaluation findings were presented at 3 provincial micro-planning meetings. 3DE also held numerous one-on-one meetings

with stakeholders to share results as needed. MoH/MCDMCH officials were given lead roles in presenting findings at various dissemination events.

Feedback preceded finalisation, which has risks. In all three cases, preliminary results were shared well ahead of final reporting, presumably in order to provide rapid inputs into decision-making, and Ministry 'decisions' based on the evidence were taken ahead of final reports in the case of the ITN and the Mama Kits evaluations. While the 3DE programme rightly focused on timeliness, as mentioned above, only for the ITN study was there an absolute urgency. Sharing results quickly raises some concern about time for data checking and appropriate interpretation of findings, especially when they are feeding into policy decisions. For example:

- the presentation of findings on the EID study suggested not only that there was no negative effect on immunisation but also a positive effect on maternal testing – a finding which was not confirmed in the final report.
- For the ITN evaluation, the provincial presentation of March 2014 does not mention the option of having no CHW hang-up visit as part of the community fixed point strategy. This may be why CHW hang-up visits were built into the strategy that was later recommended to districts despite the technical report indicating that having no hang-up visit does not lead to any difference in ITN use or retention after five to six months. Estimates of time savings also varied between the March presentation and July report.
- For the Mama Kits, the presentation of October 2013 shows a 60% increase in facility deliveries, but this is not reported in the technical report of April 2014. The presentation also reports baseline characteristics with the variable '% districts with Mother and Child Health [MCH] activity in past year'. This is 93% in comparison facilities and 60% in treatment facilities, so clearly there was a lack of sample balance along this variable. It is left out of the table in the technical report.

Also, it is not clear how well disseminated or read the final reports were.

'Decisions' were advisory. The decisions which were documented and which are key performance targets for 3DE were limited in scope and largely in the form of advisory notes (Mama Kits, EID, ITN) or reinforced existing practice (EID).

Catalysation has been limited and not always appropriate. The main catalysation, as originally conceived (operational plans, costing, helping the MoH to roll out interventions), was recorded for all three evaluations. It is not clear whether these were always appropriate given the nature of the evaluations, which do not imply policy changes or rolling out new interventions. The Mama Kits are distributed by donors, which is why an operational plan for the MoH is of questionable applicability. In the case of the EID evaluation, this was already national policy and it is not clear what the benefit of developing a post-evaluation implementation plan was or by whom it was taken up. The main catalysation activity for the ITN evaluation was the sensitisation of the provincial and district staff of the point distribution process during micro-planning events in February and March 2014. While an operational plan for implementation of the point distribution was developed in July 2014, it is not clear whether this was utilised given that the distribution had already started in most areas.

Uptake to date appears to be limited. 3DE was not resourced to follow up after 'decisions' had been made on the evidence they generated. However, informal follow-up of the ITN study (see Box 3) found that around 30% of districts had adopted point distribution in some fashion but that awareness of the new position was low. For the Mama Kits, one donor decided to use them, but adopted a more expensive version (costing $11 per kit) and in largely urban/peri-urban areas (11 districts in Lusaka and Copperbelt Province for 2014 and 2015 only). As the Mama Kits results were generated in rural areas and focused on reducing the cost of previous kits, the influence of the study would seem to be limited so far.

Box 2 EID evidence uptake

Presentation of findings
Prior to commencement of evaluation field work, the design of the evaluation and its timeline was presented at the National Paediatric ART review meeting in April 2013. Subsequently, the findings were presented by one of the co-investigators from the Ministry in Australia at the 6th HIV Paediatric Workshop in July 2014. The final findings were presented to the co-investigators and discussed in October 2014. The final reports were also presented at the Expanded Programme of Immunisation (EPI) TWG as well as the EID in March 2014.

Government response / decision
An official letter from the Permanent Secretary was circulated in March 2014 to Provincial Medical Officers (PMOs), the district community medical officers and HIV and ART implementation partners and key stakeholders as well as those under immunisation. It drew their attention to existing guidelines related to HIV treatment, prevention and immunisation, and reminded them to ensure that identification of HIV positive infant and children occurs in under-five clinics in facilities within districts and is conducted with adherence to standard HIV testing and counselling best practice. Moreover, they were instructed to ensure that monitoring of this service provision was included as part of routine monitoring activities such as performance assessment activities in facilities.

Catalysation
A draft EID–EPI post-implementation plan was produced in July 2014 by the 3DE programme.

What has happened since?
The main discussions took place at the EPI TWG meeting in March 2015, at which it was agreed that a Job Aide would be produced for health workers at under-five clinics to operationalise the guidelines for integration of EID and EPI and guide health workers at under-five clinics in providing HIV services to mothers and infants. At the time of this evaluation this activity had not taken place. There are a number of activities and initiatives taking place to tackle the system-wide barriers to ensuring existing guidelines are implemented and EPI/EDI services are integrated. CHAI through its core programmatic focus is supporting this process, for example through development of a module on EID/EPI integrated into the Paediatric HIV/PMTM training, the secondment of a paediatric HIV coordinator, and lobbying for an integrated commodities supply system.

3.4.3 Reviewing the assumptions

Assumptions for dissemination

Interviews and documentary evidence suggest that there was demand and interest in learning about the outcomes of the evaluation by the main co-investigators at least and that 3DE made efforts to present findings to wider stakeholders. The stakeholders, as highlighted, were programme staff at the MoH and partner agencies, and 3DE developed stakeholder mapping tools to identify them and develop engagement strategies to work with them. Given the focus of the evaluations, which was on operational issues, this is judged to be the right audience (see Box 4). The findings were presented to the right stakeholders, but with the exception of the EID evaluation the *final* findings of the completed evaluations were not presented to wider stakeholders and were only presented to the co-investigators.

Presenting in an accessible manner was done largely through the PowerPoint presentations made based on preliminary findings. These products were very important as they were the main vehicle through which stakeholders were informed of the results. Some of the results tables are unclear, even to the evaluation team. For example, the presentation of the preliminary results for the Mama Kits at the MDSR Meeting highlights its cost-effectiveness and in one slide provides a comparison of two logistic regressions that is meant to show the initial analysis of impact of the Mama Kits intervention. It is not an easy slide to interpret. However, the presentation was the basis of the initial response by the government to encourage districts to consider ways to incorporate Mama Kits into their activities.

Some posters for conferences were produced, along with a manual and short video clip on the 3DE approach, but there is no further evidence of tailored products for domestic audiences (e.g. policy briefs on the findings, with the exception of the Mama kits evaluation, although briefs were produced earlier as part of the sourcing process). These might have been useful, given the presentations were done prior to final reports being produced, and few decision-makers are likely to have read the final technical reports. Interviews suggest that even those closely involved in the evaluations may not have fully understood the findings:

> *In a nutshell without doing much there was an increase in the number of tests that were conducted and immunisation coverage was not affected. There was no negative impact but rather improved testing of HIV testing for children*. (KII, Zambia)

Box 3 ITN evidence uptake

Presentation of findings
The evaluation was introduced in November 2013 to the ITN TWG, which contains key stakeholders involved in procurement and mass distribution of the nets, to garner their interest and support. Evaluation findings were then presented on 25 February 2014 to the ITN TWG, where they agreed to provide this as an option to the provinces and districts as they were best placed to decide on the best distribution strategy given consideration of local conditions. The preliminary findings of the evaluation and use of point distribution were discussed in two provincial micro-planning meetings in February and March 2014.. The point distribution strategy was presented as an 'alternative distribution strategy districts could consider in planning for and conducting the mass distribution.' In the meeting it was also acknowledged that the decision had not formally been taken. Moreover, the importance of having a letter from the Permanent Secretary (PS) of the MoH was highlighted in provinces considering point distribution.

Government response / decision
A letter dated 5 May 2014 by the PS to PMOs' states encouraged them *'to explore whether the CFPD strategy can complement the door-to-door distribution strategy in your specific geographical setting'*. They were referred to the ITN distribution micro-planning guidelines with an addendum on the CFPD net distribution strategy (which was attached). They were also advised to inform and work closely with the National Malaria Control Programme before implementing the community fixed point distribution.

Catalysation
A draft operational distribution plan for community fixed points was produced in July 2014.

What has happened since?
Most micro-planning was conducted during March/April and before the official letter from the PS had been signed. According to a post-evaluation implementation report, only 17 districts (out of 57 districts that responded) reported using point distribution for some or all their areas. Moreover, of the 48 districts that knew about the point distribution only 17% had complete knowledge of the programme.
The National Malaria Control Centre has not decided whether it will consider point distribution for the mass distribution or not.

Assumptions for catalysation

The assumption that results have clear policy implications was not found to hold in practice, although the answer depends to some extent on how the term 'policy' is understood. The evaluations focused on operational questions, and to that extent were not aimed at changing policy, if that is understood as a "statement of goals, objectives and courses of action outlined by the government to guide its intended actions".[9] In addition, the operational implications of the evaluation findings were not always clear, in line with the questions over external validity raised above, and which were raised by stakeholders at the time (noted in minutes) and in interviews with the evaluation team.

[9] Government of Zambia (2010), in Zambian Governance Foundation (2012), 'The Policy Formulation Framework in Zambia', September.

In terms of staff who are able to technically engage with the evaluations, this was generally the case but the MoH partners faced time and capacity limitations in Zambia, with very few staff dedicated to research in the MoH (only two, who play more of a coordination role), while the MCDMCH has no research team at present.

Ownership by the MoH/MCDMCH of the evaluation findings was mixed. There was a clear sense of having been involved by 3DE in the process, but that did not mean that comprehension or acceptance of findings was complete:

> *We discussed this at the dissemination. We cannot say we cannot generalise these results because we really don't know what the situation would be, for example in Copper Belt. There is some bias in the sense that this is an area we already had EID which was very strongly done and they were using sites where there was a lot of mentoring and mentorship taking place. We do not know what would happen in a place which is really naïve – we don't know if that would affect it and we don't the situation of the PMTCT [Prevention of Mother to Child Transmission] programme elsewhere and how that would affect the results. So we cannot guarantee that the results can be generalised to the Zambian population.* (KII Zambia)

Moreover, districts play an important part in programme implementation and the ability of the 3DE programme to link with them was limited by funding and lack of presence:

> *The Ministry is assumed to have capacity but may need help for longer to carry recommendations to district level. CHAI has had limited time and budget to follow up on findings, which is frustrating. It also has no RMNCH [Reproductive, Maternal, Newborn and Child Health] programme and so is less well placed to catalyse findings. […] And at district level, how is it supposed to work? If they recommend to districts, do they have the support or information to act? There is no facility in CHAI to do that. (KII, Zambia)*

Given the lack of major policy change implications of the evaluations, the assumptions that the Ministry requires and requests support or accepts proposed support and that the support is provided on time and according to their needs were not directly applicable. There is every evidence that the 3DE had technical skills in providing catalysation support, knew the stakeholders well and was able to engage well with the MoH and partners – the work done by the wider CHAI programme in Zambia is evidence of this. However, these skills could have been supplied by CHAI but not requested by the Ministry.

Box 4 Overview of TWGs in health sector in Zambia

There are numerous TWGs within the MoH and MCDMCH, including TWGs for child health, safe motherhood, malaria, HIV, TB and other smaller TWGs within them. The TWGs are a sub-set of the Sector Advisory Group Meetings (SAGs), joint high-level consultative forums through which sector-level dialogue, alignment, harmonisation and managing for results are expected to take place. The MoH SAG brings together the MoH and all its partners, including all the relevant government ministries[10] and departments, the private sector, civil society[11] and CPs.

As noted above, the TWGs are thematic and are a forum for coordination of activities between government, donors and implementing partners, the sharing of information and discussion of issues arising from implementing of programmes and initiatives related to that thematic area.

Government officials (usually directors or principal officers) chair these meetings and set the agenda. Members are invited at the behest of the government or as requested by an organisation. While the agenda is set by the government, some of the larger representatives (in terms of funding and levels of activity) also have considerable input. These meetings are largely attended by

[10] These include the MCDMCH, Ministry of Finance, Ministry of Justice, Ministry of Agriculture, Ministry of Education and Ministry of Local Government and Housing.
[11] The Churches Health Association of Zambia and health sector civil society organisations.

implementation partners, civil society, some private sector actors and CPs, with a very small number of government officials present.

The TWGs are focused on implementation of existing policies and mainly operational in nature. Evidence and findings are discussed and presented in the group but these are largely driven by CPs and implementing partners, who are often the sponsors of these studies, evaluations and assessments. The consultants for these studies often present their study design, progress and findings to garner support and buy-in.

TWGs' feedback to the Ministry is through the chairs and any issue worth further consideration is escalated to the deputy/director level and then further above to the PS.

The TWGs are not the only mechanism where implementing partners and donors discuss operational issues. These are also raised bilaterally with directors, or even PSs depending on relationships built and the level of access implementing partners and donors may have. Sponsors of studies also arrange specific dissemination events and invite key stakeholders to them.

Source: analysis of selected TWG minutes and interview responses

3.5 Achieving outcomes

3.5.1 Theory

Consideration and use of 3DE findings by policy-makers

As highlighted in the ToC, the MoH's consideration and use of the 3DE evaluation findings in making policy decisions rely on a number of assumptions. It requires that the sourced evaluation questions be important and relevant to the Ministry, for the findings to be credible and acceptable to the main stakeholders,[12] and for there to be a willingness to act on these findings. Moreover, it requires the findings to be aligned with the policy priorities and decisions of the Ministry and for them to be provided in a timely manner and within the policy formulation timeframe of the Ministry. Additionally, the evaluation findings need to be clear, have obvious policy implications, for the policy process to be flexible in allowing the incorporation of the findings and, finally, for there to be resources available in the event of policy change or reformulation. Resource availability is also considered at the sourcing stage of the pilot.

Contextual factors needed to support this process include organisational and institutional structures that provide incentives for demand for evidence and its use and a conducive political environment. These factors will be explored as part of the PEA described below.

There are a number of proximate indicators to gauge whether the outcomes of the programme are achieved or more likely to be achieved. These include whether the outputs have been presented at the right TWGs, whether there is evidence of them being discussed, whether the 'owners' of the 3DE (e.g. PIs or co-investigators from the Ministry) are part of the policy process, whether the work is developed as part of the process, and whether new policy plans reference the 3DE evaluation outputs.

Consideration and use by CPs and other organisations

Although not explicit in the original ToC of the programme, one of the potential outcomes of the 3DE model is the use and adoption of the 3DE evaluation findings by the CPs and other

[12] If the findings are not acceptable, significant brokering activities will be needed to convince the audience otherwise.

organisations involved in the health sector in Zambia. This is especially pertinent in the case of Zambia, where significant health financing resources come from these partners.

For the CPs and other organisations to consider and adopt the findings of the 3DE evaluation it assumes they are interested in and engaged with the 3DE evaluation questions and the evaluation findings and that they are disseminated to them. Moreover, the findings are expected to be aligned to their priorities, and they are assumed to have the resources to support programmatic changes and the flexibility to respond to the findings of the evaluation. Again there are a number of proximate indicators to assess the likelihood of this output being achieved.

Enhanced evaluation capacity and interest at the Ministry

A central assumption for the 3DE model to enhance the capacity and interest of the MoH in evaluation uptake is that the 3DE model explicitly supports the development of individual and organisational skills and capabilities in commissioning evaluations, engaging in their implementation, and in understanding and interpretation of their findings through an elaborated training plan/capacity-building initiative that has been implemented by CHAI.

Beyond this a number of contextual factors influence this intended outcome, including the incentives provided by the organisational and institutional structures for demand for evidence and use by the Ministry and, related to this, the availability of appropriate positions and the recruitment of suitable cadres of staff that have strong research and evaluation skills.[13]

As shown in the ToC, indicators of change in behaviour or interest include whether there is interest from the evaluation 'owners' to present evaluation findings at the TWGs (or equivalent forums) or with CPs or other appropriate avenues, whether there is demand for presentation of the 3DE findings and methodology in other areas of the Ministry, and whether there are organisational strategies developed for 3DE or similar evaluations to be procured.

3.5.2 What happened

Evaluation outputs were indeed presented as preliminary or final findings at the TWGs and relevant forums for the three completed evaluations. The minutes show that the 3DE outputs were discussed and were presented by MoH staff or the other PIs involved in their development. MoH staff were able to interrogate the evidence, at least in some cases, and suggest modifications (e.g. adding items to the Mama Kits to address clinical concerns). However, there was limited need for catalysation, as highlighted above, and no policy or budget development process for which the outputs had implications.

In relation to influencing development partners, there was evidence of interest in the 3DE evaluations but also some scepticism about the findings and the extent to which they should be incorporated in wider plans. None of the other 'good to see' or 'like to see' indicators were identified, although hidden evidence trails may exist:

> *We have a Reproductive Health TWG in country where different interventions to increase access and use of Maternal and Newborn Health services including Mama Kits were presented and discussed based on the literature and feasibility studies in Zambia and other countries. We were not convinced that this intervention was cost-effective or sustainable for our initiative… we discouraged our implementing partners from using Mama Kits; there were other clinical, social and community interventions that encouraged women to utilise facilities for maternity services. (correspondence with a CP)*

[13] If there is political demand and interest for evidence, it is more likely that the Ministry will have positions for generation or gathering of evidence that are filled and done so with appropriately skilled individuals. If there is little demand or interest in evidence these posts may exist in name only and be vacant or be filled with staff without the appropriate skills.

3DE did not have a specific capacity-building plan beyond working closely through the stages of the programme with MoH/MCDMCH partners and this was not an explicit requirement of the pilot as stated in the original business case. In terms of strengthening evaluation capacity and interest at individual and organisational level, interviews indicate that individuals who worked closely with 3DE did benefit in terms of capacity development – particularly for the FCD evaluation in Uganda, where the MoH team were closely engaged at all stages. More broadly, there is an expression of latent demand for evidence, though not necessarily for evaluations specifically. In Zambia, a Health Strategy Plan and Act have been developed to support coordination and setting of health research priorities, but this has not as yet been effectively implemented. The re-alignment of the two ministries (MCDMCH/MoH) has further diluted the research functions of the MoH while no research functions have been allocated to MCDMCH. In Uganda, the MoH lacks a wider strategic approach to evidence and research and there is no indication that this has changed as a result of 3DE:

> *It is hard to say if there has been wider change, but the individuals with whom CHAI has worked closely in the MoH on the FCD study have grown. There are no quick wins. You have to work one individual at a time. It is a long game. Really tough.* (KII, Uganda)

Although it was not a focus of the programme, the organisations implementing the programme have learned along the way – in CHAI's case around the requirements and techniques for doing impact evaluations, which it is now undertaking in other projects in Zambia, and in the case of IDinsight, expanding their portfolio and experience of working with different ministries and in new contexts:

> *There has been great impact from 3DE in capacity building [for CHAI]. We have always provided evidence to help shape policy, but not rigorous evidence. CHAI is now thinking about more rigorous studies to generate evidence; this has improved CHAI's capacity too. This is not traditionally how CHAI has worked.* (KII, Uganda)

3.5.3 The review of assumptions

Assumptions for achieving outcomes

In relation to assumptions regarding MoH uptake of findings, the evaluation found the evaluation questions were aligned with priority areas within the Ministry. The evaluations provided some useful evidence on operational aspects of specific interventions required as part of wider set of interventions on issues that are of priority to the government. 3DE staff were seen as professional, understanding and flexible in their approach. However, as has been mentioned above, some question marks were raised over the applicability of findings in non-study contexts, and some addressed programme areas where the MoH has limited funding or budget flexibility (e.g. for Mama Kits):

> *Feasibility and acceptability of integration were questioned [by PMOs] when in reality there were numerous HR challenges, including shortages. An additional comment was the need/importance of community-level awareness. Male involvement would also be key. Supply chain was key and a priority, because once health workers were sensitised about what to do, stock-outs would reduce their morale to provide the services, undoing any gains from training/mentorship*. (Notes from 26 February 2015 EID PMO Option B+ Update meeting)

> *The question mark is in the places where they did point distribution – are they [community volunteers] going to follow up to see if the volunteers have hung the nets? Because the follow-up in Lufonza was done for the study and was closely monitored and the question is are they going to be doing that elsewhere? For the distribution they are given incentives but what about the follow-up? Should they be given incentives? […] I can't really say what is*

> *going to happen in the next distribution in 2017 and what will be recommended.* (KII, MoH, Zambia)

The MoH and districts have some flexibility to re-plan funds within budget years, but overall resource constraints do affect evidence uptake – for example, understaffing and supply chain security were among the concerns highlighted in relation to implementing the integration that was supported by the EID evaluation.

As highlighted in the PEA, institutional and organisational incentives for evidence demand and use are not evident in Zambia given the centralisation of power and lack of resources allocated to research within the health sector, despite the rhetorical support (No equivalent study was conducted in Uganda). The low-level implications of the evaluations meant that they did not threaten any interest groups or imply any shift in resource allocation, and so the requirement for political responsiveness to new evidence was not tested. The evaluations have been operational in nature and thus operate below the political radar. Issues around financing and the availability of donor funding play an important role in influencing these managerial decisions and interventions. MoH, MCHMCH, district and donor interests all have to be aligned in ensuring uptake. Key informants were able to give some examples of decisions that were influenced by evidence, but not many.

Partners were engaged early on in the 3DE studies, and were well represented in the TWGs where findings were discussed. They have more flexibility to respond to findings where these intersect with donor programmes.

In terms of building demand and capacity for evaluations, this was not a primary focus of the 3DE programme: as mentioned above, there was no elaborated plan of training or capacity building. Although the proposal for 3DE describes a methodology of embedding staff in the MoH, working alongside MoH staff and gradually shifting responsibility to them, this was not the model followed in practice, which ultimately was focused on the supply of evidence. Such capacity as was built (largely in a few close counterparts) appears to have been achieved through the direct involvement of collaborators on the evaluation processes. The wider environment was also challenging: in Zambia, the research team comprises two individuals in the MoH. The need to build capacity (in the ministries, among researchers, and through development of brokerage functions and research networks) was highlighted in many interviews:

> *The capacity is very limited… they are trying their best. They have limited capacity because they have limited personnel. They will still need a lot of capacity building and technical support to build the research portfolio. They can't do everything. So if 3DE came to do that that is very much welcome… For knowledge management there is essentially ZAMFOHR [the Zambia Forum for Health Research), which really needs help. That was the only institution that was developed for knowledge management and translation (founded by the current Minister of Health but before he became a Minister). But it is limping because of inadequate funding.* (KII, local organisation, Zambia)

> *There is need for capacity building in the MoH at all levels – central, province and district. More could be done to help people appreciate research findings – the capacity and coordinating mechanisms are not there. We are hoping that if we create the National Health Research Authority then we will have a department specifically for supporting this, to recruit and develop young researchers. I was hoping that we could have another department completely responsible for the regulatory framework, to look at generating the resources and money to allow people to conduct research.* (KII, MoH, Zambia)

> *We wanted to have a semi-autonomous institution to help us translate research findings into a policy language which we can use. Some of our colleagues in policy don't understand technical research language, so you need to translate it for them.* (KII, MoH, Zambia)

3.6 Outcomes to impact

For the intended outcomes of the programme to result in improved health outcomes a number of assumptions need to hold, according to our ToC. These include improved health policies as a result of the 3DE model, availability of financial resources, and implementation of the policies. Moreover, the programmes are assumed to be implemented as planned and targeted at the appropriate population group and geographical area. The target populations are also assumed to respond to and utilise the programmes or interventions. Finally, it is assumed that the predicted gains in terms of efficiency, equity and outcomes are actually realised once programmes and policies have changed and, where gains are made (e.g. increased efficiency), those are retained and reinvested in the health sector.

As the expected timeframe for change from the outcomes of the programme to impact were expected to be significantly longer than the actual duration of the pilot, the current evaluation did not aim to test the impact of the 3DE pilot on health outcomes. However, based on the evaluation findings and actions taken so far, the expected impact would not be transformational. Some suggest a reinforced implementation of the status quo (EID). Others suggest potential for some cost savings, though only in some contexts in Zambia (ITN), and with careful attention to ensure replicability of results. Others suggest a potential saving but largely for donors, as the provision of Mama Kits is not currently seen as likely to be taken up by the government.

3.7 Other contextual and internal explanatory factors

The 3DE programme and its achievements are also affected by wider contextual factors including the extent of currently unmet demand for evidence, especially impact evaluation evidence, by the MoH and the government more broadly. The contextual factors were explored through interviews, a literature review and in the PEA in Zambia, so as to better understand how evidence is demanded/used, by whom and how in the policy formulation process, and how this relates to officials engaged in the question-sourcing process and is aligned with the 3DE programme. The interviews also probed internal factors – explanatory factors linked to how the 3DE programme was designed, established and managed. Summaries of main external and internal factors are presented below.

3.7.1 External factors

Prioritisation and funding of research

There was concern in Zambia that, although some effort and progress has been made in shifting priorities for research, the overall emphasis of research priorities was still heavily skewed by the bias of the funding available from external partners.

One factor is the limited decision space the MoH has, particularly given the limited flexible funding available. Districts, for example, in Zambia are encouraged to put research questions into their annual plans, but it is not clear how often these are funded – the impression is that most are cut when funding shortfalls occur.

Nine other ongoing health sector RCTs in Zambia and Uganda were picked up in our literature review (some of which were conducted by IDinsight and CHAI but outside of 3DE), but all are externally funded. An analysis of budget sub-heads within the Department of Disease Surveillance, Control and Research indicates that, despite the provisions for research, the departmental activities under each sub-head do not have research components factored in (see Annex F).

Clarity of inter-ministry roles and capacity

Another factor relates to the clarity of roles and capacity. The MoH/MCDMCH split in 2012 has left a considerable legacy of confusion around functions, and within the MCDMCH there is no research team. Given the operational nature of this Ministry, such a research and M&E group, had it existed, would have been a natural partner for 3DE:

> *The MoH was split into two in 2012. MCDMCH didn't have the budget or time to organise itself; it went from around 1,500 staff to around 7,000. A massive explosion of personnel and structure. I came in 2014 and they are still grappling with the role, running from pillar to post. There are some strong individuals and a director who is dynamic and competent, but the vision and strategy of the organisation is all over the place and there is a lot of firefighting. In this context it is very difficult for the programme to be demand-driven and it is mainly looking for opportunities and pushing for things. (KII, Zambia)*

Furthermore, despite there being a Directorate for Disease Surveillance, Control and Research, the new MoH structure provides for only two officers that are charged with the responsibility for research. In order to improve the contribution of research to health outcomes and health equity, the MoH has been working on developing a national health research system that links actors, resources and stakeholders. Under the current institutional arrangements, there is a provision for the National Health Research Advisory Committee (NHRAC), whose overall responsibility is to advise the MoH on all matters related to health research in Zambia. However, the NHRAC secretariat is considered weak in that it does not have a specific office and nor does it have operational funds. The National Health Research Act (2013) ratifies the establishment of the National Health Research Authority but the government is still in the process of establishing its Board.

Within Uganda, there is reported to be a plan for a cross-sectoral research agency to provide evidence to support the national development plan. According to key informants, a central MoH research strategy and a coordination function are currently absent.

Research and evaluation supply and networking

Despite its importance, health research is often a fragmented, competitive and highly specialised activity with researchers in different disciplines working in isolation. A Study on the Demand for and Supply of Evaluation in Zambia (CLEAR 2013) discusses several challenges. It notes that the supply of evaluation expertise needs further development before evidence can be used and evaluation done. Consultancy firms and individuals have arisen with specific areas of strengths in response to demand from funding partners but it is observed that qualified staff are leaving for better-paid positions elsewhere.

Feeding research into policy and practice is also challenging, especially for local researchers with less funding and access. Key informants note that MoH staff have limited time to engage with research results. Local researchers report the difficulty of maintaining engagement with MoH staff. There are bi-annual conferences to share research results in Zambia, but these are more for facilitating networks between researchers than as a linkage to policy-makers. There is, however, a general shift noted by some in the direction of evidence-based decision-making:

> *I think it [evidence-based policy] has improved. If you look at proposals there is always a reference to evidence, and most health programmes are implemented by partners who are required to put in a bit of M&E. I'd say that on a scale of one to 10 we are at five now.* (KII, 3DE, Uganda)

Demand and incentives

The study on the Demand for and Supply of Evaluation in Zambia (December 2013) also notes that there is very little actual demand from stakeholders outside funding/development partners, and that it is unclear how the gathered information feeds back into accountability and ultimately into

performance. It mentions that evaluations touching on sensitive areas such as resource allocation or infrastructure need to be taken with 'consideration', suggesting that there might be potential conflicts of interests at play. This is also highlighted in a comment on the increase in political pressures in Uganda:

> *Overall in Uganda there has been a declining appetite for the use of evidence in policy-making since the 1990s due to the reintroduction of party politics. The political space has become increasingly fractured, causing policy-making to be increasingly politicised: driven by personal allegiances and individual priorities… The overall impression is that there is interest at some level in the production and use of evidence, but uptake is limited due to political factors, constrained budget and personnel changes.* (KII, international, Uganda)

The donor funding landscape and government/donor relationships

Changes in the donor funding landscape affected the MoH's programmatic options, which affected the actionability of priority evaluation questions, according to the 2013 annual review. In Uganda, the shift of Global Fund funding priorities to commodities was said to limit the National Malaria Control Programme's resources and control for scaling supporting interventions, which in turn narrowed the options for impactful 3DE questions.

Government/donor relationships are key to evidence uptake in contexts, like Zambia, where donor agreement and funding are crucial to the implementation of policy. This was well understood by 3DE, but managing these complex dynamics at various stages of the programme still required intensive efforts.

3.7.2 Internal factors

Strong starting base of CHAI in Zambia

In the case of Zambia at least (without a field visit to Uganda it was harder to assess there), one of the factors supporting the effective engagement with the MoH was the very established presence of CHAI and its strong track record of working with the Ministry on health system challenges prior to 3DE. Although 3DE hired new programme staff, they were able to benefit from the wider institutional experience.

Lack of resources to follow up decisions

In the programme design, there was no funding for 3DE to assess any follow-up to decisions made, or to support implementation beyond the stage of the MoH reaching a decision based on the evaluations. This has implications in terms of not allowing an assessment of or support for impact. Where the programme management has other programmes embedded in the MoH, as CHAI does, some follow-up can be expected; however, without this there is a risk that evaluation findings lack an ongoing external 'champion'.

Lack of presence at the district level was also highlighted by some as a factor mitigating against being able to support the implementation of evaluation findings, since districts have some discretion and their own planning functions in the health sector.

Lack of funding for intervention implementation

The constraint of not having funding for implementation was mentioned by key informants in Uganda, thus compounding the difficulty of finding suitable evaluation questions. Others saw this as a positive feature – making later scale-up more likely, in the event of positive findings.

Staffing

Different staffing patterns were found in Uganda and Zambia for the 3DE programme, which may in part explain the different outcomes (at a basic level, the Zambia programme met its target of four evaluations, while in Uganda only one will be produced – it is assumed – by the end of the project cycle). The Zambia programme certainly benefited from the presence of the AAT and a full-time programme coordinator from the start, as well as the IDinsight team, which set up an office in Lusaka at the start of the 3DE programme. In Uganda, the question-sourcing stage was managed by a part-time CHAI member of staff, who was also heavily involved in other malaria-related work, with visits from external supporting experts. While IDinsight was actively engaged in sourcing of questions, this was not the case in Uganda, where CHAI took the lead. It seems that this arrangement failed to provide the momentum the programme required, as well as leading to 3DE focusing exclusively on malaria in the first period in Uganda.

Prior interests and expertise

While the question sourcing was open to all programme areas in Zambia (unlike in Uganda), CHAI had a strong track record working in supply chain management and communicable diseases, and many of the topics did finally fit quite closely with these areas of expertise (with Mama Kits being the exception). It is indeed practical to focus on areas where the organisation not only has expertise but can also follow up, in terms of future support to the Ministry with implementation. There may, however, be a tension between being driven purely by demand and having specific skill and interest sets.

Partnership breakdown

The partnership between CHAI, which was meant to focus originally on the sourcing of questions and catalysation, and IDinsight, which brought in impact evaluation skills, was ended in July 2014, just as three of the evaluations were completed. This may have affected some of the outcomes; although the hand-over was well managed, there is likely to have been a loss of institutional memory in the change-over of personnel.

3.8 Implication for assumptions

One of the goals of the evaluation was to test and revise the ToC that had been developed in the inception report. Having applied it to the context of the 3DE programme, most stages, nodes and assumptions were judged to be relevant questions to ask of future comparable programmes. We have, however, made some minor changes (see the revised ToC in Annex D). In summary:

- In relation to sourcing, the need for an agreed question list seemed less important than arriving at questions which the MoH saw as relevant and high priority and we have rephrased as such.
- The need for the Ministry to be involved in developing, weighting and applying criteria for prioritisation has been added to this slide, along with the (rather basic) requirement that enough questions must exist which meet the criteria.
- In terms of the conducting of evaluations, it was not obvious that rapidity is always needed. It may be in some cases but not in others (and has costs), thus timeliness has been emphasised instead.
- On reporting, the need for presentations to be robust, clear and accurately reflect data findings, strengths and limitations has been added.
- In relation to catalysation, the need for policy implications might be broadened to operational implications, if evaluations are focusing – as in this case – on programme delivery.

- For the cooperating partners' node, the issue of credibility of evidence arose in the evaluation and has been added as an assumption.

- The indicators of capacity and interest at individual and organisational level have been tightened up.

4. Conclusion and recommendations

4.1 Main conclusions

The overall evaluation question was whether the 3DE model has been successful in its stated goal of supporting and increasing evidence-based policy-making, building capacity and changing the behaviour of Ministry staff in demanding and using evidence. The answer, based on the evidence available to the evaluation team, and given the current stage of the programme, is that there has been very limited contribution to changing evidence-based policy-making, capacity and behaviour in both countries. The main reasons behind this conclusion are two-fold:

1. This goal was inherently over-ambitious for a three-year pilot. The overall goal, particularly in terms of building capacity and changing behaviour, requires a longer timeframe and a different focus; and

2. The programme had a number of aspirations that were not all compatible with one another.

3DE aimed to be demand-led, focused on robust impact evaluations, rapid/responsive and affordable, and catalysing action. A number of tensions or trade-offs exist within and between these aspirations.

Demand-led versus methods-led

Ministry staff need evidence but only a small sub-set of this need is for impact evaluations. As 3DE was set up to test the feasibility of undertaking impact evaluations specifically, it was unable to be fully responsive to Ministry priorities and legitimate research questions, many of which required more exploratory methods, such as situation analyses, diagnostic probing of routine data, qualitative studies and operational research. They had finally to choose between being methods-led and being demand-led, and chose the former.

One of the assumptions was that demand existed but was not being met by current evidence providers and brokers. However, it is not clear that demand for impact evaluations is strong, and in that sense 3DE was trying to stimulate demand (by showing how impact evaluations could be used) rather than responding to demand.

Trying to combine responsiveness, rapidity, affordability, robustness and actionability

The evaluation finds that, while some of these qualities are possible to combine, trade-offs between them have to be managed in practice. For example, responsiveness often implies taking on questions which cannot be answered in a short time – for example, more systemic questions that require longer to assess, develop, pilot, train, embed and evaluate.

Similarly, small budgets can mean small-scale studies that sacrifice external validity and therefore some degree of usefulness to decision-makers. Rapidity may mean releasing results without thoroughly peer reviewing them. Extensive consultation on priorities as part of the responsiveness agenda leads to evaluations that are not low-resource overall (an average of £400,000 per evaluation, when the overall pilot budget is divided by the evaluations produced and about to be produced). There may also be some tension between being a neutral provider of evidence, as implied by the robustness aim, and triggering policy decisions, as is required by the programme logframe.

None of this implies that these trade-offs were badly managed by 3DE. Our main conclusion is that the model was over-ambitious, not that it was badly implemented. However, there needs to be reflection on what are the most important objectives and how to set realistic priorities for the next phase of the programme.

Lessons from the pilot

As the literature review (Annex G) highlights, the factors affecting the likelihood of evidence informing policy decisions can be categorised as those relating to characteristics of evaluation, the evaluation users, and the wider contextual factors respectively. In terms of evaluation users, the literature highlights relatively weak but growing demand. The low demand for evaluations is in part due to the issues around the characteristics of evaluations noted by the 3DE proposal (lack of timeliness, links to practical questions, robustness and clear dissemination). However, it also stems from the nature of knowledge and evidence required (i.e. not only scientific knowledge) and the skills and capacity of policy-makers to understand and engage with them. In addition to the immediate characteristics of the evaluation and evaluator, the review also emphasises the role of political and institutional factors in shaping evidence uptake by policy-makers and the importance of engagement with the political environment.

The 3DE model, as reflected in the original ToC, largely focuses on the characteristics of evaluations. Less explicit in the pilot model are the characteristics of the evaluation users and the wider political context. As a pilot, the 3DE programme has generated important insights into how feasible it is to address some of the common weaknesses of evaluations and the resources which are required to do so, but it did not address the wider needs of users or the contextual factors. To meet the goal of 'supporting and increasing evidence-based policy-making, building capacity and changing the behaviour of Ministry staff in demanding and using evidence', a broader approach would be needed.

Assessing against DAC criteria

Relevance, effectiveness and impact

The 3DE pilot has generated relevant evidence, which has been fed into the policy process. It is less clear that this has been transformational, in terms of changing the nature of demand and evidence use. This is partly because some of the underlying structural issues have not been addressed (and were never intended to be directly addressed by the design of the programme). Lack of incentives for use of evidence, lack of capacity, decision-makers lacking time and resources to devote to research and M&E – all of these barriers remain, and the 3DE model did not aim to address them directly, only indirectly via a good experience of working together on programme activities.

Moreover, it should be acknowledged that, while the model was innovative, there are many groups working in a similar way in close partnership with ministries on evidence-generation. Local research groups, and international research consortia and organisations (such as IPA, 3ie and other research consortia; see Annex F) operate in similar manner, albeit managing the trade-offs outlined above in different ways. 3DE was perceived as responsive and a good partner by the MoH/MCDMCH stakeholders, but not as being radically different from counterparts.

Other groups like ZCHARD may not go with a completely open approach to priority areas, but many of the other stages are conducted in a similar way – with collaborative working with the MoH and engagement to get results into policy, where relevant.

By engaging with mid-level operational staff, 3DE staff made a judgement about where they could most effectively engage – a judgement which was probably correct, especially since it allowed entry below the radar of ministerial politics. However, the nature of the questions addressed also limited the scale of impact, addressing very specific operational questions. Interpreting the findings for different areas of the country (compared to the study conditions) was not straightforward for evaluation users, and ownership of findings varied, with some important players (including CPs) expressing scepticism. While considerable effort was put into dissemination and catalysation, some important stakeholders were not fully clear on the findings and their implications, and in

some cases these were not well communicated (e.g. recommending delayed hang-up of nets when the evaluation findings find no hang-up to be the most cost-effective strategy).

Efficiency

The VfM assessment component of the original evaluation ToR were dropped, but in terms of assessing 3DE against the general DAC criteria of impact and efficiency, we note that the original goal was to deliver eight impact evaluations and that the original budget was estimated at £100,000 per evaluation. In the event, the direct costs of the evaluations averaged $228,000 and only five were delivered (for reasons which are explained above). The overall programme expenditure (including other stages such as question sourcing and catalysation) was £400,000 per evaluation. As noted above, they were often not especially rapid either, and addressed very specific delivery.

Sustainability

Ministries have a variety of evidence gaps, ranging from better routine data provision to more sophisticated research questions. As was outlined above, impact evaluations cannot respond to all needs (and probably can only respond to a small proportion of them). A time-limited programme led by an international organisation, however well connected to the Ministry, as CHAI is, at least in Zambia, is unlikely to represent a strategic long-term solution unless it is coupled with specific actions to embed capacity either in the Ministry or a local organisation that can develop and maintain expertise over time.

Equity

Equity was not prioritised in the original 3DE model, which focused on potential large-scale impact rather than marginalised groups. The focus on finding questions ‘where the money is’ (in order to favour scale-up and actionability) does have implications for neglected areas – meaning that an overlooked condition, such as poor mental health, is likely to continue to be overlooked. This is one of the many trade-offs that had to be managed within the programme. Being ‘demand-led’ may also not favour raising the profile of areas that are currently ‘orphaned’.

4.2 Key recommendations

Ten main recommendations are provided, largely targeted at DFID and relating to programme design. The first five suggest potentially different approaches. The second five could be implemented by making more minor changes to the existing model.

Recommendation I – Agree on focus and design accordingly

In the next phase, it will be important to agree on the core objectives of the programme, and to tailor it accordingly.

Different objectives imply different models. If, for example, capacity building is the core need, then a programme which focuses more closely on training, working with and within ministries, and providing support for ministerial units would be most appropriate.

If the diagnosis is that there is a lack of supply of quality evidence for ministries, then investment should focus on developing local academic units, connecting them within research networks and establishing local brokerage of knowledge, tailored to the needs of the MoH.

If the focus is on improving service delivery, then more resources should be provided for following up research with implementation support to governments or other providers.

The 3DE programme appears to have been implicitly about generating demand for impact evaluations – not so much being demand-led but creating an awareness of and willingness to

engage in 'robust' research. Demand generation is also a valid function, but different in its needs from the models above. As the literature review highlights, it may sometimes be necessary to motivate demand for evaluation evidence through various strategies, such as the carrots, sticks and sermons described by Mackay (2007) or the capacity-building approach of the CLEAR initiative (see Annex G).

Recommendation II – Tailor to context

Clearly not all countries will have the same evidence needs and so a starting point for programming should be an understanding of the local institutional and market context, to understand what the gaps are, and what existing institutions or networks could be strengthened. Which of the nodes in the ToC are weakest in a given context? These should be the focal areas for support.

Recommendation III – Invest more in evaluative thinking and capacity

Capacity building was an intended indirect benefit in the pilot phase but should receive more priority in order to ensure a lasting legacy. The legacy of the programme should be increased evaluative thinking and capacity within MoH and MCDMCH to scope, oversee, quality assure and use evaluations. This includes:

- At *problem diagnosis*: being able to frame questions that need answering in terms of evaluations;
- At *planning*: developing a ToC, an improved operational plan and a solid resourcing framework for the intervention;
- At *implementation and monitoring*: developing improved indicators for implementation and designing a monitoring system; and
- At *outcome & impact*: defining the desired changes, effectiveness in achieving them and VfM;

It is important to distinguish between being responsive to demand and being demand-driven. A good supplier will establish what the demand is before addressing it, but this does not mean the supplier is being demand-driven: they just have a well-targeted approach to supplying the evidence that has been requested. The 3DE model, with its emphasis on scoping the question, ensures that it responds to the demand for evaluation evidence and that what it supplies is well targeted toward questions that need answering. But this does not necessarily mean that it is a demand-*driven* approach.

The demand for evidence 'encompasses both the *capacity* to find, evaluate and use different forms of evidence and the *motivation* to use them to make evidence-informed policy'. A demand-driven approach should be concerned with 'influencing the behaviour of decision-makers such that they access and use a range of research sources, not only those which they have commissioned.'[14] Building capacity and motivation means strengthening individual skills, seeding new practices, learning by doing, sponsoring champions, building networks, and supporting institutional processes.[15]

Ultimately, the definition depends on where the resources are being allocated: to stimulate a well-targeted supply of evidence, or to influence decision-makers to access a range of evidence sources, one of which is evidence from impact evaluations. 3DE has built up the supply side of the

[14] Both quotes from Newman, Fisher and Shaxson (2012) *Stimulating demand for research evidence: what role for capacity building?* IDS Bulletin Vol 43 No 5, Sept 2012.
[15] See BCURE Common Theory of Change on page 34 of Vogel, Punton and Lloyd (2015) *Evaluation of approaches to build capacity for use of research evidence*. Draft inception report, submitted by Itad to DFID January 2015.

equation by focusing on providing better evidence from impact evaluations, and by ensuring that the supply addresses important questions. But it is only stimulating demand for one form of evidence alone, which risks missing a broader opportunity to strengthen evaluative thinking and capacity within MoH and MCDMCH. However good 3DE's approach is to identifying evaluable questions, without corresponding work to build capacity at the individual, network and institutional levels it cannot really be said to be a demand-driven system (see Box 5 for an example of a more systemic approach in South Africa.)

Box 5 What are others doing? An example from South Africa

In South Africa, the Department of Planning, Monitoring and Evaluation (DPME) is building a demand-driven national evaluation system which ultimately aims to devolve responsibility for commissioning evaluations to departments. DPME develops a rolling three-year annual plan covering six types of evaluation: diagnosis, design, implementation, impact, economic, and evaluation synthesis. DPME contributes technical advice on evaluability issues and evaluation methods, and contributes up to a maximum of ZAR 500,000 (approximately £25,000) to each evaluation. Final evaluation reports are submitted to formal meetings of directors-general across government, and DPME requests that the custodian department for each evaluation draws up an implementation plan against which progress is reported at six-monthly intervals.

Unlike 3DE, DPME is a government department in its own right with considerable institutional weight. However, it has four years' experience in trying to implement a demand-driven system[16] and it is worth considering its analysis of the factors that have promoted and impeded a focus on results measurement to strengthen the 3DE model. High-level political commitment to M&E, a strong individual driver for M&E within a department, and the opportunities to learn from others' experiences have all contributed to an improved use of M&E evidence. Factors that have hindered the use of M&E evidence include a compliance culture, the fact that M&E is not seen as part of the strategic function, issues in terms of coordination between departments, the quality of administrative data, and evaluation capacity within government and evaluation suppliers.

Recommendation IV – Embed in local institutions

Whatever the focus chosen, the programme should be embedded in local institutions, with support provided externally as needed but with key staff commissioning, coordinating or brokering based within the Ministry or local research networks and organisations. Where new and complex skills are being developed, there should be a co-working period, but alongside staff in local institutions (a 'build–operate–transfer' model). This would also allow more flexibility about seizing policy 'windows', rather than having to identify them within the constraints of a short-term programme.

Recommendation V – Change the performance targets

In the 3DE programme, contributing to a policy decision was a key performance target. While this kept minds focused on the need to get take-up of research there is also a potential conflict of interest between being a supplier of research and helping ministries to analyse and use evidence in a neutral way. The policy 'decisions' which 3DE had to influence and document were somewhat artificial and just one part of a continued debate and evolution of programming strategies. Is policy change what DFID really wants? Or is it increases in the MoH's ability and willingness to take informed decisions using 'good enough' evidence? If it is actually the latter, then the performance metric would need to be different.

More specifically, if 'policy decision' is used as a target, then it should be broadened to include implementation. Many of the changes potentially implied by 3DE's work were operational, rather than at the policy level.

[16] See Phillips, Goldman, Gasa, Akhalwaya and Leon (2014) *A focus on M&E of results: an example from the Presidency, South Africa*. Journal of Development Effectiveness 6:4, 392–406.

Recommendation VI – Enlarge the toolkit

We question the privileging of impact evaluations as a higher form of knowledge. They have their own limitations, particularly in terms of generalisability, and commonly fail to provide good insights into the 'how, why and in what contexts' questions. Ministries rightly look for a range of information, including on the equity, sustainability etc. of interventions. Demand-generation or evidence-supply programmes should focus on supporting and providing appropriate tools for different questions.

Recommendation VII – Timeliness, not rapidity, should be the goal

Evidence should fit with policy needs, but rapidity has costs and is not always required or appropriate to the question. Timeframes should follow on from the question for which the MoH needs an answer – not dictate the question. In some cases, having a longer time period would generate more useful and valuable information for the MoH than one with artificially constrained fieldwork periods.

Recommendation VIII – Monitor VfM

Information on expenditure in 3DE was not reported for the different stages of the programme, with the result that the cost-efficiency of different stages could not be assessed (we cannot say, for example, how much of the budget was spent on question sourcing, which would be interesting, given that this was a distinctive feature of the programme). In the next phase, this information should be systematically reported.

Recommendation IX – Ensure quality assurance at all relevant stages

In the pilot programme, the peer review of products appears to have been at the stage of developing protocols, while at report-writing stage there was no quality assurance process that the evaluation team is aware of. Peer reviewing of final products is important to ensure that findings are robust and accurately presented.

Recommendation X – Take a broad approach and ensure adequate support

The differential success in Uganda and Zambia – both environments judged to be initially receptive to an evidence-based approach – suggests some practical lessons for the next phase, including the wisdom of taking a broad approach to ministerial needs (rather than being locked in to relationships with specific programmes) and also of ensuring adequate staffing to drive forward what has been an intensive process, if a similar approach is adopted.

References

A. Hyder, A. Corluka, P. Winch, A. El-Shinnawy, H. Ghassany, H. Malekafzali, M. Lim, J. Mfutso-Bengo, E. Segura and A. Ghaffar (2010, "National policy-makers speak out: are researchers giving them what they need?," Health policy and planning.

Centre for Learning on Evaluation and Results Anglophone Africa (CLEAR-AA) (2013) 'STUDY ON THE DEMAND FOR AND SUPPLY OF EVALUATION IN ZAMBIA', http://www.theclearinitiative.org [27 April 2015]

CHAI and IDinsight (2012), Demand-Driven Evaluations for Decision: Technical Proposal, July.

CLEAR Secretariat (2013) 'Building Blocks of CLEAR's Capacity Development Strategy Change Agents – Capacity Outcomes – M&E Capacity Development Activities', http://www.thecelarinitiative.org [27 April 2015]

CLEAR Secretariat (2013) 'CLEAR STRATEGY (2013 - 2018) - Development Anchored in Evidence, Learning, and Mutual Accountability', http://www.theclearinitiative.org [27 April 2015]

DFID (2012), Business Case and Intervention Summary: Demand-Driven Evaluations for Decisions (3DE).

DFID (2013), CHAI 3DE Annual Review – 2012/13, October/November.

DFID (2014), CHAI 3DE Annual Review – 2013/14, October/November.

ePact (2014) 'CLEAR Mid Term Evaluation, Final Evaluation Report', http://www.thecelarinitiative.org [27 April 2015]

ePact (2014) 'CLEAR Mid Term Evaluation, Inception Report', http://www.thecelarinitiative.org [27 April 2015]

European Commission, Evaluation Unit of the Directorate General for Development and Cooperation, EuropeAid (2014) 'Study on the uptake of learning from EuropeAid's strategic evaluations into development policy and practice', https://ec.europa.eu [3 May 2015]

Fisher and Shaxson (2012) *Stimulating demand for research evidence: what role for capacity building?* IDS Bulletin Vol 43 No 5, Sept 2012.

International Initiative for Impact Evaluation (2014) 'Annual report 2014: Evidence – Influence – Impact ', http://www.3ieimpact.org [3 May 2015]

International Initiative for Impact Evaluation (2014) 'Replication Paper 1: Quality evidence for policymaking - I'll believe it when I see the replication', http://www.3ieimpact.org [3 May 2015]

J. Bossuyt, L. Shaxson and A. Datta (2014), "Study on the uptake of learning from EuropeAid's Strategic Evaluations into Development Policy and Practice. Final Report," On behalf of the European Commission, June.

J. Nabyonga-Orem, F. Ssengooba, R. Mijumbi, C. Kirunga Tashobya, B. Marchal and B. Criel (2014), "Uptake of Evidence in Policy Development: the Case of User Fees for Health Care in Public Health Facilities in Uganda," BMC Health Services Research, vol. 14, no. 639.

J. Rutter (2012), "Evidence and Evaluation in Policymaking: A problem of supply or demand?," Institute For Government.

K. Johnson, L. Greenseid, S. Toal, J. King, F. Lawrenz and B. Volkov (2009), "Research on Evaluation Use: A Review of the Empirical Literature from 1986 to 2005'," American Journal of Evaluation, vol. 30, no. 3, pp. 377-410.

K. Mackay (2007), "How to Build M&E Systems to Support Better Government, Washington D.C," the World Bank, Independent Evaluation Group (IEG).

K. Oliver, S. Innvar, T. Lorenc, J. Woodman and J. Thomas (2014), "A systematic review of barriers to and facilitators of the use of evidence by policymakers," BMC health services research, p. 14.1.

M. Liverani, B. Hawkins and J. Pankhurst (2013), "Political and Institutional Influences on the Use of Evidence in Public Health Policy. A systematic Review," Plos One, vol. 8.10.

Mayne. 2008. Contribution Analysis: An approach to exploring cause and effect. ILAC Brief 16. May 2008

Mayne. 2010. Contribution Analysis: Addressing Cause and Effect. In R.Shwartz et al, Evaluating the Complex. New Brunswick, NJ; Transaction Publishers

N. Jones, A. Datta and H. Jones (2009), "Knowledge, policy and power: Six dimensions of the knowledge-development policy interface," Overseas Development Institute (ODI).

Phillips, Goldman, Gasa, Akhalwaya and Leon (2014) *A focus on M&E of results: an example from the Presidency, South Africa*. Journal of Development Effectiveness 6:4, 392–406.

Phillips, S. et al. (2014) 'A focus on M&E of results: an example from the Presidency, South Africa' in Journal of Development Effectiveness Vol. 6 (4), 392-406

Presidency of the Republic of South Africa, Department of Performance, Monitoring and Evaluation (2013) 'DPME Guideline 3.1.4: Improving the Operation of M&E in Offices of the Premier', http://www.thepresidency-dpme.gov.za [27 April 2015]

Presidency of the Republic of South Africa, Department of Performance, Monitoring and Evaluation (2011) 'National Evaluation Policy Framework', http://www.thepresidency-dpme.gov.za [27 April 2015]

Presidency of the Republic of South Africa, Department of Performance, Monitoring and Evaluation (2011) 'STRATEGIC PLAN 2011/12 – 2015/16', http://www.thepresidency-dpme.gov.za [27 April 2015]

Punton and Lloyd (2015) *Evaluation of approaches to build capacity for use of research evidence*. Draft inception report, submitted by Itad to DFID January 2015

Regional Centres for Learning on Evaluation and Results (CLEAR) (2013), "Demand and Supply: Monitoring, Evaluation, and Performance Management Information and Services in Anglophone Sub-Saharan Africa: A synthesis of nine studies".

Republic of Zambia, MINISTRY OF HEALTH (2010) 'NATIONAL HEALTH STRATEGIC PLAN 2011-2015', http://www.moh.gov.zm [28 April 2015]

S. Sutcliffe and J. Court (2005), "Evidence-based Policymaking: What is it? How Does it Work? What Reklevance for Developing Countries?," Overseas Development Institute.

Treasury Board of Canada Secretariat (2012), Theory-based approaches to evaluation: Concepts and practices.

UKaid (2014) 'EVIDENCE INTO ACTION TEAM PROGRAMME GUIDE: A guide to programmes funded by the Evidence into Action Team', https://www.gov.uk [29 April 2015]

United Nations Development Programme (2013) 'Millennium Development Goals - Progress Report Zambia', http://www.za.undp.org [28 April 2015]

World Bank (2012), "Impact Evaluations: Relevance and Effectiveness," Independent Evaluation Group, Washington DC.

World Bank Group (2014) 'RBF – A Smarter Approach to Delivering More and Better Reproductive, Maternal, Newborn, and Child Health Services', http://www.rbfhealth.org [28 April 2015]

World Bank Group, Independent Evaluation Group (2006), ECD Working Paper Series No. 16 'Experience with Institutionalizing Monitoring and Evaluation Systems In Five Latin American Countries: Argentina, Chile, Colombia, Costa Rica and Uruguay', http://ieg.worldbankgroup.org [4 May 2015]

World Bank Group, Independent Evaluation Group (2014), ECD Working Paper Series No. 29 'Monitoring and Evaluation System: The Case of Chile 1990–2014', http://ieg.worldbankgroup.org [4 May 2015]

Websites consulted:

http://devpolicy.org

http://idinsight.org

http://ieg.worldbankgroup.org

http://www.3ieimpact.org

http://www.afrea.org

http://www.africaevidencenetwork.org

http://www.bu.edu

http://www.cabinetgovernment.net

http://www.drussa.net

http://www.fhi360.org

http://www.gdn.int

http://www.inasp.info

http://www.isrctn.com

http://www.moh.gov.zm

http://www.nationalservice.gov

http://www.rbfhealth.org

http://www.theclearinitiative.org

http://www.thepresidency-dpme.gov.za

http://www.unicef.org

http://www.za.undp.org

http://www.zambart.org

https://becureglobal.wordpress.com

https://ec.europa.eu

Annex A The Original Terms of Reference

Below is the original Terms of Reference (ToR). This ToR is superseded by the Inception report that constitutes the new terms of reference for this evaluation.

Terms of Reference

Independent evaluation of the Demand-Driven Impact Evaluations for Decisions (3DE) Pilot, implemented by the Clinton Health Access initiative on behalf of DFID

Overall Purpose

The main purpose of this evaluation is to help DFID and partners to learn about the innovative (rapid and demand-driven) evidence model of the 3DE pilot. The specific focus of the evaluation is to develop and test the Theory of Change (TOC) of the pilot and to assess its overall efficiency, effectiveness and impact. This should include a rigorous and independent assessment of the quality and relevance of 3DE evaluations and their uptake by policymakers; an analysis of 3DE's value for money; and identifying lessons and implications for DFID as it looks at future options for commissioning innovative evaluation programmes. Evaluation findings are expected to be used by DFID and CHAI for future programming decisions.

Background and Context

3DE Programme Outline
The UK has provided £2,000,000 over 4 years (2012/13 – 2015/16) to the Clinton Health Access Initiative (CHAI) to pilot a demand driven approach to health impact evaluations. The funding is to finance impact evaluations of health interventions, generated by demand from Ministries of Health in Uganda and Zambia (see section 11 for the 3DE Business Case).

The expected results of the pilot are a rigorous evidence base that is generated through local demand, from which improved policies and programmes to address health outcomes can be developed and implemented. It is expected that because the questions for the 3DE impact evaluations are selected based on local policy priorities and demand for evidence, the findings are more likely to be utilised than from donor driven initiatives for evidence.

A minimum of 5 impact evaluations are expected to be completed by CHAI during the pilot (output). The primary outcome of the pilot is that evaluation findings are used to inform a managerial decision (outcome). The business case specifies that 4 types of potential decisions can be made based on evaluation evidence:

1. Conduct more research
2. Reduce or abandon an intervention
3. Adjust the design of an intervention
4. Increase the scale of an intervention

It is expected that evaluation findings will influence a managerial decision in one of the four ways listed above.

A secondary expected outcome is that the 3DE pilot changes how Ministries of Health think about and use evidence in the process of innovating, learning and improving their policies and programmes

DFID expects to achieve the following impacts over a longer period:

- Improved delivery of national health programmes through better use of evidence and research.
- Improved health outcomes for the poorest.

Context

There has been considerable interest in rigorous impact evaluations in recent years, due to increasing drive for cost-effective interventions and value for money, and to gain a better understanding of the effects of interventions. This movement towards evidence-based approaches hinges on field experiments, including randomised controlled trials, to rigorously quantify the impact of programs and interventions. However, the typical multi-year evaluation time frames are usually too slow for pressing policy deadlines, and high evaluation costs are impractical for most resource limited settings. This means rigorous evidence is only available for a small subset of high-profile programs and policies, with topics often determined by evaluating institutions and the global research community rather than in-country policy makers. As a result, field experiments may fail to optimally benefit managers in developing countries responsible for implementing large-scale health programs, and large portions of the billions of dollars committed to global health are inefficiently spent each year.

DFID has been at the forefront of efforts to make better investments in health and development through the use of evaluation and research, including through 3ie[17] and the Strategic Impact Evaluation Fund.[18] However, there are still significant obstacles to ensuring that the growing number of robust evaluations and studies have a meaningful impact on major policy and spending decisions. In response, 3DE provides a potentially powerful additional tool to overcome those obstacles, complementing support to 3ie and similar groups with an intensive focus on country policy-maker needs and the use of targeted evaluations to address them. The 3DE approach is coherent with DFID's commitment to commissioning research that seeks to find better ways of delivering existing health interventions and scaling those health solutions to more people.

In line with the Paris Declaration DFID recognises that national ownership is fundamental in formulating and implementing evaluation findings. DFID's policy on evaluation already states that stakeholders must be involved in evaluations throughout the process. However, there are concerns that donor-driven initiatives may not always be able to engender the support required from national ministries to accept and take up the findings of evaluations.

The existing international literature identifies a number of factors that influence the use of evaluation findings. These can be split as either relating to the characteristics of the evaluation (for example, relevance, credibility, quality, findings, communications and timeliness) or the context in which the evaluations are conducted (for example, commitment, political and financial climate, information needs, competing information, personal characteristics, decision-making climate and personalities). A range of publications and leaders in the field of evaluation have recognised both the growing demand for evidence from policy makers and managers and the disconnect between the demand and the structure, timing, and cost on which typical research and evaluation is conducted. There is, however, limited evidence on the assumptions from which the 3DE model is taken (as set out in the business case), both in Zambia and Uganda, as well as in the international development context more broadly.

Reviews and Other Documents

This evaluation will build on earlier reviews and other background documentation provided by CHAI and DFID (some of the key documents are attached in section 11 of this TOR). Two annual reviews of 3DE have been carried out by DFID in November 2013 and 2014 respectively. For the last annual

[17] http://devtracker.dfid.gov.uk/projects/GB-1-200135/
[18] http://devtracker.dfid.gov.uk/projects/GB-1-203933/

review, a DFID Evaluation Adviser travelled to Lusaka and interviewed a wide range of stakeholders and partners. The review has found that overall 3DE is on track to deliver expected results as operationalised in the Logframe. However, it also identifies a number of challenges around the basic assumptions underlying the pilot's theory of change, including for example the impact of political and institutional constraints upon the implementation of decisions made based on 3DE evidence. Also, a year after the 3DE pilot's launch DFID and CHAI have agreed to focus efforts on Zambia where roughly 80% of 3DE's work is taking place now. The TOR for the independent evaluation will build on the findings from the previous monitoring missions, which are provided along with other key documents in section 11.

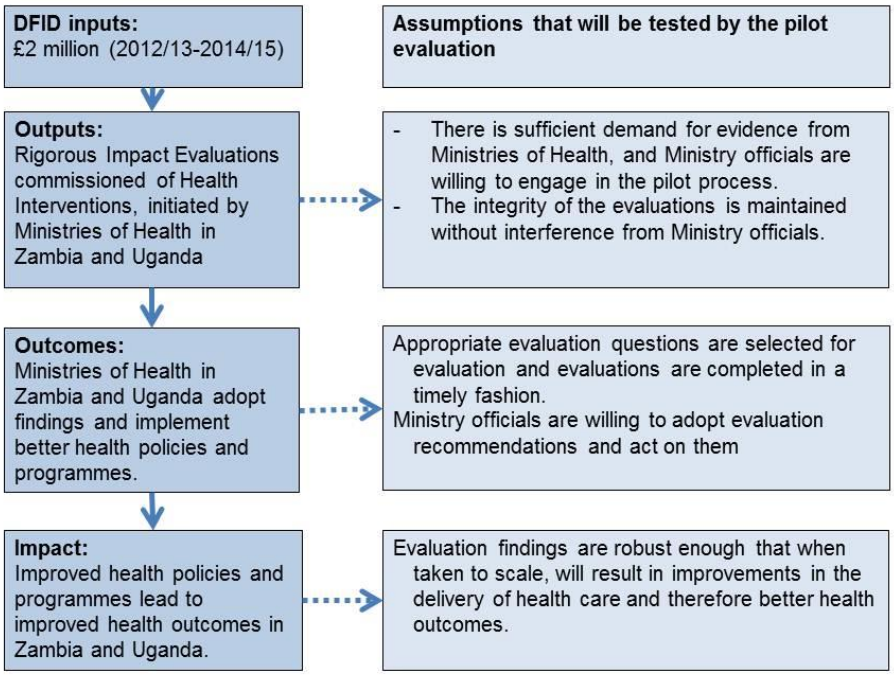
Objectives, Scope and Evaluation Questions

Objectives and Scope

The evaluation has three main objectives, listed in order of priority, and focusing on three out of the five OECD DAC evaluation criteria (efficiency, effectiveness and impact):

a) To refine and develop in more detail the TOC of the pilot and to test whether its assumptions hold in practice, specifically regarding evidence uptake and the role of political economy challenges, as well as the reasons for differences in 3DE outcomes and impact between Zambia and Uganda (see graph below for the initial TOC outline)

Theory of Change Outline from the 3DE Business Case



The 3DE business case outlines a broader set of outcomes, which could be explored and reflected in a refined Theory of Change and its verification. They include e.g.: Higher ownership over evaluation questions by national stakeholders increases the probability of evidence uptake; Relevant National Ministries adopt and implement findings from 3DE evaluations; Increased in-country demand and capacity to use evidence; Increased global awareness of 3DE model and findings. The TOC outline in the business case and the pilot's logframe do not articulate these components sufficiently, yet they seem fundamental to achieving the broader impact level objectives of this pilot.

b) To analyse the Value for Money of the 3DE pilot and the five evaluations conducted within the project, with a focus on efficiency and cost-effectiveness (DFID is currently developing a systematic approach for valuing costs and benefits of evaluations. If feasible, the 3DE evaluation might partly build on this approach).

c) To assess the quality of the 3DE pilot products (the impact evaluations), and understand what the impact of the innovative rapid and demand driven model is on their quality.

The scope of this evaluation covers the period from April 2012 until 2015 (time of the commissioning of the evaluation).

Evaluation Questions

Building on the objectives outlined above, and on the recommendation from the 2013/14 annual review of the 3DE pilot the primary evaluation questions are:

a) Innovative 3DE evidence model: Do the main assumptions of the innovative (demand-driven and rapid) 3DE evidence model hold in practice, especially at the output and outcome level (e.g. has the demand-driven character of the pilot led to increased evidence uptake)? How did the 3DE pilot engage with different stakeholders, and how did this contribute to the pilot's outputs and outcomes (e.g. during question sourcing and dissemination)? Are the benefits that the five 3DE evaluations anticipate from evidence uptake actually occurring? What are the reasons for the differential outcomes of the 3DE pilot in Zambia and Uganda in practice? *(OECD DAC Criteria: Effectiveness, Impact)*

b) Value for Money: Is it cost-efficient and cost-effective (given the significant initial set up costs) to try and influence policy with the 3DE model? Are impact evaluations the most appropriate and cost-efficient type of evidence for the 3DE model? *(OECD DAC Criteria: Efficiency)*

c) Quality of 3DE outputs: Are the 3DE pilot evaluations rigorous and in accordance with international quality standards (e.g. OECD DAC quality standards)? On quality, the evaluation should analyse whether data quality of 3DE evaluations is sufficient, and how the demand-driven character of questions and rapid collection of data impacted upon the rigour (e.g. internal and external validity) of the final studies altogether. Drawing on answers to questions under b) it will also be of interest to understand the balance between rigour and costs of the 3DE evaluations. *(OECD DAC Criteria: Efficiency, Effectiveness)*

- How well did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included? *(OECD DAC Criteria: Relevance)*

Other potential evaluation questions include:

- To what extent has the 3DE pilot contributed to evaluation capacity building among Ministry officials and other partners? The business case states that "the primary objective of the pilot is to test the 3D Evaluation model. However, CHAI will also work to ensure the sustainability of the pilot's impact by building the relevant capacity of the MoHs that participate in the pilot".
- What, if any, are the main unintended outcomes (positive and negative) of the 3DE pilot?
- To what extent could the model of the 3DE pilot be sustained by partner ministries after donor funding ceased?
- Is a focus on impact evaluations, as opposed to other types of evaluations or research, the most relevant form of evidence for the 3DE model?
- On relevance, it might also be of interest what the process is for evaluation question sourcing, involvement and engagement with national stakeholders, and how relevant agreed evaluation questions are for different types of Zambian and Ugandan stakeholders.
- What was the impact of the 3DE global learning component, e.g. have the 3DE manual and presentations at conferences led to increased global awareness of 3DE model and findings?

The tenderers are expected to refine the priority evaluation questions and select and refine some of the secondary questions put forward above.

Users and audience of evaluation

The main users of the evaluation will be DFID, CHAI and the 3DE partners, especially the Ugandan and Zambian government. It is expected that the findings of the evaluation will deliver insights on CHAI's achievement and challenges encountered, and on the innovative model of rapid and demand-driven evaluations.

More specifically, the target audience for this evaluation includes:

- DFID Evaluation Department and Evaluation Advisers
- DFID Research and Evidence Division
- CHAI 3DE Management and implementing partners
- Partner governments in Zambia and Uganda
- Other donors with an interest innovative models of evidence and evaluation

It is expected that evaluation findings will influence programming decisions by DFID and CHAI. Furthermore, findings could influence future decisions on the commissioning of evidence by partner ministries in Uganda and Zambia. Relevant partner ministries will be involved in the evaluation process from the beginning (see section 8).

Methodology and Data Sources

Tenderers should spell out in detail the evaluation design and methodology they propose to use, the potential risks and challenges for the evaluation and how these will be managed. DFID does not endorse a particular methodology(ies) for the 3DE evaluation, but would expect that priority questions under 3.2.a) of this TOR will be answered with a theory-based approach, whereas priority questions b) and c) will require other methodologies. Therefore, tenderers are invited to propose approaches and methods which they believe will most effectively and efficiently answer the different priority questions and meet the purpose of the study within the time available. The successful tenderer will then refine this proposal within the first month of the contract, in consultation with DFID, CHAI and other relevant stakeholders. Tenderers should note that we are committed to quality and rigour in line with international good practice in evaluation, as set out in DFID's evaluation policy. As per DFID evaluation policy, the evaluation should adhere to international best practice standards in evaluation, including the OECD DAC International Quality Standards for Development Evaluation, the OECD DAC principles Standards for Development Evaluation, and DFID's Ethics Principles for Research and Evaluation.

The methods and assessment frameworks employed for this evaluation should facilitate the collection and analysis of data, be relevant to the questions outlined in section 3 above, and make optimal use of existing data. The evaluation may need primarily to use retrospective evaluation methodology techniques.

Sources that will be used in the evaluation would, at a minimum, include:

- *Document review:* Review of key documents including those outlined in Section 2. A table of key programme and project documents will be prepared by CHAI/DFID and provided to the evaluator with further assistance available if required (Section 11 in this document includes some of the key documents)
- *Quality assurance/peer review of the five impact evaluations conducted under the 3DE banner:* Analysis of data collection documents, 3DE data analysis methods (e.g. random check of STATA do-files and outputs) and final reports. The Evaluation Team may wish to consult / include key health or impact evaluation expert(s) in the bid to assist in assessing quality of 3DE evaluation outputs.
- *Interviews with key partners and users:* Interviews with key stakeholders such as national, district-level and local health policy makers in Uganda and Zambia (governments, donor and civil society), other researchers and practitioners (health researchers, data analysts, former sub-contractor within the 3DE pilot) and key staff members from CHAI and DFID.

These interviews may be done in person if feasible, but most likely by telephone or internet based communication.

- *Surveys or other data collection methods:* to solicit input from additional stakeholders external to CHAI. If surveys are used, these should be rigorously designed with appropriate sampling methods and expectation of acceptably high response rates. Alternative or complementary approaches, such as online discussion fora, could be considered. The evaluator should also consider field trips to investigate possible impact of 3DE evaluations in Zambian or Ugandan districts
- For VfM assessment, data should primarily be drawn from the administrative reporting systems of CHAI, and compare 3DE's efficiency and uptake with other similar IE initiatives, where possible.
- The tenderers might also consider conducting a review of relevant other literature and findings on ensuring relevance and use of impact evaluation findings by governments and policy makers.

Available data: CHAI and DFID will provide tenderers with documentation on the policy impact of the five 3DE evaluations (e.g. operational plans, minutes from meetings between CHAI and ministries), but this will need to be accompanied by additional data collection through qualitative interviews. In particular, there is limited data available on the actual decision-making process in the relevant ministries and 3DE outputs' role in this process. DFID and CHAI will facilitate access to relevant stakeholders in Zambia and Uganda, but the evaluation team will have to make direct approaches to other stakeholders and beneficiaries who are in scope of their evaluation design.

There are some risks and challenges regarding data collection, mostly regarding information on the impact of the 3DE studies on policy decisions, as availability of this information will depend on inputs from Zambian and Ugandan government stakeholders and other partners. The tenderers are expected to identify risks and challenges more specifically in an inception report and propose a mitigation strategy.

Timetable and Milestones

DFID expects tenderers to propose a detailed timetable, having regard to the following:

Activity / Output	**Deadline**
Evaluators selected and contract signed	February 2015
Inception Report and Draft Theory of Change Submitted to Management Group Approach should be finalised in consultation with DFID and CHAI. This Inception report should include - Refined Theory of Change, - Suggestions on refinements/amendments of the evaluation questions, - Full methodology, - Implications for the degree to which the evaluation questions can be answered using a credible and robust evidence base, - Assessment frameworks - Identified sources of data - Risk management strategy - Communications / dissemination plan for the evaluation (intended user groups, dissemination documents, events etc.)	Within 3 weeks of contract starting

Activity / Output	Deadline
Management Group provide feedback, discussion on TOC	Within 5 weeks of contract starting
Data collection and analysis	Weeks 6 – 13 after contract starting
Draft Final report submitted for comments. The report should include: 1. Cover page 2. Table of Contents 3. Executive Summary 4. Purpose of Evaluation 5. Evaluation approach and methodology 6. Limitations of evaluation 7. Response to evaluation questions with supporting evidence 8. General findings, key messages and potential implications and recommendations 9. Annexes – additional supporting evidence as relevant	Within 16 weeks of contract starting
Presentation to Management Group (and others) to discuss draft findings, and further dissemination activities/outputs as proposed in the communications / dissemination plan	Within 18 weeks of contract starting
Final Report Final report should take into account comments on the draft report from DFID	Within 19 weeks of contract starting (ideally on 30 June 2014; this is a target date and alternative proposed dates will be considered)

Evaluation Outputs

The Evaluation Team will produce the following outputs:

- Inception Report
- Draft Final Report
- Presentation to Management Group and others
- Final report (30 – 50 pages with a maximum 3 page Executive Summary) that incorporates feedback obtained on the draft report
- Appendices with details on the methodology, data collection etc.
- A "policy brief" summarising the main findings of the evaluation for circulation to stakeholders, or other learning documents / events as proposed in the communications and dissemination plan

Skills and Qualifications of Evaluation Team

The essential competencies and experience that the Evaluation Team will need to deliver the work are:

- Extensive knowledge of evaluation methods and techniques, incl. thorough understanding of the methodology and design of experimental impact evaluations;

- Strong qualitative and quantitative research skills;
- Good knowledge and understanding of evaluation uptake (how evaluation and research can influence policy and practice) and/or political economy analysis and its relation with evidence
- Good understanding of value for money, and some experience in analysing the costs and benefits of research and/or evaluations
- Good understanding of health policy in the international development context
- Strong analysis, report writing and communication skills

Desirable competencies and experience are:

- Experience in the Zambian and/or Ugandan context
- Understanding of or experience in evaluations related to the Zambian and/or Ugandan political context
- Good knowledge of gender analysis

Expressions of Interest (EoI) from suitably qualified individuals, organisations and consortia are equally welcome. We would welcome bids from teams including evaluators from Zambia and Uganda, though this is not a requirement.

Evaluation Management Arrangements and Stakeholder Involvement

The evaluation will be overseen by a Management Group. This group will be responsible for approving the evaluation outputs and commenting on draft reports. The Group will include the following DFID staff: David Rinnert – lead/day-to-day point of contact for all technical issues (Evaluation Adviser, d-rinnert@dfid.gov.uk; Evaluation Department; East Kilbride); Cormac Quinn (Evaluation Adviser, DFID Zambia; Lusaka), Jonas Heirman (Evaluation Adviser, Evaluation Department; East Kilbride). DFID's Evaluation Department will be responsible for a management response to the evaluations recommendations, and for their implementation. Where relevant, recommendations from the evaluation will be forwarded to other stakeholders (CHAI, partner ministries) for their consideration. The DFID management group will have unlimited access to the material produced by the supplier.

Beyond this group, relevant other stakeholders including CHAI and Zambian/Ugandan partners will be involved early on in the evaluation process. Specifically, the relevant drafts outputs (e.g. inception report) of the evaluation should be circulated to and discussed with relevant stakeholders from CHAI and ministries in both countries (Ministries of Health, Ministry of Child and Maternal Health in Zambia). Furthermore, DFID also plans on including at least one external stakeholder (e.g. from another donor or an academic institution) with extensive relevant experience in the evaluation process for quality assurance and comments on key evaluation outputs.

Liaison will include regular meetings with DFID and one or more presentations by the evaluators. Up to two key meetings/presentations will take place in either DFID East Kilbride or DFID Whitehall, but the evaluation team is expected to use video or audio-conferencing for other/regular meetings with DFID. The tenderer is expected to budget for attendance of all core members at a minimum of two meetings in DFID East Kilbride or DFID Whitehall. The tenderers are encouraged to include one trip to Lusaka, Zambia (and possibly to Kampala, Uganda, but this seems less important as 4 out of the 5 3DE evaluation were implemented in Zambia) in their budget for data collection. Cormac Quinn (DFID Zambia) will be available for (a) meeting(s) and logistical support in Lusaka, should the tenderers plan on travelling there.

Duty of care

The Supplier is responsible for the safety and well-being of their Personnel (as defined in Section 2 of the Contract) and Third Parties affected by their activities under this contract, including appropriate security arrangements. They will also be responsible for the provision of suitable security arrangements for their domestic and business property.

DFID will share available information with the Supplier on security status and developments in-country where appropriate. DFID will provide the following:

- A copy of the DFID visitor notes (and a further copy each time these are updated), which the Supplier may use to brief their Personnel on arrival.

The Supplier is responsible for ensuring appropriate safety and security briefings for all of their Personnel working under this contract and ensuring that their Personnel register and receive briefing as outlined above. Travel advice is also available on the FCO website and the Supplier must ensure they (and their Personnel) are up to date with the latest position.

Tenderers must develop their Tender on the basis of being fully responsible for Duty of Care in line with the details provided above and the initial risk assessment matrix developed by DFID. They must confirm in their Tender that:

- They fully accept responsibility for Security and Duty of Care.
- They understand the potential risks and have the knowledge and experience to develop an effective risk plan.
- They have the capability to manage their Duty of Care responsibilities throughout the life of the contract.

If you are unwilling or unable to accept responsibility for Security and Duty of Care as detailed above, your Tender will be viewed as non-compliant and excluded from further evaluation.

Acceptance of responsibility must be supported with evidence of capability and DFID reserves the right to clarify any aspect of this evidence. In providing evidence Tenderers should consider the following questions:

- Have you completed an initial assessment of potential risks that demonstrates your knowledge and understanding, and are you satisfied that you understand the risk management implications (not solely relying on information provided by DFID)?
- Have you prepared an outline plan that you consider appropriate to manage these risks at this stage (or will you do so if you are awarded the contract) and are you confident/comfortable that you can implement this effectively?
- Have you ensured or will you ensure that your staff are appropriately trained (including specialist training where required) before they are deployed and will you ensure that on-going training is provided where necessary?
- Have you an appropriate mechanism in place to monitor risk on a live / on-going basis (or will you put one in place if you are awarded the contract)?
- Have you ensured or will you ensure that your staff are provided with and have access to suitable equipment and will you ensure that this is reviewed and provided on an on-going basis?
- Have you appropriate systems in place to manage an emergency / incident if one arises?

Budget

The allocated budget for this evaluation is max. £100,000 (incl. VAT, travel and all expenses). Tenderers are expected to prepare a budget detailing planned expenses. Value for money will be a key criterion in selection and the final budget will be agreed with the successful supplier.

Documentation / References

CHAI 3DE Business Case

CHAI Business Case and Submission-5.doc

CHAI 3DE Annual Review 2012/13

2013 completed CHAI Annual Review

CHAI 3DE Annual Review 2013/14

CHAI 3DE Annual Review 20132014.do

CHAI 3DE Logframe (Revised in Jan 2014)

Copy of 2014-01-21 3DE Logical Framewo

Annex B List of key informants

Table 4 List of Key Informants

	Name	Role	Organisation
1	Benjamin Chibuye	Programme Manager 3DE/ Program Management of 3DE Zambia.	CHAI
2	Jeff Grosz	CHAI Country Director	CHAI
3	Elizabeth McCarthy	3DE PI CHAI Global Applied Analytics Team/ PI of 3DE grant.	CHAI
4	Sarah Moberley	Senior Technical Advisor	CHAI
5	Alex Ogwal	Former 3DE Program manager/ Malaria program manager	CHAI
6	Tom Pellens	Former 3DE global project manager	CHAI
7	Marta Prescott	Senior technical adviser	CHAI
8	CJ Schellack	Uganda 3DE country manager/ CHAI 3DE manager	CHAI
9	Jan-Willem Van Den Broek	Country director Zambia	CHAI
10	Alison Connor	Senior Manager	IDinsight
11	Jeremy Fisher	Director of Finance/ Former 3DE program manager	IDinsight
12	Daniel Gastfriend	Idinsight associate	IDinsight
13	Buddy Shah	CEO	IDinsight
14	Esther Hsu Wang	CEO and founding member	IDinsight
15	Paul Wang	Founding partner. Led evaluation component of 3 3DE evaluations (EID, MK, ITN).	IDinsight
16	Barbara Asiire	MoH Officer for Paediatric and Adolescent HIV care	MoH Uganda
17	Peter Elyanu	Former National Coordinator for Paediatric and Adolescent HIV care/ Study PI	MoH Uganda
18	Ivan Lukabwe	MoH Officer for M&E of Paediatric and Adolescent HIV care.	MoH Uganda
19	Hon. Dr. Chitalu Chilufya	Deputy Minister	MoH Zambia
20	Dr Elizabeth Chizema	Director; Disease, Surveillance, Control and Research	MoH Zambia
21	Dr Mulakwo Kamuliwo	Deputy Director, National Malaria Control Centre (NMCC)/ Head of NMCC – involved in ITN scale-up.	MoH Zambia
22	Ndhlovu Ketty	Insecticide Treated Nets Principal Officer, National Malaria Control Center/ Co-investigator for ITN.	MoH Zambia
23	Mr Chikuta Mbewe	Deputy Director Pharmaceutical Services, Ministry of Health	MoH Zambia
24	Victor Mukonka	Senior Lecturer, Copperbelt University/Research Fellow, University College Dublin and Former Director of Public Health and Research Department	MoH Zambia
25	Sandra Sakala	Senior Research and Surveillance Officer	MoH Zambia
26	Mr George Kadimba	Lusaka District Pharmacist, Ministry of Community Development Mother and Child Health/ co-investigator for decongestion.	MCDMCH Zambia

	Name	Role	Organisation
27	Dr Penelope Kalesha	Deputy Director Child Health (MCDMCH), EPI National Professional Officer – Routine Immunisation (WHO Zambia). Co-investigator for EID/EPI National Program Officer	MCDMCH Zambia
28	Dr Mary Nambao	Maternal Health Specialist	MCDMCH Zambia
29	Dr. Carolyn Phiri	Director Mother and Child Health, Ministry of Community Development Mother and Child Health/ Co-investigator for MK.	MCDMCH Zambia
30	Dr Vincent Kanyamuna	Senior Planner	Ministry of Finance, Zambia
31	Joseph Musonda	Acting Principal Planner	Ministry of Finance, Zambia
32	Uzoamaka Gilpin	Health Advisor	DFID
33	Anna Henttinen	Evaluation department. Worked on first Annual Review	DFID
34	Robinah Lukwago	Health Adviser for DFID in Uganda	DFID
35	Cormac Quinn	Evaluation and Results Adviser	DFID
36	David Rinnert	Evaluation adviser. Worked on second Annual Review	DFID
37	Bethany Freeman	Director of Research Operations	Centre for Infectious Disease Research in Zambia (CIDRZ)
38	Margaret P. Kasaro	Clinical Scientist	Centre for Infectious Disease Research in Zambia (CIDRZ)
39	Helen Mulenga Bwalya / Mpande Mukumbwa-Mwenechanya	Head Pharmaceutical Services Department, Centre for Infectious Disease Research in Zambia (CIDRZ) / POPART Program Manager, Centre for Infectious Disease Research in Zambia (CIDRZ)/ Co-investigator for EID.	Centre for Infectious Disease Research in Zambia (CIDRZ)
40	David Rider Smith	Adviser to Prime Minister's Office, Uganda (2007-12)	Prime Minister's Office, Uganda
41	Dr. Penelope Kalesha	Deputy Director Child Health, Ministry of Community Development Mother and Child Health); EPI National Program Officer, WHO Zambia	WHO
42	Dr Godfrey Biemba	Executive Director	Zambia Centre for Applied Health Research and Development (ZCAHRD)
43	Dr Lastone Chitembo	HIV & AIDS specialist / Health and Nutrition	UNICEF
44	Miranda Mhere	Member of EPI Technical Working Group	World Vision
45	Esther Bouma	Attaché - Manager Health and Social Sector EU Delegation to the Republic of Zambia and COMESA	European Union

	Name	Role	Organisation
46	Hon Munji Habeenzu	MP and Member of health committee	Member of Parliament

Annex C Evaluation framework

The evaluation questions are structured around the three main questions listed in the ToR of this assignment and linked to the OECD DAC criteria for evaluation.

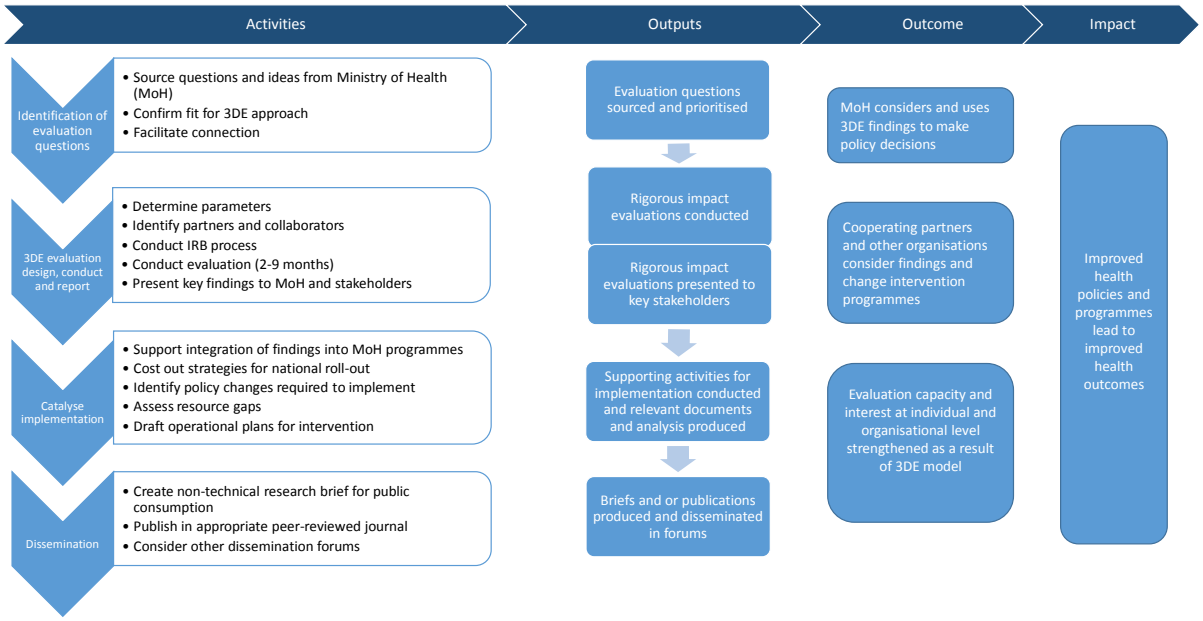
Table 5 Evaluation questions

Main evaluation objectives	OECD DAC Criteria	Research questions	Analytical approach	Data collection tools/methods
Refine and develop ToC and test underlying assumptions and casual mechanisms	Relevance, effectiveness and impact	• Do the main assumptions of the innovative (demand-driven and rapid) 3DE evidence model hold in practice, especially at the output and outcome level (e.g. has the demand-driven character of the pilot led to increased evidence uptake)? • What is the nature of evidence gap and how is this related to impact evaluations? Are there other ways to more strategically, effectively and more efficiently fill this gap? • What was the process for engaging with different stakeholders in sourcing of evaluation questions, conducting evaluations and in dissemination of findings and results? • How were the findings and results of the evaluations communicated to non-technical audience? And in what events or during which period? • What is the process for policy formulation and resource allocations and how are these linked to use and uptake of evidence? • What are the institutional processes for use and uptake of evidence in policy formulation and resource allocation? And what are the incentives around this? • What is the capacity for understanding and engaging with sourcing of evaluation questions, design and implementation of evaluations and interpretation and use of findings at individual and organisational level? • What is the unit within the Ministries responsible for generation and management of evidence (including evaluations) and how do these units interact with the rest of the Ministry? How did these units interact with the 3DE pilot? • Are there any evidence of any evaluations conducted in past 2-3 years being used by policy makers and the health Ministries in the two countries? Are the 3DE evaluations seen differently or have had different outcomes? • How have the donor agencies and organisations used the findings of the evaluations and how have these affected or changed their strategies and intervention programmes and projects? • What are the reasons for the differential outcomes of the 3DE pilot in Zambia and Uganda in practice?	Revised theory of change and systematic testing of assumptions drawing on contribution analysis Political Economy Analysis	Literature review, document analysis and KIIs

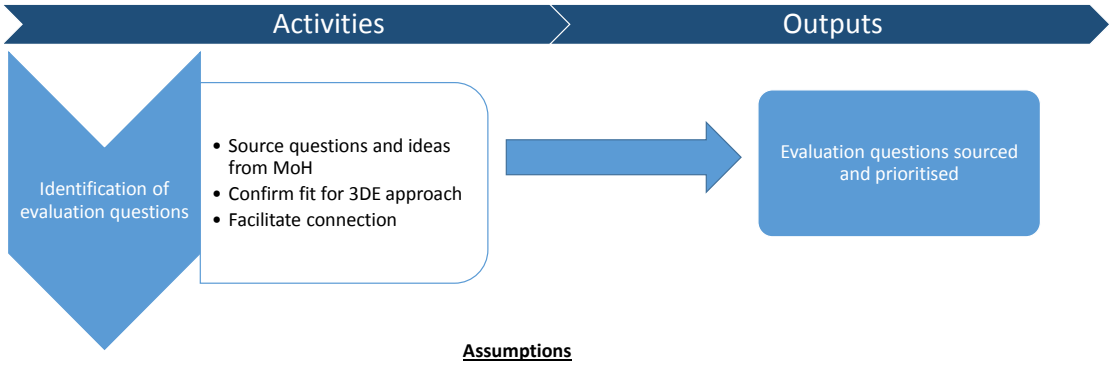
Assessing quality of 3DE pilot products	Relevance, effectiveness and efficiency	• Are the 3DE pilot evaluations rigorous and in accordance with international quality standards (e.g. OECD DAC quality standards)? • What is the effect of the 3DE modality (e.g. demand driven and rapid data collection) on the quality of the data generated? Is the data sufficiently rigorous and credible? • How well did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?	Development and application of a series of questions and comparison with international standards drawing on of SEQAS and ePact quality assurance guidelines	Document review and expert opinion

Annex D Revised Theory of Change

Overall ToC



Identification of evaluation questions



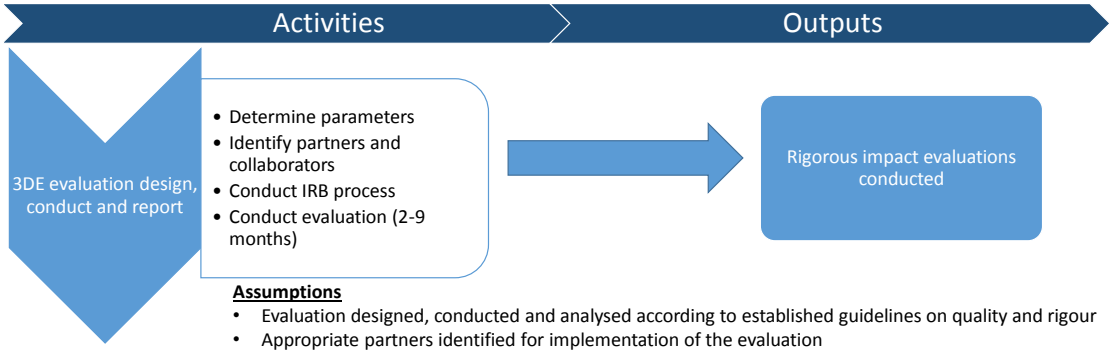
Contextual factor

- There is sufficient demand in MoH for evidence and in particular evaluations

Assumptions

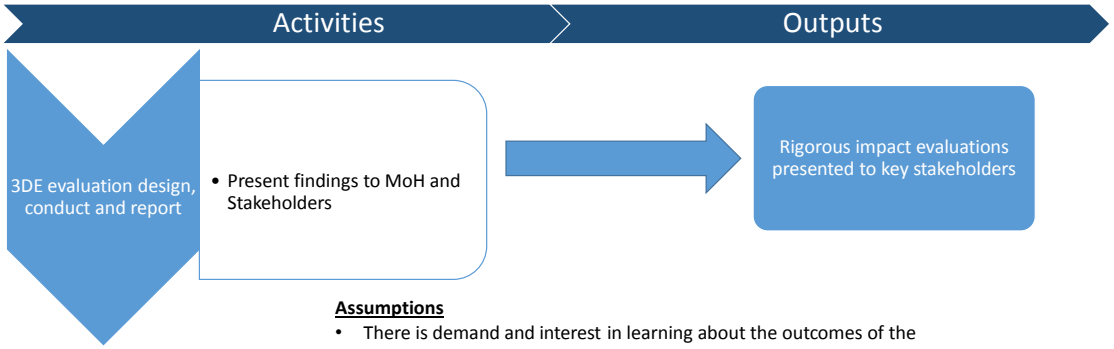
- There is interest from Ministry to engage with sourcing of questions
- The Ministry has a number of questions it needs answers to (ideally situated in a wider research/evaluation policy)
- Good relations are established with appropriate sections within MoH
- Appropriate officials or units are identified and are part of question sourcing and prioritisation
- The Ministry is involved in developing, weighting and applying criteria for prioritisation
- Enough questions which meet the criteria exist
- The Ministry recognises chosen questions as relevant and high-priority

3DE Evaluation design, conduct and report



Assumptions

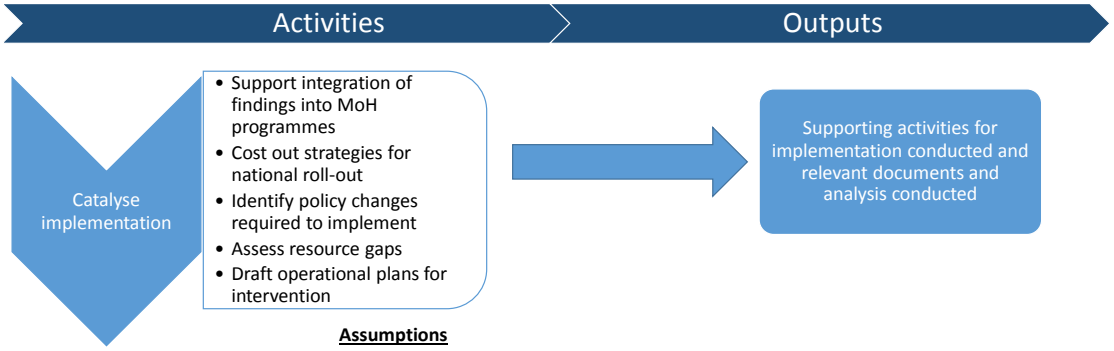
- Evaluation designed, conducted and analysed according to established guidelines on quality and rigour
- Appropriate partners identified for implementation of the evaluation
- Funding and implementing partners are willing to implement programme in accordance with evaluation design and protocol
- Evaluation goes through ethical board and is approved in a timely manner
- The evaluation is conducted in a timely way to meet policy deadlines or windows of opportunity for scale-up, if applicable



Assumptions

- There is demand and interest in learning about the outcomes of the evaluation by the main stakeholders
- Findings are presented to the right stakeholders
- The key stakeholders are present at the dissemination events organised and or fully informed and briefed
- The findings are presented in a accessible manner and tailored to the different audiences
- Presentations are robust, clear and accurately reflect data findings, strengths and limitations

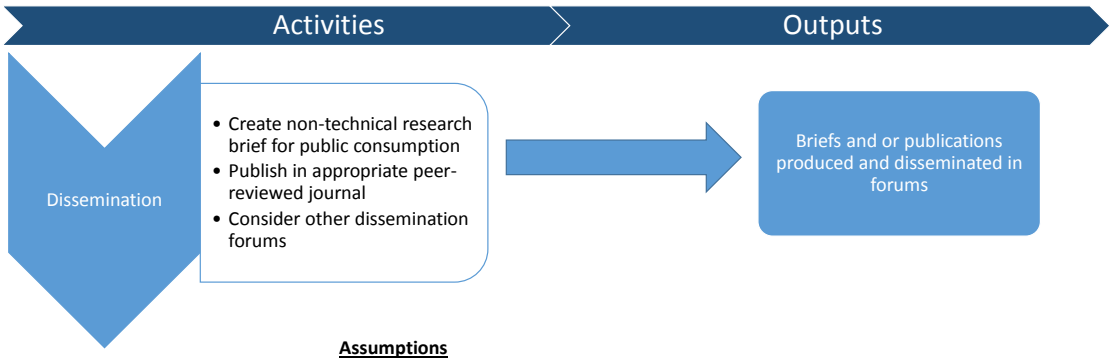
Overall Catalyse implementation



Assumptions

- The results have clear policy (or operational, if relevant) implications
- There are staff who are able to technically engage with the evaluations
- The ministry accepts the findings of the evaluation (takes ownership of findings)
- The ministry requires and requests for support (accepts proposed support)
- The support is provided on time and in accordance with the needs and requirements of the ministry
- The 3DE team is made up of individuals with the right technical skills and thematic areas of interest
- 3DE team engages with the right individuals and stakeholders within the ministry

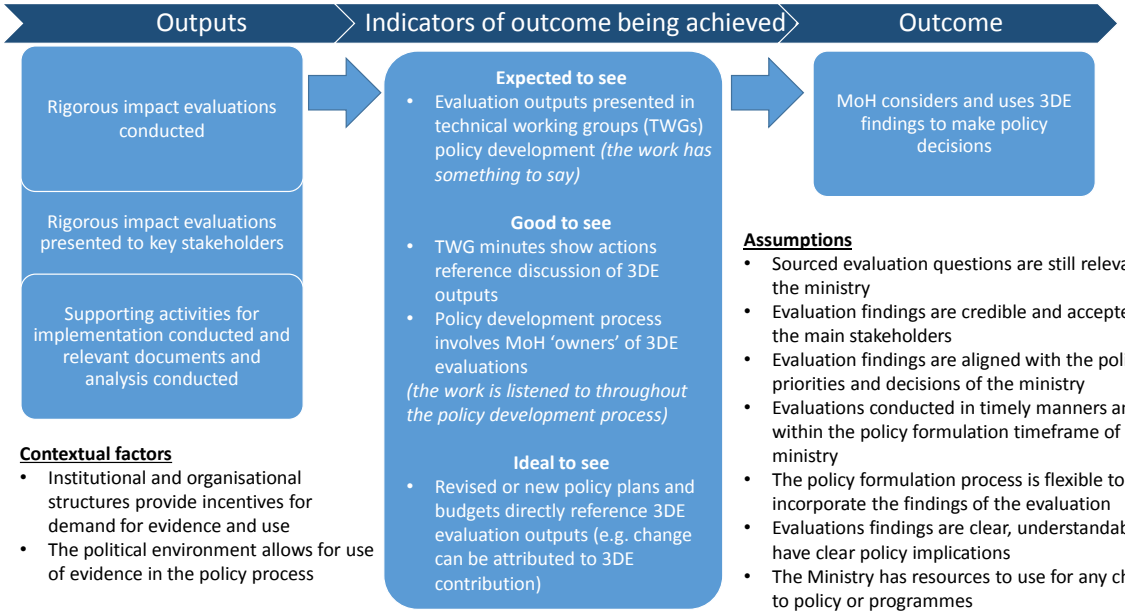
Dissemination

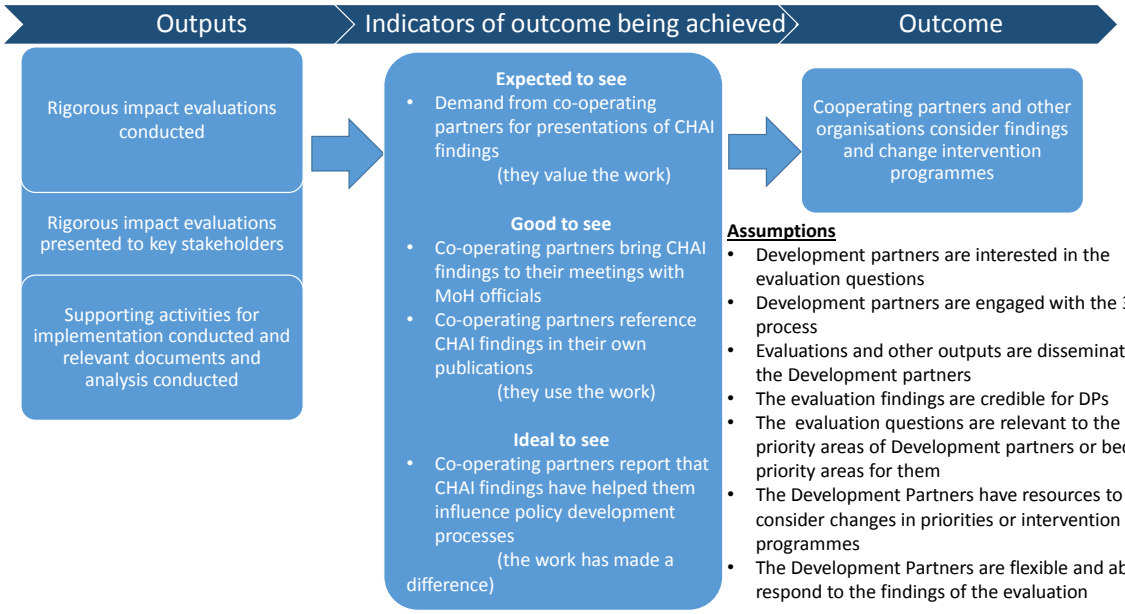


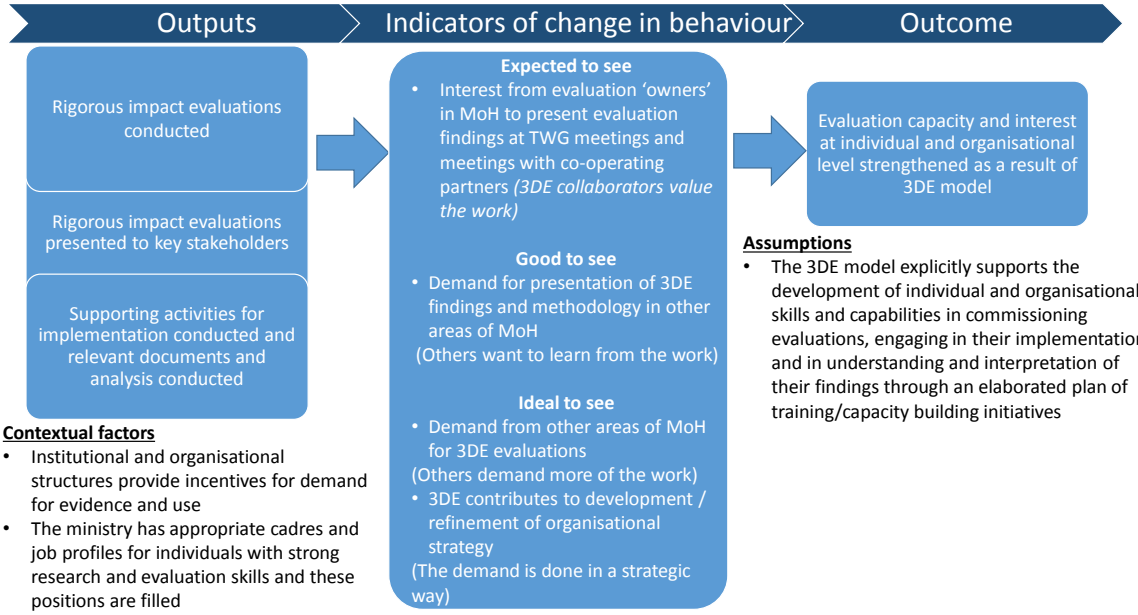
Assumptions

- The need for publication of non-technical research briefs identified
- Appropriate journal identified, papers produced and accepted by targeted journals
- Appropriate dissemination forums identified and 3DE team invited to attend

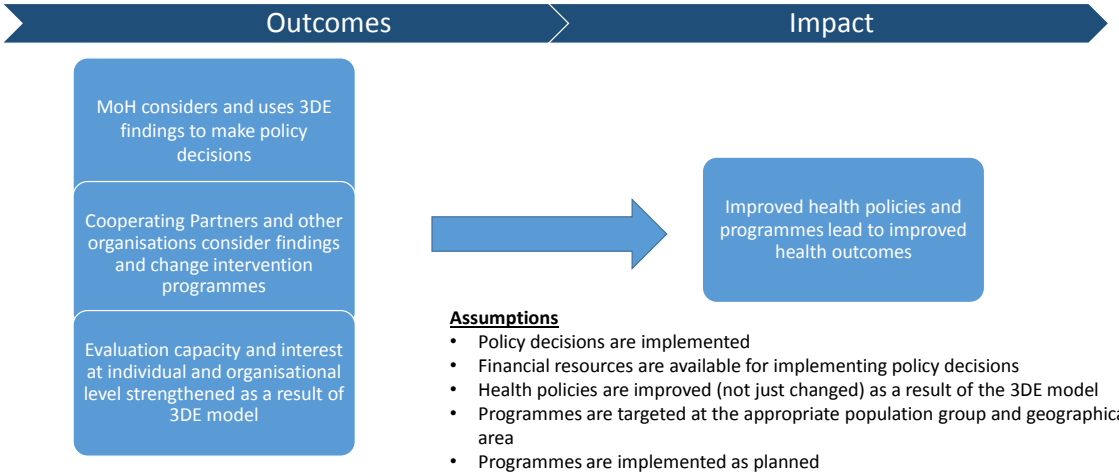
Outcomes


Outputs
Indicators of outcome being achieved
Outcome
Rigorous impact evaluations conducted
Rigorous impact evaluations presented to key stakeholders
Supporting activities for implementation conducted and relevant documents and analysis conducted
Expected to see
• Evaluation outputs presented in technical working groups (TWGs) policy development (the work has something to say)
Good to see
• TWG minutes show actions reference discussion of 3DE outputs
• Policy development process involves MoH 'owners' of 3DE evaluations
(the work is listened to throughout the policy development process)
Ideal to see
• Revised or new policy plans and budgets directly reference 3DE evaluation outputs (e.g. change can be attributed to 3DE contribution)
MoH considers and uses 3DE findings to make policy decisions
Contextual factors
• Institutional and organisational structures provide incentives for demand for evidence and use
• The political environment allows for use of evidence in the policy process
Assumptions
• Sourced evaluation questions are still relevant to the ministry
• Evaluation findings are credible and accepted by the main stakeholders
• Evaluation findings are aligned with the political priorities and decisions of the ministry
• Evaluations conducted in timely manners and within the policy formulation timeframe of the ministry
• The policy formulation process is flexible to incorporate the findings of the evaluation
• Evaluations findings are clear, understandable and have clear policy implications
• The Ministry has resources to use for any changes to policy or programmes


Outputs
Indicators of outcome being achieved
Outcome
Rigorous impact evaluations conducted
Rigorous impact evaluations presented to key stakeholders
Supporting activities for implementation conducted and relevant documents and analysis conducted
Expected to see
• Demand from co-operating partners for presentations of CHAI findings
(they value the work)
Good to see
• Co-operating partners bring CHAI findings to their meetings with MoH officials
• Co-operating partners reference CHAI findings in their own publications
(they use the work)
Ideal to see
• Co-operating partners report that CHAI findings have helped them influence policy development processes
(the work has made a difference)
Cooperating partners and other organisations consider findings and change intervention programmes
Assumptions
• Development partners are interested in the evaluation questions
• Development partners are engaged with the 3DE process
• Evaluations and other outputs are disseminated to the Development partners
• The evaluation findings are credible for DPs
• The evaluation questions are relevant to the priority areas of Development partners or become priority areas for them
• The Development Partners have resources to consider changes in priorities or intervention programmes
• The Development Partners are flexible and able to respond to the findings of the evaluation


Outputs
Indicators of change in behaviour
Outcome
Rigorous impact evaluations conducted
Rigorous impact evaluations presented to key stakeholders
Supporting activities for implementation conducted and relevant documents and analysis conducted
Expected to see
• Interest from evaluation 'owners' in MoH to present evaluation findings at TWG meetings and meetings with co-operating partners (3DE collaborators value the work)
Good to see
• Demand for presentation of 3DE findings and methodology in other areas of MoH
(Others want to learn from the work)
Ideal to see
• Demand from other areas of MoH for 3DE evaluations
(Others demand more of the work)
• 3DE contributes to development / refinement of organisational strategy
(The demand is done in a strategic way)
Evaluation capacity and interest at individual and organisational level strengthened as a result of 3DE model
Contextual factors
• Institutional and organisational structures provide incentives for demand for evidence and use
• The ministry has appropriate cadres and job profiles for individuals with strong research and evaluation skills and these positions are filled
Assumptions
• The 3DE model explicitly supports the development of individual and organisational skills and capabilities in commissioning evaluations, engaging in their implementation, and in understanding and interpretation of their findings through an elaborated plan of training/capacity building initiatives

Impact


Outcomes
Impact
MoH considers and uses 3DE findings to make policy decisions
Cooperating Partners and other organisations consider findings and change intervention programmes
Evaluation capacity and interest at individual and organisational level strengthened as a result of 3DE model
Improved health policies and programmes lead to improved health outcomes
Assumptions
• Policy decisions are implemented
• Financial resources are available for implementing policy decisions
• Health policies are improved (not just changed) as a result of the 3DE model
• Programmes are targeted at the appropriate population group and geographical area
• Programmes are implemented as planned
• Target populations respond to and utilise the programmes or interventions
• Predicted gains in terms of efficiency, equity and outcomes are realised (and any unintended negative consequences are limited)
• Where gains are made, such as increased efficiency, those are retained and reinvested in the health sector

Annex E Assessment of quality of 3DE evaluations

This Annex provides a detailed assessment of the quality of each 3DE evaluation against a list of criteria.

E.1 Mama Kits evaluation

Appropriateness of evaluation questions

Low facility delivery rates in Zambia are identified as a central issue which there is a clear need to address. A gap in evidence on the effectiveness of non-monetary incentives to tackle this problem is also identified. A Theory of Change was provided by 3DE in response to peer review panel comments, which outlines some of the demand-side constraints to facility delivery that Mama Kits are expected to alleviate.

To provide a stronger justification for the intervention, it would still be useful if more background or evidence could be given to show why Mama Kits are thought to be an effective strategy in the potential scale-up region in Zambia. For example, the ethics protocol identifies three possible 'delays' to facility delivery: delay in the decision to seek care, delay in reaching care and delay in receiving adequate healthcare. If delay in reaching care is the primary constraint in Zambia, due to poor transport or long distances to health facilities, then there may be only a weak case for Mama Kits.

Robustness of evaluation design

The following are the main findings of the quality assessment around the robustness of design:

1. The sample frame does not perfectly match the evaluation question.

The evaluation aims as stated in the introduction are framed in terms of assessing intervention effects on institutional delivery rates. However the evaluation design only tests for the causal effect of Mama Kits among the women who visited a study facility for ANC, rather than all pregnant women. If women who attend ANC visits respond differently to Mama Kits than women who do not attend ANC, the evaluation design will not capture the effect of Mama Kits on institutional delivery rates in the population overall. This concern is mitigated by the fact that according to 3DE's response to peer review comments, over 90% of Zambian women do attend at least one ANC visit.

2. The primary outcome measured is suitable for the short time horizon, but it would have also been interesting to measure longer term outcomes.

The study focuses on an intermediate outcome (institutional delivery) which could plausibly be expected to respond to the intervention over a period of months. A longer evaluation period would have added value to the study, since it is possible that the impact of Mama Kits may vary over time. On the one hand Mama Kits may have sizeable long term effects if they contribute to a gradual shifting of established social norms and practices in favour of institutional delivery. Conversely, the effect may weaken over time, if kits lose their value as an incentive, or if SMAGS reduce their effort, for example.

3. The study area is small, causing findings to have limited generalisability to other contexts.

The evaluation is only carried out in 2 districts and the sample is not representative of any wider population in Zambia.

Conduct and reporting of evaluation

The following are the main conclusions of the quality assessment around conduct and reporting:

1. Based on available evidence, the data that was collected was high quality.

Back check surveys to validate delivery data did not uncover any inconsistencies. Moreover qualitative data collection tools provided to us suggest that focus groups were carried out well and covered sensible topics. However the technical report doesn't describe in detail in what way administrative data was physically collected by evaluation staff and if there was any oversight to this process. We don't have enough information to confirm that this was done well.

2. The results of power calculations are poorly presented.

The statement of how many women were intended to make up the final sample is ambiguous. This makes it difficult to confirm that the final sample size was as large as intended.

3. The quality of the presentation of analysis and results is variable.

Many of the variables for primary analysis listed in Table 1 are poorly defined. Tables 2 and 3 do not provide enough information to demonstrate that the randomisation produced treatment and control groups with similar characteristics before the intervention was rolled out. However the main analytical results are well presented and a good description of study limitations is included.

4. Little interpretation is given for the results.

The qualitative findings provide some clues as to which constraints to facility delivery Mama Kits may have reduced in this study area, but the focus group discussions are not well integrated with the quantitative results. The implicit reason is that Mama Kits encouraged facility delivery because they alleviated a concern among mothers that they needed to bring certain items with them to health facilities in order to deliver. Given the limited generalisability of the experimental findings the qualitative results are critical to help understand how and why Mama Kits had the observed effect in that study area. This is necessary to assess the likely effects of scaling up the programme to other areas, and to support the claim made in the closing section of the technical report that the results are generalizable to other rural African settings.

Table 6 Detailed assessment of quality of Mama Kits evaluation

Category	Proposed questions	Comments
Planning and context	1.1 How relevant are the evaluation questions to the priority questions of the Ministry?	The technical report provides a good argument for why promoting institutional delivery may be considered a policy priority in Zambia and the current lack of rigorous evidence on the efficacy of non-monetary incentive strategies.
Introduction	2.1 Is the evaluation question(s) written simply and clearly?	The evaluation aims are stated in the background section of the technical report. The technical report should emphasise somewhere that the evaluation will only assess the impact of Mama Kits on institutional delivery rates among the population of women who have attended first visit ANC, as opposed to all pregnant women.
	2.2 Are the evaluation questions suitable given the short duration of the evaluation period?	Yes. The study does not attempt to identify impacts on high level outcomes such as maternal and child health, which could not plausibly be expected to change in a detectable way over a limited time horizon. It instead focuses on an intermediate outcome (institutional delivery).
	2.3 Is there an adequate description of the intervention to be evaluated (this should include detail on the intervention's target groups, timescale, geographical coverage, anticipated impact, outcomes and outputs, intervention logic and/or theory of change)?	Yes the intervention is well described and no key information is missing. A stronger justification for the Mama Kits intervention could be given. The technical report doesn't provide evidence around which barriers to facility delivery women in the potential scale up region in Zambia face, and how Mama Kits are thought to address them. There are potentially many constraints to facility delivery and Mama Kits would not be expected to be a good response in all cases. For example, the ethics protocol identifies 3 'delays' to facility delivery: delay in the decision to seek care, delay in reaching care and delay in receiving adequate healthcare. If delay in reaching care is a strong constraint in Zambia, due to poor transport or long distances to health facilities, than delay in the decision to seek care, then there may be only a weak case for Mama Kits. The Theory of Change provided by 3DE in their responses to peer review comments is helpful in understanding some of the intervention logic and would be useful to include in the technical report too.
	2.4 Is there a discussion of other programmes or interventions that may also affect impact, outcome and output indicators?	None mentioned.
Method	3.1 Is a RCT the most appropriate method to answer the evaluation question	An RCT is in theory a reasonable choice. The evaluation considers a policy relevant question which is sufficiently in equipoise (according to the literature review) to justify a randomised design. Undue harm for the control group is not anticipated, as the intervention is an incentive design only which doesn't prevent control group women from accessing any services. However given the limited resources available to carry out the evaluation it is not clear that an RCT was in fact the best choice in this case. The study could only cover a small sample in a restricted geographical area, so the findings have weak external validity to other regions in Zambia and cannot stand alone as a meaningful input to a policy decision. An in-depth qualitative study is required alongside these results to explore key mechanisms and contextual factors, which help understand whether similar findings might obtain in other areas. We therefore feel that there is only a modest case for an RCT to answer these evaluation questions. It is not clear that the findings were more useful to policy than what could have been generated from broader research.
	3.2 Is the unit of randomisation appropriate?	Yes. Randomising at the health facility level rather than the individual level lowers the risk of spillovers or contamination that could jeopardise the results. Randomisation at the individual level would not in any case have been suitable, since it is not likely to be feasible to expect health facility staff to distribute Mama Kits to some patients and not others.

Category	Proposed questions	Comments
	3.3 Did the randomisation produce treatment and control groups that were similar at baseline?	Balance between treatment and control groups is presented in Tables 2 and 3. The information given is inadequate to be able to judge whether treatment and control groups were truly balanced at baseline since the tables are missing P values associated with a t-test of the difference in means between treatment and control groups. This is required in order to demonstrate that the differences in means are not statistically significant. The dates associated with the data used to construct each mean should also be presented to confirm that the characteristics are really from the pre-intervention period.
	3.4 Are issues related to spillover effects/externalities (untreated individuals are affected by the treatment) considered and dealt with appropriately?	There is a limited risk of spillovers associated with this evaluation design, given the short evaluation duration and the fact that the intervention was rolled out at the health facility level. Facility level randomisation reduces the chance that women can be influenced by the behaviour of others, since it places more geographical distance between women assigned to the two treatment groups than would be the case if kits were randomised to individuals. However, it is possible that women attending control group facilities for ANC are influenced to choose a facility delivery by their observation of the behaviour of treatment group women. According to the Theory of Change, a key mechanism through which Mama Kits are expected to work is that women interpret the fact that they are being offered an incentive as a signal of the quality of care that they can expect to receive at the facility. To the extent that the provision of mama kits to treatment group women is visible to control group women, this mechanism might be expected to influence the behaviour of both and would cause the effects of the intervention to be understated. This is not a major concern if the findings still reveal a large enough effect to justify a scale-up of the intervention.
	3.5 Are issues related to imperfect compliance (people in treatment group not being treated, or people in control group being treated) considered and dealt with appropriately?	Non-compliance does not appear to have been an issue in this evaluation. The extent to which women attended different facilities for ANC as for delivery is investigated, and only 6 women are found to have delivered in a treatment facility after receiving ANC from a control group facility. A second possibility that is not explored is that treatment group women might share the contents of Mama Kits with control group women, and this affects the decision of control group women over whether to deliver in a facility themselves. The qualitative findings indicate that a key barrier to institutional delivery is the belief that women need to come to a health facility equipped with certain items. If control group women are able to obtain kit contents from treatment group women this barrier may be removed. However, as before this concern is likely to be minimal due to the short evaluation period and the randomisation of the intervention at the facility rather than individual level. Even if present, this mechanism would serve to underestimate the effect of the intervention on treated women. Again, this is less problematic than overstating the effect of the intervention if the aim is to identify a policy-relevant effect.
	3.6 Are local and national contextual factors that could affect the evaluation considered?	Relevant contextual factors that might influence the effect of the intervention include facility capacity for delivery, attitudes towards maternal healthcare, the ease of access to health facilities for rural women and any major political, economic or climate events that may have changed outcomes in the region over the evaluation period. The qualitative findings pick up on some of these factors, and this is useful to help interpret the quantitative results. For example they indicate that women appreciate the benefits of delivering in health facilities, and that it is not lack of knowledge which causes them not to attend.
	3.7 Is the timing of the data collection appropriate given the timing of the intervention?	Yes, the timing of collecting information on women attending ANC between October and August 2013 was appropriate to identify women with an expected delivery date in the required range.

Category	Proposed questions	Comments
	3.8 Can the findings be expected to have reasonable external validity to inform a wider policy or programmatic decision?	As noted, weak external validity is a major limitation with this study. The sample is drawn from only 2 districts in Zambia (out of 72). Without further discussion of how the study areas and sample facilities chosen compare with the rest of Zambia it is not plausible to assume that similar results would obtain if the intervention was rolled out elsewhere.
	3.9 Were there any trade-offs in design due to the relatively short time frame of the evaluation, and if so what were they?	A benefit of a longer evaluation period would have been the ability to assess the sustainability of the intervention and see whether the magnitude of effect changes over time. There are several reasons why the effect of Mama Kits may be expected to be different in the long term than in the short term. On the one hand, the marginal value of distributing additional Mama Kits may decrease over time as community availability of kits increases, and women can share contents amongst themselves to obtain the items without needing to deliver at a facility. This would cause the intervention to become less effective over time. On the other hand, if the provision of Mama Kits contributes to a gradual process of shifting social norms around delivery and maternal health the intervention may have larger effects in the longer term. Given these differing possibilities it would have been interesting to be able to evaluate the Mama Kits over a longer time span. A longer evaluation period would also enable the evaluation to explore the effects of the intervention on the final welfare impacts for children and mothers of increased institutional delivery.
	3.10 Are there other significant methodological limitations (not mentioned above)?	Methodological limitations are generally well documented in the technical report and ethical proposal. These include the fact that the findings only relate to the population of women who attend ANC, the reliance on administrative data of possibly unknown quality, being unable to perfectly match records across ANC and delivery registers, and the fact that the evaluation did not test for the effects of different contents or values of Mama Kit. The decision not to test different kinds of Mama Kit package in separate treatment arms is justifiable in this case. Adding more treatment arms to the study would have reduced the power to detect statistically significant policy relevant effect sizes. The kit package that was tested was cheap relative to other Mama Kits that have been used in the past, and the fact that positive effects were still observed provides useful evidence that a low value kit package can be effective. Although tables comparing the characteristics of treatment and control groups before the intervention are not clear (as discussed in 3.3), we understand from our key informant interviews that tests were performed to confirm sample balance at baseline. The risk of bias caused by spillover and non-compliance is also low for this study design, therefore internal validity is expected to be acceptable for this evaluation despite the reasonably small sample size of 15 clusters per treatment arm.
Data	4.1 Were the most suitable data sources selected? If primary data collection was undertaken, were the most suitable data collection methods selected?	The evaluation made a good use of administrative data from facility records to identify women who attended ANC, women who delivered in health facilities and women who received Mama Kits. A weakness of relying on these data sources is that it proved difficult to match women across ANC and delivery registers. This is acknowledged in the technical report and is an acceptable trade off in exchange for not having to conduct full scale primary data collection. It would be useful if the technical report could give more detail on how data was actually collected from these administrative sources by the evaluation team. Suitable quality assurance was conducted to validate delivery data and confirm matching of women across facility registers. The study is not able to validate the ANC data. The ethical proposal suggests that a household survey for all women in selected villages (2 per facility) was originally planned, which would have helped to verify data on ANC visits and estimate the extent of home births.
	4.2 Have the sampling frame and the sampling populations been correctly defined?	The sampling frame does not perfectly match the evaluation question as articulated, since it only tests for the effect of Mama Kits on institutional delivery rates among the population of women who attend ANC, rather than all pregnant women. The evaluation question does not indicate that that the study is focused on this restricted sample.

Category	Proposed questions	Comments
	4.3 Is the sampling procedure rigorous and appropriate? (What is the sample representative of?)	The sample selection is not representative of any wider population in Zambia because it is too small and covers a narrow geographical area. Representativeness was not taken into account in the sampling procedure. Districts were selected according to how appropriate they were considered to be for rolling out this intervention, and health facilities were chosen on the basis of having a low ratio of deliveries to ANC visits. The information provided on the characteristics of sampled health facilities in the peer review responses document is helpful to indicate how the sample may compare with Zambia more generally. It would be useful to include a short discussion on this in the technical report as this helps understand whether the results might be generalizable to potential scale up areas.
	4.4 If primary data collection was undertaken, are survey instruments well-constructed (clear, robust skip patterns, relevant answer codes) and are they adequately described?	The surveys for the home spot check survey are presented in the ethics protocol and appear well-constructed. It is interesting that the spot check survey contains several additional questions other than those needed to confirm a facility delivery, such as the reasons why facility delivery was chosen and satisfaction with the service provided. The findings from these questions are not given anywhere in the technical report so it is not clear why they were included (or perhaps the questionnaire that was ultimately used was shortened). We have seen a topic list, and moderator guide for focus groups, and received information from stakeholder interviews to suggest that focus groups were carried out well. The technical report would benefit from describing in more detail the content and conduct of focus groups, as little information is given.
	4.5 Are secondary data sources adequately described and has their quality been checked to determine the data is reliable?	Delivery data was verified using spot check surveys. ANC data wasn't checked, but validating delivery data was of higher priority to ensure that the effects of the Mama Kits are not overestimated, so this was a reasonable decision given the constraints to resources or budget.
	4.6 Were sample sizes adequate?	The results of the power calculations are unclear so we cannot establish whether the sample size was adequate. See 4.7
	4.7 Were sample size calculations done well and are they presented?	The power calculations are poorly presented. It is not clear what is meant by the overall sample size of '200 women per facility per quarter'. This seems to imply an intended sample of 6000 women 'per quarter' (200*30 facilities), which over the course of the period of 11 months over which ANC records were gathered (October 2012 – August 2013) would mean an intended sample size of nearly 18,000 women. The final sample contained only 2219 women. Since Optimal Design was used it would have been useful to see the output that was produced.
	4.8 If primary data collection was undertaken, are any biases from non-response discussed?	Yes this is well discussed. The only primary data collection undertaken in the evaluation is the home spot check surveys and qualitative focus group interviews. Bias from non-response is less of an issue for the purposes of quality assurance and qualitative data collection than it would be for a quantitative survey. The technical report nevertheless indicates the number of women who could not be located for the spot check survey and the reasons why.
Data Collection	5.1 If primary data collection was undertaken, were data collected in an appropriate and respectful manner, taking into account cultural, ethical, as determined from the protocols submitted for ethical approval, the field	Yes, appropriate ethical concerns have been taken into account. The sample spot check surveys and focus group materials provided to us indicate that informed consent was obtained from respondents and that surveys are carried out confidentially.

Category	Proposed questions	Comments
	manual and the characteristics of the data collectors?	
	5.2 If primary data collection was undertaken, were the instruments tested and validated (e.g. pre-testing of questionnaires)?	It is not clear from the technical report whether there was any pre-testing of the home spot check surveys or focus group interviews, or if a pilot was conducted.
	5.3 If primary data collection was undertaken, were the instruments translated and back translated?	We understand from our key informant interviews that back translations were carried out and the consistency between the original English versions and back translations was checked.
	5.4 Were field teams trained to gather data before the start of the intervention? If primary data collection was undertaken, were the field teams trained by the same people who made and tested the survey instruments?	The technical report does not indicate whether evaluation staff were trained to carry out home spot check surveys or collect and enter data from facility registers.
	5.5 Has there been an appropriate level of oversight and data quality assurance in the data collection?	Some oversight of data collection was undertaken. The technical report notes that 10% of home spot check surveys were randomly selected for resurvey to check for interviewer error. It is not clear whether there was any oversight to the process of collecting data from facility administrative records.
Data entry and cleaning	6.1 If a survey was undertaken on paper, was the data double entered and were discrepancies between the two entries systematically resolved by checking the hard copies?	The main data for analysis was not collected on paper. There is no evidence that the home spot check survey was double entered.
	6.2 Was the data cleaning done is a robust, clear and transparent way and does it include both range and consistency checks?	Nothing is mentioned on how data was cleaned and what the main issues were that arose.
Data analysis	7.1 Are primary analysis methods appropriate? If regressions are used, are they correctly specified and are standard errors calculated correctly?	A logistic regression model is an appropriate choice for the analysis. It would be useful to see the regression written out in equation form and to define an odds ratio, to help readers understand the analysis and interpret the results. Huber White cluster-robust standard errors are used to account for the effects of correlation between outcomes in the same cluster. However, the asymptotic assumptions that underpin the definition of the standard errors are may not be justified when the number of clusters is small (15 in each treatment arm in this case). The weak performance of cluster-robust standard errors when there are few clusters is discussed in Bertrand, Duflo and Mullainathan (2004)[19]. In view of this, the regression model may have been improved by further adjustment to compensate for the small number of clusters, such as by implementing the wild cluster bootstrap procedure outlined in Cameron, Gelbach and Miller (2008)[20].
	7.2 Are the key indicators clearly defined including how they are	A weakness of the technical report is that it doesn't define key indicators very clearly.

[19] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-In-Differences Estimates?*." *The Quarterly journal of economics* 119.1 (2004): 249-275.
[20] Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. "Bootstrap-based improvements for inference with clustered errors." *Review of Economics and Statistics* 90.3 (2008): 414-427.

Category	Proposed questions	Comments
	calculated, and are they suitable to measure the outcomes of interest?	The primary outcome variable and data from which it is calculated are not clearly defined in the technical report, though it appears to be a binary variable indicating whether a women who attended at least one ANC visit delivered in a health facility. The definition of variables in Table 1 is also extremely unclear. For example, does the variable '#deliveries/#pregnancies' capture the proportion of pregnancies that resulted in a live birth or the proportion of pregnancies that resulted in delivery in a health facility? In either case it is not clear where the data to calculate previous pregnancies would be gathered from. Does the number of past ANC visits refer only to her current pregnancy, or to all pregnancies? Gravida should also be defined. No justification is provided for the covariates chosen. It appears that some of the covariates are intended to capture the effects of other potential barriers to institutional delivery that are not addressed by Mama Kits (for example, % of staff that are male, # trained staff). The findings in relation to some of these variables would be interesting to include in the final discussion. The role of some other covariates is less clear. For example, is the proportion of ANC visits that are outreach intended to be a proxy for health facility quality, or for the dispersion and level of isolation of some households in the facility catchment area?
	7.3 Have sampling weights been used correctly?	Sampling weights are not needed in this case since the analysis included all pregnant women who attended ANC and had an estimated delivery date; there was no individual level sampling.
	7.4 Are departures from the full sample size (non-response) explained and has any non-random attrition been identified and dealt with correctly?	The sample of women is based on the ANC record, so one possible loss of sample occurs if women attend ANC but are not recorded as such. It is not clear to what extent this may have been a problem. A second issue is if women who delivered in health facilities are missing from facility registers. In this case, there is no loss of sample, but these women would be classified as having had a home birth. This would cause the incidence of facility delivery to be underestimated, which is less problematic than would be the case of the effect was overestimated and the intervention was wrongly attributed with having a significant positive effect.
	7.5 Have any differences between treatment and control groups at baseline been accounted for in measures of impact?	The technical report asserts that there are no differences at baseline to account for, the summary statistics table does not provide enough information to confirm this.
	7.6 Is the analysis disaggregated to show outcomes and impact on different groups and sexes? Did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?	The data is not disaggregated to show results on different subgroups. In this case, gender effects do not apply since the entire sample is women. The possibilities for subgroup analysis are limited by the nature of the administrative data collected, which does not include socioeconomic variables that may have formed a basis for disaggregated analysis. Women who did not attend ANC or who attend outreach services are not included in the sample for this evaluation. To the extent that these women may be expected to be on average poorer than women who attend ANC at health facilities, the evaluation design is unlikely to capture the effects of the intervention on the most vulnerable groups.
Reporting	8.1 Are quantitative results presented systematically and logically?	The main results table (Table 4) is fairly well presented. It shows three specifications of the regression model and indicates which variables were found to be statistically significant in each case. It would benefit from the inclusion of sample sizes.
	8.2 How clear are the links between data, interpretation and conclusions?	The claim that Mama Kits improved institutional delivery rates in this setting is supported by the quantitative results presented. Qualitative findings could be better integrated with the quantitative results to provide more interpretation for the findings. The implicit conclusion is that Mama Kits were effective because they alleviated a concern among mothers that they needed to bring certain items with them to health facilities in order to deliver. However, Mama Kits did not address other identified barriers, including the distance to the health facility, early delivery and lack of female staff at facilities. Since women were not found to lack knowledge of the health benefits of institutional delivery, this suggests that the 'signal of quality' mechanism outlined in

Category	Proposed questions	Comments
		the Theory of Change did not obtain in this setting. The report would have been strengthened by picking up on some of these issues, and linking the findings from focus groups directly to the intervention.
	8.3 Were negative/discrepant results addressed or ignored?	There were no negative results.
	8.4 Are the final recommendations and conclusions plausible?	The conclusion of the technical report suggests that the Mama Kits can be used to improve rural facility delivery rates in Africa, which is implausible on the basis of these results. The discussion of results should be very clear that the findings presented only obtain in a very selective population and cannot be easily generalised to a wider setting.
	8.5 Have alternative explanations been explored and discounted?	Very little interpretation for the findings is given, since as discussed the qualitative findings are not developed in depth. These findings could have been used to probe some of the mechanisms initially outlined in the Theory of Change provided in the peer review comments document.

E.2 ITN evaluation

Appropriateness of evaluation questions

The research questions are relevant and well timed to provide evidence on optimal distribution strategies ahead of the planned allocation of six million ITNs across Zambia.

Robustness of evaluation design

The following are the main findings of the quality assessment around the robustness of design:

1. Aside from external validity concerns, the evaluation design is appropriate for the budget.

Ideally, to answer the research the evaluation would have included different treatment arms to separately test the effects of a community fixed point distribution strategy with a door to door method. The second best option of comparing the findings with an existing evaluation testing the door to door method is appropriate given the budget constraints. An efficient use of CHW time was made by combining data collection activities with routine ITN registration, distribution and hang-up activities.

2. Aside from external validity concerns, the design of the RCT was appropriately tailored for the time horizon and budget.

The primary outcomes measured were suitable for a short time horizon. Ideally the two follow-up surveys would have been carried out at the same time of year in different years to prevent seasonality effects from influencing the results. But since this was not possible within the time horizon, the study still benefited from having a second follow-up within the same year, as a check on the possibility that the findings from the first follow-up were primarily driven by seasonal effects.

3. The level of randomisation (to individuals) is appropriate to obtain high statistical power.

4. The study area is small, causing findings to have limited generalisability to other contexts.

The evaluation is only carried out in 3 communities in 1 district.

Conduct and reporting of evaluations

The following are the main conclusions of the quality assessment around conduct and reporting:

1. The evaluation appears to have been well conducted.

Data collection for the primary analysis was well executed, following suitable ethical protocols and used well designed survey instruments. Data was quality assured to a good degree through back check surveys and interviewer observation to confirm information on ITN installation and use. Few inconsistencies were found, indicating that data was of high quality.

2. We do not have enough information to comment on the quality of fieldwork.

We have not had access to information around the training that CHWs received, whether a pilot was undertaken, what the field team structure was, how much oversight CHWs had and the process by which CHWs physically located households. It would be useful if the technical report could have described some of this in more detail. However we do understand that there were no reported issues around CHWs being unable to find households or visiting the wrong households, so this is some evidence that fieldwork processes were sound.

3. The sample size appears to have been smaller than what was intended.

The final sample of households appears to have been smaller than the intended sample size, according to the power calculations. It appears that not enough villages were selected into the sample.

4. The quality of the presentation of results is variable

Most results tables are presented well, with no key information missing. However, not enough information is provided to determine whether the different treatment groups in the study had similar characteristics before the intervention was rolled out. This is necessary to confirm that the final results were caused by the effects of the intervention alone. The graphical results on ITN retention are not separated by treatment group, and supporting numbers from analysis are not given. This is not sufficient to verify the conclusion that there was no significant difference in retention rates between households assigned to hang up and non-hang up groups. Crucially, details of the recently evaluated door to door intervention against which the community fixed point strategy is tested are not provided, including what the findings were and where it was conducted. It is essential to describe this evaluation since the findings form the basis for the main conclusions of the study.

5. The quality of the interpretation of results is variable

On the whole, lessons from the evaluation are well presented and reflect the analytical results. However, although noted in various places in the report, the final discussion of results reports the finding that delaying hang-up visits is more cost effective than scheduling them sooner. This is misleading since the results do not show that there is a case to be made for having CHW hang up visits at all.

Table 7 Detailed assessment of quality of ITN evaluation

Category	Proposed questions	Comments
Planning and context	1.1 How relevant are the evaluations questions to the priority questions of the Ministry? (explored as part of validation of the ToC)	Understanding effective ways to ensure high ITN use and retention is a priority question for the MoH ahead of the planned distribution of 6 million nets across the country.
Introduction	2.1 Is the evaluation question(s) written simply and clearly?	Yes, the objectives of the evaluation are clearly set out in the technical report.
	2.2 Are the evaluation questions suitable given the short duration of the evaluation period?	Yes, the evaluation questions focus on short and medium term outcomes and so are within the scope of the short evaluation period. See 3.8 for description of the limitation of conducting the evaluation over a limited time period.
	2.3 Is there an adequate description of the intervention to be evaluated (this should include detail on the intervention's target groups, timescale, geographical coverage, anticipated impact, outcomes and outputs, intervention logic and/or theory of change)?	The intervention is well described. The only part that is not fully clear is whether there were any restrictions on the households that were eligible to be registered to receive ITNs. Was this all households in the selected neighbourhood zone, or just some of them? There is no Theory of Change or intervention logic model, which might have been useful to help justify the intervention. The technical report explains why fixed point distribution may be preferable to a door to door strategy in certain contexts, but doesn't fully describe what the barriers are to ITN ownership and use in the first place. Are ITNs often found to be in short supply, or is the problem that there is a lack of demand to buy nets?
	2.4 Is there a discussion of other programmes or interventions that may also affect impact, outcome and output indicators?	None are mentioned.
Method	3.1 Is a RCT the most appropriate method to answer the evaluation question	An RCT is an appropriate approach to understand the effects of the intervention. There is a high priority research question to address, for which the technical report currently indicates there is a gap in evidence. There is also no reason to expect that any of the study groups will be unduly disadvantages by the evaluation, since access to ITNs is not withheld for any group. However, it is not clear that the value of an RCT is justified in this case where the resources available to carry out the study are so limited. The study was only able to roll out to three neighbourhood zones in a single district. This means that the findings have weak external validity. It also means that the study was not able to

Category	Proposed questions	Comments
		directly compare the effects of a community fixed point distribution strategy with a door to door strategy. This would have required the inclusion of more clusters in the study, since it is likely to be unpractical to overlay the use of the two strategies within the same area.
	3.2 Is the unit of randomisation appropriate?	Yes, individual randomisation is a good choice in this case to maximise the power of the study to detect statistically significant changes in outcomes. There is no strong justification for randomisation at a higher level since the risk of spillovers and contamination is small for this intervention, and there are unlikely to be any tensions within communities caused by variation in hang-up visits among individual households.
	3.3 Did the randomisation produce treatment and control groups that were similar at baseline?	The results of tests for statistically significant differences in characteristics between the different treatment groups of the study are not presented, so it is not possible to determine whether the groups were similar at the start of the evaluation period. Presenting means and standard deviations is not enough to demonstrate balance. The table should also record the characteristics across all treatment groups, rather than just groups 1-4 (hang – up visits) against group 5 (no hang up). Even though the five groups are not separated out in some of the analysis, this check is still necessary to confirm that the randomisation was successful.
	3.4 Are issues related to spillover effects/externalities (untreated individuals are affected by the treatment) considered and dealt with appropriately?	A spillover in this context could occur if households assigned to the no hang-up or delayed hang up groups observe hang-up activities happening for their neighbours in other treatment groups, and are influenced by this to hang up their own nets. This is discussed in the Ethics Protocol, which argues that the likelihood is minimised by the encouragement provided to all household to hang up their nets as soon as they receive them, and the fact that hang-up visits will be spread out over a number of weeks. The technical report also notes that large distances between the rural households in the sample make it less likely that households are influenced by their neighbours. The Ethics Protocol makes the valid point that any spillovers which do occur in this study are likely to result in increased installation, use and retention of nets among the no-hang up or delayed hang up groups. This would cause the effects of CHW hang up visits being underestimated, which is less problematic from a policy perspective than erroneously measuring a false positive effect.
	3.5 Are issues related to imperfect compliance (people in treatment group not being treated, or people in control group being treated) considered and dealt with appropriately?	The risk that CHWs visit the wrong households for hang-up visits, or visit at the wrong times is low if there is a good system in place for identifying households. It would be useful if the technical report could describe in more detail the process by which households were located by CHWs and whether there were any cases of the wrong households being visited.
	3.6 Are local and national contextual factors that could	Yes the technical report contains a good discussion of how seasonal effects might have influenced the findings in the two follow-up surveys.

Category	Proposed questions	Comments
	affect the evaluation considered?	
	3.7 Is the timing of the data collection appropriate given the timing of the intervention?	The timing of data collection is reasonable given the constraints posed by the short evaluation time frame. It was a good choice to carry out two follow-up surveys to investigate whether findings changed over time.
	3.8 Can the findings be expected to have reasonable external validity to inform a wider policy or programmatic decision?	Low external validity is the most important limitation with this evaluation, which is only carried out in three neighbourhood zones of a single district of Zambia. As noted in the Ethics Protocol, this makes it very unlikely that the study population is perfectly statistically representative of the potential scale up population of Zambia. The technical report notes at least two reasons why the findings in Rufunsa may be more positive than could be expected elsewhere – firstly because of large distances between households and villages, which pose challenges for the alternative door to door approach, and secondly because previous malaria sensitisation activities had been conducted in the region. It would be useful if the technical report could provide a fuller discussion of the situations in which a community fixed point distribution strategy is likely to be effective in comparison to an alternative approach, so that the findings can be applied more easily to practical policy questions.
	3.9 Were there any trade-offs in design due to the relatively short time frame of the evaluation, and if so what were they?	Limitations of the short evaluation duration include the fact that it is not possible to measure the longer term effects of community fixed point distribution after 6 months, and that the two follow up surveys couldn't be carried out at the same time of year in different years to account for seasonal effects in analysis. Measuring longer term effects of the intervention is potentially of interest as there are reasons to expect that the effects of the intervention may change over time. On the one hand, the effects may be strengthened over time if a culture of ITN use becomes ingrained and households are influenced by one another to maintain ITN retention. On the other hand, the effects of the encouragement to use and hang up nets may weaken over time as it is difficult for households to directly observe the benefits of taking preventive healthcare measures.
	3.10 Are there other significant methodological limitations (not mentioned above)?	Overall, the design of the evaluation was good given the resource constraints faced. The methodological limitations that there were could generally only have been mitigated if the RCT had a larger scope. A weakness associated with the small sample area covered by the evaluation, in addition to low external validity, is that the design did not allow for a comparison to be made between a door-to-door strategy and the community fixed point strategy. This is mentioned in the technical report. As above, it is likely that the evaluation design would have needed expand to cover more areas and randomise at a higher level in order to compare these two strategies. Given that this was infeasible, the evaluation made a reasonable effort to compare the results with other studies that examined door to door methods. There is a low risk of attrition from the sample in this evaluation, where households included in the initial surveys are not found at follow-up because they have moved away from the area or couldn't be located. This is

Category	Proposed questions	Comments
		unlikely given the short evaluation duration and the fact that the evaluation works with CHWs within the local community to identify households.
Data	4.1 Were the most suitable data sources selected? If primary data collection was undertaken were the most suitable data collection methods selected?	The choice of data was good, and made efficient use of CHW time by combining the distribution and hang-up activities with data collection. It was sensible to ask that CHWs observe nets hanging where possible to confirm information on net use and retention. As outlined, it would be useful to have described in more detail how households in each treatment group were located by CHWs so that the efficacy of this process could be appraised.
	4.2 Have the sampling frame and sampling populations been correctly defined?	Yes the sampling frame and sample populations have been correctly defined. As discussed, clarity over whether all households were eligible to register for ITN nets, and therefore be included in the sample, would be helpful.
	4.3 Is the sampling procedure rigorous and appropriate? (What is the sample representative of?)	A representative sample of neighbourhood zones was not selected – sites were instead chosen purposefully so that there would be variation in characteristics such as distance to the nearest Rural Health Centre. Overall the study area is too small and specific to be considered representative of any wider population.
	4.4. If primary data collection was undertaken, are the survey instruments well-constructed (clear, robust skip patterns, relevant answer codes) and are they adequately described?	Survey instruments are shown in the Ethics Protocol and are well constructed.
	4.5 Are secondary data sources adequately described and has their quality been checked to determine the data is reliable?	The only secondary data used in the study was the registration data from the NMCC which was used to define the sample of households. It is difficult to see how the quality of this data could have been checked in a practical way.
	4.6 Were sample sizes adequate?	The sample sizes appear to be insufficient. According to the sample size calculations given in the Ethics Protocol, 662 households were intended to make up the sample, but only 560 were ultimately included. This could suggests that not enough villages were selected into the sample.
	4.7 Were sample size calculations done well and are they presented?	Sample size calculations are given in the Ethics Protocol. It would be useful also to report these as an annex to the technical report. The calculations are not fully clear. The outcome variable on which the main calculation is based is not clearly defined, it is described as an 'ITN-level' outcome when what is meant is household level. According to the power calculations, stratification is done by village, yet in the technical report the stratification level is given as the registering CHW – if the two definitions coincide this should be made clear. Finally, it could have been

Category	Proposed questions	Comments
		helpful to present the graphs produced by the Optimal Design software. We are unsure how such precise sample size numbers could have been generated by this method. The power calculations should have ideally been supported by some discussion or evidence of whether the proposed sample sizes are likely to have been feasible in the study areas under consideration. This could have alerted the evaluation team to the possibility of not being able to find enough households in the proposed study area to make up the sample.
	4.8 If primary data collection was undertaken, are any biases from non-response discussed?	Of the original 562 households who were pre-registered, only 2 moved away prior to the distribution of ITNs. However only 514 and 502 households were included in the 7-11 week and 5-6 month surveys respectively, suggesting some degree of non-response. The expected reasons for this loss of sample are not outlined, but would be helpful to understand whether there is a potential risk of bias. Missing values for specific questions are reported under Tables 4 and 5 in analysis, and are minimal.
Data Collection	5.1 If primary data collection was undertaken, were data collected in an appropriate and respectful manner, taking into account cultural, ethical, as determined from the protocols submitted for ethical approval, the field manual and the characteristics of the data collectors?	Yes, data collection appears to have been carried out with an appropriate degree of ethical oversight. The Ethics Protocol indicates that informed consent was sought from households prior to surveying, and that precautions were taken to ensure confidentiality.
	5.2 If primary data collection was undertaken, were the instruments tested and validated (e.g. pre-testing of questionnaires)?	It is not clear if instruments were pre-tested or if a pilot was undertaken.
	5.3 If primary data collection was undertaken, were the instruments translated and back translated?	We understand from our conversations with some Key Informants that back translations were made and consistency between the original English language surveys and back translations checked for errors.
	5.4 Were the field teams trained to gather the required data before the start of the intervention? If primary data collection was undertaken, were field teams trained by the same	The Ethics Protocol reports that CHWs were trained in data collection methods and hang-up techniques, but few details are provided on what the nature of this was.

Category	Proposed questions	Comments
	people who made and tested the survey instruments?	
	5.5 Has there been an appropriate level of oversight and data quality assurance in the data collection?	Data was quality assured to a good degree through back check interviews and interviewer observation to confirm information on ITN installation and use. Oversight to the distribution process by CHWs evaluation staff and community leaders was also arranged to help ensure that nets were assigned to the correct households. The Technical Report notes very few cases of inconsistency between the original surveys and back checks, indicating that the data was of high quality. It would be useful if a more thorough description of fieldwork processes could be given somewhere, to show what the team structure was and if further oversight or supervision was given to CHWs as part of the normal course of their operations.
Data entry and cleaning	6.1 If a survey was undertaken on paper, were the data double entered and were discrepancies between the two entries systematically resolved by checking the hard copies?	The evaluation used a mixture of data collection methods. The short survey completed by CHWs during their visits to hang up nets were completed on paper and follow up surveys were carried out using Open Data Kit application on mobile phones. There is no full description of data entry processes for the paper based surveys, and if further consistency checks or oversight to data collection was carried out beyond the back check visits.
	6.2 Was the data cleaning done is a robust, clear and transparent way and does it include both range and consistency checks?	There is no discussion of data cleaning, for example how inconsistent of impossible results were treated in analysis or whether surveys were designed to prevent such responses on the ground.
Data analysis	7.1 Are primary analysis methods appropriate? If regressions are used, are they correctly specified and are standard errors calculated correctly?	Regressions appear to be well specified. It is a good choice to base the main analysis on two primary 'treatment' groups (groups 1-4 and group 5) rather than separating out all treatment groups since sample sizes in individual groups are relatively small and could cause results to be measured imprecisely. The chosen strategy exploits variation in the date of hang-up across groups 1-4 in combination to assess the effects of the days between the hang up visit and net distribution on self-installation, retention and use of nets. It is also sensible to include the households who attended distribution events but were not pre-registered in group 5 as a robustness check rather than in the main analysis. Including these households in the main analysis would risk interfering with the randomisation process, since households who did not pre-register to receive ITNs may have different characteristics from those who did. The analysis does not involve comparing before and after measures of the same outcome variable. In most cases this would not have been possible, since outcome variables are generally measured in terms of the number of ITNs that were distributed – and there is no baseline measure for this. However, the percentage of sleeping spaces covered could in theory have been measured using a differences in differences specification, i.e comparing changes before and after the intervention in the two treatment groups. This would have required CHWs to add a question on current sleeping space coverage during the data collection that was done at the

Category	Proposed questions	Comments
		community fixed point distribution event. The ability to make use of baseline observations for the same variable would strengthen the statistical power to detect changes in outcomes caused by the intervention. The report does not document which formula for standard errors was used; however since this was not a cluster RCT we assume that a conventional formula was applied.
	7.2 Are the key indicators clearly defined including how they are calculated, and are they suitable to measure the outcomes of interest?	Yes, indicators are generally clearly defined and there is some discussion around the shortcomings of the chosen variables in terms of how well they are able to capture actual ITN behaviours. Some weaknesses in variable definitions are discussed. These issues would generally cause main outcome measures to be underestimated, which is preferable to causing spurious overestimates if policy relevant effects are still detected. For example, the 7-11 week survey was conducted during a busy farming period when some household members may have moved away, taking their nets with them, and also during the rainy season when some households reported taking down ITN nets due to leaking roof or heat. The measure for the %ITNs hanging is also an underestimate if households were distributed more nets than there are sleeping spaces. We believe there is an error in the definition of one of the variables in the presentation of ITN use regressions in Appendix C. The variable 'days between hang up visit and follow-up visit' should be 'days between distribution and follow-up visit'. It is otherwise not defined for households in the no follow-up group, and some assumptions would have to be made in order to specify the regression. This variable is given as 'days between visit and distribution' in the results Tables 3, 4, 10 and 11, which is ambiguous, as 'visit' could refer either to the hang-up visit or follow up survey visits.
	7.3 Have sampling weights been used correctly?	Sampling weights cannot be constructed for this evaluation since the selection of neighbourhood zones was done in a purposeful manner, and all households who registered for ITNs were sampled.
	7.4 Are departures from the full sample size (non-response) explained and has any non-random attrition been identified and dealt with correctly?	There is some discussion of different forms of non-response in the evaluation, for example households leaving the sample area after registration or not being identified for back-check surveys. Regression tables also indicate the number of missing values for each variable. In general minimal issues relating to non-response are reported. As before, it would be useful to detail whether there were any issues around CHWs being unable to locate some households to hand out nets (i.e the 'mop-up' visits for households who missed the main distribution event), hang-up nets or carry out follow up interviews.
	7.5 Have any differences between treatment and control groups at baseline been accounted for in measures of impact?	The technical report states that there are no differences between baseline characteristics of the treatment groups, and therefore no need to adjust analysis accordingly, however the balance tables do not provide enough information to confirm this since significance of differences is not reported.
	7.6 Is the analysis disaggregated to show outcomes and impact on different groups and sexes? Did	According to the Ethics Protocol, data from the district and facility level will provide baseline information that will allow the team to calculate whether hang-up activities had different impacts according to certain characteristics.

Category	Proposed questions	Comments
	the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?	However the technical report does not suggest that differences in effects according to household socio-economic variables were tested.
Reporting	8.1 Are quantitative results presented systematically and logically?	The quantitative results on retention are not well presented. Figure 2, showing ITN retention rates at each follow-up period, should distinguish between households assigned to the hang-up and non hang-up groups as done in Figure 3 (percentage of sleeping spaces covered). This would help to support the key claim made in the text that there was no significant difference in retention between households assigned to the hang-up and non hang-up groups. In addition to the graphical representation, the numbers generated by the analysis that led to this conclusion should also be presented. A key issue is that results of the door-to-door distribution study are not presented anywhere in the technical report. The final recommendations of the report hinge on the claim that the ITN behaviours observed as a result of the community point distribution intervention compare well with these door to door results, so it is a major weakness of the technical report that these findings are not shown. It would also be necessary to describe details of this intervention and its evaluation, including a discussion of how the study area in that evaluation compares with the present case. The regression tables are otherwise well presented and no major supporting evidence is missing,
	8.2 How clear are the links between data, interpretation and conclusions?	Overall lessons from the evaluation are well presented and on the whole reflect the results well. There appears to be some caution in drawing out one of the main lessons of the quantitative results, that there are no long term benefits of having a CHW hang-up visit on the two measures of ITN use after 5-6 months. Although this finding is written up in various places in the technical report, the 'final discussion of results' section does not mention this and reports instead that delaying hang-up visits to at least 10 days is more cost effective than scheduling them sooner. This is misleading since the results actually do not show that there is a case to be made for having CHW hang-up visits at all. We appreciate that the findings on CHW hang-up are likely to be very sensitive to the particular characteristics of different communities, and should not form the basis of a blanket recommendation to apply to all districts. However the discussion should make this point very clear rather than discussing time savings resulting from delayed hang-up as through this were a key recommendation.
	8.2 Were negative/discrepant results addressed or ignored?	As above, the discussion could bring out further the finding that hang-up visits did not lead to higher rates of ITN use in the long term.
	8.4 Are the final recommendations and conclusions plausible?	Yes the conclusions are plausible. Given the specificity of the study area, the technical report would have benefited from a more thorough discussion of the particular characteristics of that area and how these might have made it more or less suitable for a community fixed point distribution approach. More useful conclusions could be generated from the study if the findings from this sample area were clearly linked to its particular features.

Category	Proposed questions	Comments
	8.4 Have alternative explanations been explored and discounted?	The research question is simple and concerns only whether behaviour changed as a result of the intervention, not why. This is acceptable given the scope of the evaluation; there is no need to fully consider different explanations for the findings.

E.3 EID evaluation

Appropriateness of evaluation questions

Poor progress against MoH guidelines on infant and maternal HIV testing rates are highlighted as an important issue in Zambia which there is a need to improve. The interventions under test seem reasonable to address the identified problem, and the idea to align HIV testing and treatment with routine immunisation care points is well justified.

However as before, a stronger justification for the research question could have been given by presenting more evidence on the likely reasons for low HIV testing rates in this setting. There is no Theory of Change for these interventions, which we recognise was not necessarily the responsibility of 3DE to develop. Nonetheless some discussion of whether supply chain issues, or demand side constraints such as lack of knowledge, stigma and fear of HIV testing, are considered to be prevalent in the potential scale-up region in Zambia would have helped to motivate the case for the interventions.

Robustness of evaluation design

The following are the main findings of the quality assessment around the robustness of design:

1. The study design may have been too ambitious given the limited budget.

Including three treatment arms rather than two reduces the power of the evaluation to detect statistically significant policy relevant effects. We agree with the peer review comment that a compelling reason is not provided to justify this choice. The rationale might be that the two intervention types are thought to address different possible constraints to postpartum and infant HIV testing. Alternatively, there may be uncertainty as to whether the Simple intervention would be sufficient to change outcomes. In either case it would be useful if the technical report could outline the basis for the decision, providing evidence where appropriate, otherwise it is not clear that this was a reasonable choice given the limited available budget. 3DEs response to the peer review comments does present evidence of supply side failures in HIV testing from a situation assessment, to justify the Simple Intervention arm. But what are identified as demand side barriers also seem to be caused by failures in processing of patients at the health facility side.

2. The intervention may have been too ambitious for the short evaluation duration.

The Comprehensive intervention involved testing a change in how patients are processed and treated at health facilities. Health systems reform such as this may be more ambitious than interventions involving the delivery of some service to a target population. Evaluation findings highlighted some significant challenges in implementing the intervention, including supply stock-outs and high workloads for facility staff. This indicates that the intervention may have been still undergoing an adjustment period during the evaluation period. The findings may not therefore have accurately reflected its potential effectiveness.

3. The level of randomisation (to health facilities) is suitable for the intervention type.

4. Other evaluation limitations are acknowledged in the Technical Report.

Issues around supply stock-outs, possible ways in which Control group facilities may have been influenced by the intervention and the risk that patients purposefully switch health facilities in response to the intervention allocation are all well described. The study was not powered to detect a small increase in DBS testing rates, which may explain why no effect was ultimately detected. A larger sample may have been required to detect the true impact.

Conduct and reporting of evaluations

The following are the main conclusions of the quality assessment around conduct and reporting:

1. There were some challenges around implementing the intervention, which may have affected the results.

Challenges included high staff workload and frequent stock outs of testing equipment. These issues may have contributed to the finding that the intervention did not lead to a statistically significant increase in DBS testing rates. Data collection was planned and carried out well. The technical report indicates that a pilot was conducted, which included supervisor visits to troubleshoot difficulties, and that evaluation staff were trained in data collection activities. The process of inputting data from facility registers is well described, and a range of quality checks were included to validate the data. Qualitative data collection also appear to have been well done, using sensible question guides and following good ethical protocols.

2. The quality of the presentation of results was generally good.

The study presents detailed results in connection with all aspects of the quantitative analysis. Qualitative results from exit interviews, staff interviews and focus group discussions are also presented in detail. The definition of key outcome indicators is sometimes ambiguous. Outcomes are measured as ‘average values per facility, per month’, and it is not clear what period the average is taken over. Presenting time trend graphs as well as quantitative results is helpful to understand the data. Results tables should have included sample sizes.

3. The results are fairly well discussed and interpreted.

All quantitative findings are written up in the text. A good interpretation for results is given, building on the findings of focus groups and exit interviews. There were some ways in which the interpretation of results could have been developed further. For example, we infer that a key possible reason for the failure of the intervention to cause an increase in DBS testing rates is the persistent supply shortages. However this is not provided as an explanation for this particular finding. Secondly, the reasons why the effects on maternal retests appear not to have been sustained over time are not discussed in detail, which seems crucial to understand since this has clear implications for the efficacy of scaling up the programme.

Table 8 Detailed assessment of quality of EID evaluation

Category	Proposed questions	Comments
Planning and context	1.1 How relevant are the evaluations questions to the priority questions of the Ministry? (explored as part of validation of the ToC)	A good justification for the intervention is given. There is an identified need to improve Zambia's performance against maternal retesting and EID guidelines.
Introduction	2.1 Is the evaluation question(s) written simply and clearly?	Yes the evaluation questions are presented clearly in the technical report.
	2.2 Are the evaluation questions suitable given the short duration of the evaluation period?	The evaluation questions focus on short and medium term outcome indicators, which is appropriate for the short evaluation duration since higher level impacts may take more time to be realised. However the intervention itself could be too ambitious to warrant a rapid evaluation. It does not involve simply implementing a programme, but instead aims to improve health facility systems in a fundamental way. It is likely that this kind of intervention would undergo an adjustment period before routine operations could be established. There is a risk that the evaluation period is too short to capture the full potential of this intervention to alter outcomes in a positive way,
	2.3 Is there an adequate description of the intervention to be evaluated (this should include detail on the intervention's target groups, timescale, geographical coverage, anticipated impact, outcomes and outputs, intervention logic and/or theory of change)?	Yes, the intervention is well described and no key information is missing.
	2.4 Is there a discussion of other programmes or interventions that may also affect impact, outcome and output indicators?	There is no explicit discussion of other interventions being conducted in the study area. The report does note that evaluation districts were purposefully chosen to ensure absence of conflicting research projects.
Method	3.1 Is a RCT the most appropriate method to answer the evaluation question	In the manner in which the evaluation questions are framed, an RCT is appropriate since the questions emphasise identifying a causal impact of the intervention on a range of outcomes. However, if the research questions were not articulated in this way it is possible that a different kind approach may have been suitable. For example, an operational study or process evaluation could also have helped to understand the strengths and challenges around implementing the intervention. Given the difficulties that were ultimately found in relation

Category	Proposed questions	Comments
		to ensuring consistent supplies to facilities, a process oriented study could potentially have added great value to the evaluation by exploring the reasons for this in greater depth.
	3.2 Is the unit of randomisation appropriate?	Yes. The choice to allocate the intervention at the health facility level is appropriate, although the limitations with this evaluation design given in the technical report an in 3.4 and 3.5 are important to keep in mind. It would be infeasible to randomise this intervention to individuals since it involves changes to service delivery which can only be carried out in health facilities as a whole, and therefore individual treatment and control group women would not be well defined. Rolling out the intervention at a higher level than the health facility is also not possible given the relatively small size of the study area for the intervention. There are unlikely to have been enough higher level units to create a well powered study.
	3.3 Did the randomisation produce treatment and control groups that were similar at baseline?	Yes. Randomisation schemes that would have resulted in statistically significant differences in baseline averages for DBS tests, DPT1 doses and first ANC visits were purposefully rejected. Table 2 presents summary statistics for facilities in each of the three treatment arms and confirms that there are no statistically significant differences in any characteristic.
	3.4 Are issues related to spillover effects/externalities (untreated individuals are affected by the treatment) considered and dealt with appropriately?	Yes these issues are well discussed in the technical report. The technical report indicates that spillovers may have had an important influence on the results of the evaluation. Firstly, the supply reinforcement provided to treatment facilities may have boosted district-wide supply stocks, allowing control facilities to receive more retesting kits than they would have otherwise. Tis could have caused the intervention effects on DBS testing to be underestimated, which may be part of the reason why the final results do not show any significant improvement in this outcome. Presumably it would have been difficult for the evaluation to measure changes in District level stocks and allocation decisions, otherwise the effect of this spillover on the findings could have been assessed.
	3.5 Are issues related to imperfect compliance (people in treatment group not being treated, or people in control group being treated) considered and dealt with appropriately?	Possible non-compliance is noted in the technical report as a limitation of the evaluation design. The evaluation did not attempt to estimate the extent to which women may have transferred between facilities, and if there was any substantive evidence of women who would originally have attended control group or simple intervention facilities choosing to attend comprehensive intervention facilities instead. If switching between health facilities was widespread, there is a risk that the results observed at the level of each health facility reflect the underlying characteristics of the population of women attending those facilities. This runs the risk that the evaluation finds that intervention facilities led to higher HIV retest rates, when in fact if the intervention were scaled up to all health facilities in the region the gains would not be so large.
	3.6 Are local and national contextual factors that could affect the evaluation considered?	Yes, the period of national stock out of DBS tests is noted as having an important impact on results, and is included as a covariate in the regression on the number of DBS tests per month.

Category	Proposed questions	Comments
	3.7 Is the timing of the data collection appropriate given the timing of the intervention?	The timing of data collection is appropriate. We note that the baseline period (between January 2012 and July 2013) is much longer than the intervention period (October 2013 – March 2014), but this is related to the intervention period available from which to gather data.
	3.8 Can the findings be expected to have reasonable external validity to inform a wider policy or programmatic decision?	The evaluation is only conducted in 3 districts in one province of Zambia, so external validity is a concern in the absence of a description of how the study area and population compare with the potential scale-up region. The choice of districts was done according to some combination area characteristics (e.g. geographical dispersion, urban/rural characteristics) which may have been done to ensure that some variety was represented, but the basis of the choice is not given.
	3.9 Were there any trade-offs in design due to the relatively short time frame of the evaluation, and if so what were they?	As discussed, the short duration of the evaluation may have been problematic if the routine operations put in place by the intervention took time to become established. The technical report suggests that this was the case since test kit supplies were an issue throughout the intervention period, especially at the start. Exit interview findings also showed that the intervention may not have been well communicated to women. If more time were required for the health systems change introduced by the intervention to become settled, the findings of the evaluation would be misleading of the actual results that the intervention could expect to achieve in the long term.
	3.10 Are there other significant methodological limitations (not mentioned above)?	The technical report includes a well-developed section outlining methodological limitations with the study, which includes potential spillovers affecting Control facilities, the risk of patients switching between health facilities, possible data quality issues and a lack of available data for some potentially interesting outcomes. The peer review panel commented that it is not clear that the benefits gained from having a three arm design was large enough to compensate for the loss of power associated with this. We would agree that the justification for rolling out two separate interventions is not fully provided. If, as suggested by the peer review panel comments, the intention is to separately test the effects of supply side reinforcement alone with the effects of supply side reinforcements plus additional demand side incentives, the reasons why both demand and supply side constraints might be expected to be important barriers in this context should be fully outlined. We are not clear from the response given by 3DE to this comment what additional constraint the comprehensive intervention is seen to alleviate. What 3DE describe as 'demand-side' barriers appear to be more to do with failures in processing from the health facility side. Overall it would be useful if the report could argue more fully why a 3 arm study was chosen instead of a two arm study, drawing on existing evidence or preliminary research to inform the research design.
Data	4.1 Were the most suitable data sources selected? If primary data collection was undertaken, were the most suitable data collection methods selected?	A good use was made of secondary data to determine quantitative results. The process of inputting data from facility registers into Open Data Kit surveys using mobile phone applications is well described. The scope of the data only allows individual level analysis to be performed for a very limited range of variables. The only individual level outcome included in analysis is the age at first DBS test.

Category	Proposed questions	Comments
		A limitation with the qualitative data collection that it does not include the perspectives of women who did not attend U-5 services. Exit surveys are carried out for women after their U-5 appointments, and focus group discussions with mothers who attended static or outreach services, so those who accessed no kind of service were not represented. The technical report notes that attempts were made to include these women but it was not possible to locate them.
	4.2 Have the sampling frame and the sampling populations been correctly defined?	The sampling frame and sample population (all women and their infants attending static services) is correctly defined. A high proportion of possible health facilities for the study were selected into the sample (60 out of 77), meaning that the proportional sampling of facilities would have only had a limited effect on the composition of the final sample. The study would have been strengthened if the initial population of health facilities was larger than 77, but this would have been infeasible under the resources allocated to the evaluation.
	4.3 Is the sampling procedure rigorous and appropriate? (What is the sample representative of?)	The sample is only representative of three districts in one southern province, and does not have wider generalisability to a larger population in Zambia.
	4.4 If primary data collection was undertaken, are the survey instruments well-constructed (clear, robust skip patterns, relevant answer codes) and are they adequately described?	Yes, the instruments for focus group discussions and data verification interviews are presented in the Ethics Protocol and are well constructed and clear.
	4.5 Are secondary data sources adequately described and has their quality been checked to determine the data is reliable?	All data sources used in the evaluation are well described and thorough quality assurance was planned to check the accuracy of secondary data. A limitation of the quality assurance process is that some verification checks are only possible for data collected during the intervention period in treatment areas.
	4.6 Were sample sizes adequate?	The study sample size was designed for ensure no deleterious effects on immunization We do not have enough information to assess whether sample sizes were large enough relative to what was intended. Results tables don't show the sample size that was achieved, and nor does the presentation of sample size calculations indicate the number of observations that was intended. Note that the number of DBS samples, retests and immunisations as reported in the text doesn't correspond to the sample size used for analysis, which used monthly figures. We suspect that sample sizes are too small to enable robust detection of results, since graphs of time trends are not smooth. We also note that the study was not powered to detect small increases in DBS testing rates, which may be part of the reason why the final results did not reveal any increase.

Category	Proposed questions	Comments
	4.7 Were sample size calculations done well and are they presented?	The sample size calculations are included as an annex to the technical report. They are not clear in some respects. It is confusing to report the intra-cluster correlation coefficient, since in this case the primary outcome for analysis is measured at the facility level, which is the same level as the allocation of the intervention. This kind of correlation (which is intended to account for the measurement of repeated monthly measures from the same facility) would be more usually defined as an intertemporal correlation. Figures of 0.5-0.7 would be considered very high for an ICC as it is usually understood. The minimum effect sizes expected also appear to be extremely large, and it would be useful to justify why such large effect sizes were anticipated. The report could perhaps also explain why an alpha of 0.1 is chosen, since a value of 0.05 is more customary. Finally, the actual sample sizes that are intended to achieve these minimum effect sizes are not really clear. The table doesn't show the number of observations of each outcome variable per facility that need to be obtained in the sample.
	4.8 Are any biases from non-response discussed?	The technical report does not discuss whether there were any major issues associated with missing facility records, or interviewers being unable to complete data verification surveys.
Data Collection	5.1 Were data collected in an appropriate and respectful manner, taking into account cultural, ethical, as determined from the protocols submitted for ethical approval, the field manual and the characteristics of the data collectors?	Yes. The technical report notes that confidentially was sought during the conduct of exit surveys, and that informed consent was sought in all cases. Ethical standards adhered to are outlined in the Ethics Protocol.
	5.2 If primary data collection was undertaken, were the instruments tested and validated (e.g. pre-testing of questionnaires)?	There was a six week pilot of the interventions between August and September 2013. We assume that this included testing of the HIV activity sheets and larger U-5 tally sheets. The technical report notes that supervisor visits were scheduled during the pilot period to troubleshoot difficulties, which may have included challenges around data collection. It is not clear whether the data verification interviews and exit interviews were pre-tested.
	5.3 If primary data collection was undertaken, were the instruments translated and back translated?	Yes. We understand from our key informant interviews that back translations of survey instruments were completed, and consistency with the original English was checked.
	5.4 Were the field teams trained to gather the required data before the start of the intervention? If primary data	Yes, training was undertaken in all intervention facilities and appears to have been thorough. It is not indicated whether this training was carried out by the same team who developed the intervention data collection tools.

Category	Proposed questions	Comments
	collection was undertaken, were field teams trained by the same people who made and tested the survey instruments?	
	5.5 Has there been an appropriate level of oversight and data quality assurance in the data collection?	Yes, a good effort was made to quality assure the data sources used for the evaluation. Some checks, for example verification of the average monthly number of maternal retests using HIV activity sheets, were only possible in intervention facilities during the intervention period. It appears that there were some reasonably large discrepancies between the main data sources and verification. For example, the average monthly number of DPT1 doses recorded on facility registers and U-5 tally sheets differed by 10% or more in 52% of cases. The technical report notes that inconsistencies were resolved using established data cleaning rules, but it is not clear what these were.
Data entry and cleaning	6.1 If a survey was undertaken on paper, was the data double entered and were discrepancies between the two entries systematically resolved by checking the hard copies?	Not relevant for this study.
	6.2 Was the data cleaning done is a robust, clear and transparent way and does it include both range and consistency checks?	Data cleaning rules are not documented.
Data analysis	7.1 Are primary analysis methods appropriate? If regressions are used, are they correctly specified and are standard errors calculated correctly?	A differences in differences approach was chosen to analyse the data. This was an appropriate choice, since the summary statistics table does indicate some fairly large differences in mean characteristics between the different treatment arms (though none significant at the 10% level). Standard errors were calculated using bootstrap methods, which was also a suitable choice given the small number of clusters in the study. When the number of clusters is small, classic cluster-robust adjustments to standard errors may not be sufficient to overcome the bias that can be caused by correlation between outcomes within the same cluster. Bootstrapping may in these circumstances be an improvement. Data from the pilot and training period was dropped in order to calculate the effects of the intervention when it was operating as normal. However it might have been worth considering re-running the results with the pilot and training data included as intervention period observations, since Figure 6 suggests that increases to maternal retests were largely realised during this time and actually dropped in both intervention arms after the pilot period was over.
	7.2 Are the key indicators clearly defined including how they are	Yes.

Category	Proposed questions	Comments
	calculated, and are they suitable to measure the outcomes of interest?	
	7.3 Have sampling weights been used correctly?	Sampling weights were not used in the analysis to account for the semi-proportional sampling of the original 60 facilities from a possible 77. However with such a high proportion of facilities selected into the sample this would not have had significant effects on the final analysis. Individual level weights are not required since data from all U-5 visits are included in the sample.
	7.4 Are departures from the full sample size (non-response) explained and has any non-random attrition been identified and dealt with correctly?	The technical report does not mention whether missing data records were a problem for the study.
	7.5 Have any differences between treatment and control groups at baseline been accounted for in measures of impact?	A differences in differences analysis accounts for the effect of unbalanced treatment arms at baseline.
	7.6 Is the analysis disaggregated to show outcomes and impact on different groups and sexes? Did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?	The primary analysis for the study is focused on determining the effects of the intervention on HIV testing rates for women and children. Results are not disaggregated according to other criteria, such as socioeconomic status of women or of the facility catchment area. However it is not clear that this would have added meaningful results to the study, since the small sample sizes used in the study limit the options for further disaggregation, and the study does not have access to rich individual level data.
Reporting	8.1 Are quantitative results presented systematically and logically?	The study presents detailed results in connection with all aspects of analysis. Presenting time trend and descriptive analysis alongside regression results is helpful to understand the data and findings better. However, sample sizes used for analysis are missing from the majority of the results tables.
	8.2 How clear are the links between data, interpretation and conclusions?	All key quantitative findings from the results tables are mentioned in the text, and all statements made in the technical report are supported by data. Qualitative results from exit interviews, staff interviews and focus group discussions are presented in detail. There were some instances in which the interpretation of results could have been developed more in the discussion section. For example, shortages in the supply of DBS testing kits and other supply challenges seem crucial to the findings. This is touched on in the technical report in the treatment of outlier facilities in analysis. However it is not given as a reason for why the intervention did not lead to any increases in DBS testing rates,

Category	Proposed questions	Comments
		even though this seems a highly plausible explanation. In the discussion section this finding is instead attributed to spillover effects and the low power of the study. The reasons for positive results on maternal retests are generally well explained in the discussion section. But the possible reasons why the effects appear not to have been sustained over time are not discussed in detail, other than the suggestion that regular retraining for staff may be required to sustain the results. Fully exploring the possible reasons for this downward trend seems crucial, since it has important implications for the likely success of programme scale up. Finally, the discussion could investigate more thoroughly the reasons for the observed decrease in the average number of DBS tests administered per month in control facilities. The report states that this was largely driven by one facility, but it would be interesting to consider whether other reasons for this could have included a migration of patients from control facilities to intervention facilities to receive tests, or the national supply stock-outs of DBS testing kits affecting un-supported control facilities more than intervention facilities. As noted in the review panel comments, the strength of interpretation of these findings of this evaluation would be improved if a Theory of Change had been developed detailing the binding constraints to EID and maternal retest rates that the two interventions are designed to mitigate.
	8.3 Were negative/discrepant results addressed or ignored?	The study purposefully tests for possible negative results on immunisations. This is valuable to provide support for the claim that the intervention did not have damaging effects on the outcomes it did not target. It does not ignore the finding that there was no increase in DBS testing rates in intervention facilities compared with control. However the finding that the initially positive effects on total postpartum maternal retests decrease considerably over time in both treatment arms is not reiterated in the discussion section.
	8.4 Are the final recommendations and conclusions plausible?	The key conclusion from the evaluation is that reinforcing supplies and patient flow in health facilities can make a positive difference to HIV care for mothers without negatively affecting immunisation rates. It also notes the importance of redressing key supply chain failures and maintaining operational support to the integration of HIV testing activities with under-5 visits to yield maximum benefits from this innovation. These conclusions are all plausible given the findings. As noted, more discussion could be reserved for the reasons why there was no observed increase in DBS tests.
	8.5 Have alternative explanations been explored and discounted?	As noted, although the study does provide several possible reasons for some of its main findings, there are cases where further explanation or consideration of alternative reasons could be given.

E.4 Decongestion evaluation

Appropriateness of evaluation questions

Facility congestion, linked to large increases in the number of patients deemed eligible for ART care, is highlighted as a key concern in the Study Protocol. Stakeholder interviews unanimously indicated that Ministry demand for this evaluation was high. However, it is not in fact clear that this research question will provide useful information in relation to this problem. Initial assessments undertaken in Lusaka by IDinsight suggested that there may be limited scale-up potential for this intervention, since few facilities visited were found to have a need for it according to pre-established criteria.

> *"it wasn't a compelling question that was useful to policymakers without the ability to carry out a longer follow-up– saying after you provide intensive support, is there a change or not? It is not a certainty but we felt that it was pretty likely that changes would occur with someone standing over your shoulder telling you what to do. For this kind of intervention to be useful to policymakers it a longer follow-up period that is important. If you stop follow-up, what will policymakers conclude from that and what can we recommend? The more compelling question is what happens in a 4-5 month window"*

Robustness of evaluation design

The following are the main findings of the quality assessment around the robustness of design:

1. The choice of indicator is not a good proxy for the primary outcome variable.

The main outcome indicator is the proportion of stable patients receiving 3 month refill prescriptions. This is a poor measure of facility congestion unless the stock of patients visiting facilities remains fixed. If the number of patients attending the health facility were to increase over the evaluation period it is plausible that the proportion of 3 month refill prescriptions and facility congestion could increase simultaneously. Even if the number of patients is fixed, it is not possible to make comparisons across different facilities using this measure of congestion. A facility with few patients and relatively many staff may be less congested than a facility with many patients and relatively few staff even if the proportion which receives 3 month refills is lower. This outcome indicator would have been valid if the evaluation was intended to measure progress against Zambia's national guidelines to prescribe 3 month refills to eligible patients, rather than facility congestion. It is not clear why patient waiting times weren't used as a more direct indicator of congestion, since this data was collected.

2. The intervention being evaluated does not clearly answer the evaluation question.

The primary aim of the evaluation is stated to be assessing the impact of improved service efficiency and quality of pharmacy ART supply on facility-level congestion. However since improvements to facility supplies and efficiency are made in both treatment arms of the study, this is not in fact what is tested for. The RCT in fact tests for the additional impact of having a designated officer working in health facilities to oversee the process of improving service delivery. It is not clear from the evaluation question or results of the Assessment Phase why the effects of this innovation are particularly of interest. Some important details, such as whether this is a new post or not, are also unclear.

3. The study area is small, causing findings to have limited generalisability to other contexts and risking poor internal validity.

Only 16 ART clinics in the Lusaka region are covered by the evaluation. The findings are likely to be too specific to this restricted population to draw wider conclusions for other areas in Zambia. With only eight facilities in each intervention arm there is also a risk that the two groups are not sufficiently comparable before the start of the evaluation to be confident that observed changes in outcomes are due to the effects of the intervention alone (despite the pair-matching design).

4. An RCT was not required to provide useful evidence for policymakers.

The Decongestion evaluation is the clearest case among the 3DE evaluations of where an RCT was not the best approach. The assessment phase, in which supply chain challenges in health facilities were diagnosed and solutions proposed, was considered to be the most useful aspect of the study by key informants who spoke to us. Given the weaknesses to the RCT outlined above, this unlikely to add appreciable value to the study.

Table 9 Detailed quality assessment of the Decongestion evaluation

Category	Proposed questions	Comments
Planning and context	1.1 How relevant are the evaluations questions to the priority questions of the Ministry? (explored as part of validation of the ToC)	Although the study protocol explains why facility congestion is a relevant problem that the Ministry of Health may be interested in addressing, it does not suggest that the policy being tested by this evaluation (nominating a specialised Quality Improvement Officer in health facilities) is particularly of interest.
Introduction	2.1 Is the evaluation question(s) written simply and clearly?	Yes, the evaluation aims and objectives are set out clearly in the Study Protocol.
	2.2 Are the evaluation questions suitable given the short duration of the evaluation period?	Facility congestion might reasonably be expected to vary over a medium time horizon and is therefore an appropriate outcome to track for the evaluation. However, it is not clear that the intervention under test is suitable for a rapid evaluation because it implements an ambitious programme of health systems reform. Ensuring adequate supply and efficiency at health facilities is likely to be a challenging task which may take some months to become established before it has the potential to be effective.
	2.3 Is there an adequate description of the intervention to be evaluated (this should include detail on the intervention's target groups, timescale, geographical coverage, anticipated impact, outcomes and outputs, intervention logic and/or theory of change)?	The components of the intervention are well described. The timing of the evaluation as shown in the Study Scheme (6.1.2) is a bit confusing as it does not show the initial assessment phase, which provides the basis for randomising facilities into treatment groups. It also mentions a midline analysis at one point in the protocol which is not described elsewhere. A further issue with the intervention description is that the intervention logic is not made clear. A number of challenges surrounding supply shortages and clinic efficiency are identified at various points in the protocol. However the evaluation does not directly test for the effects of reinforcing facility supply chains, since basic improvements are made to facilities in both treatment arms. Instead what is being evaluated is the impact of having a designated person at a health facility to oversee supply chain improvements and ensure that guidelines are followed. The rationale for nominating a person to perform this role should be explained somewhere in the protocol, as it is not clear why this is of interest as the central innovation that the evaluation will assess. To help justify the intervention under test, the protocol would have been greatly strengthened by presenting the findings of the baseline assessment.
	2.4 Is there a discussion of other programmes or interventions that may also affect impact, outcome and output indicators?	The protocol notes that facilities participating in other trials will not be eligible for selection into the sample.

Category	Proposed questions	Comments
Method	3.1 Is a RCT the most appropriate method to answer the evaluation question	An RCT may not be the most appropriate method for this study. We would agree with some of the concerns highlighted by IDinsight in their evaluation assessment. Given the limited budget available for the study, the evaluation is only practically able to cover facilities in the Lusaka region. This means that there are few facilities available to be included in the evaluation, which runs the risk that the study will be underpowered to detect significant changes in facility congestion and not internally valid. It also means that the evaluation has limited generalisability to other areas in Zambia, which limits how useful the findings of the study will be to inform a policy-relevant decision. Secondly, we gather from our key informant interviews that the aspect of this study which was most appreciated by its intended users was the initial assessment phase to diagnose supply chain issues and propose potential solutions. Although the results of the RCT had not been shared at the time of writing, we did not get the impression that there was considerable interest among stakeholders in learning the findings. The implications of the initial assessment phase were instead highlighted as being very useful for enabling better programming decisions to be made in view of the particular challenges facing different facilities. It is our view that the addition of an RCT over a limited sample does not add much value to this study, which may therefore have been better implemented as an assessment phase followed by an operational pilot of suggested improvements.
	3.2 Is the unit of randomisation appropriate?	Yes, randomising at the health facility level is appropriate. Since the intervention involves changes to health facility systems, randomisation at the individual level would not have been an option, and randomising at any higher level would have required a larger study area.
	3.3 Did the randomisation produce treatment and control groups that were similar at baseline?	N/A. Lack of balance at baseline is a risk for this evaluation design since there are only 8 facilities per study arm. The pair-matched design may however help to deliver improved balance despite the limited sample size. The protocol could usefully provide more information about how pair matching was performed and what data was used to do it.
	3.4 Are issues related to spillover effects/externalities (untreated individuals are affected by the treatment) considered and dealt with appropriately?	A spillover in this context could arise if Simple intervention facilities benefit from the activities of the Quality Improvement Officer in Comprehensive Intervention facilities, for example from the stock checks at the clinic level and other supply chain management support.
	3.5 Are issues related to imperfect compliance (people in treatment group not being treated, or people in control group being treated) considered and dealt with appropriately?	Non-compliance in this setting could arise if people who would usually visit a Control group facility switch to a Treatment group facility in order to have a better chance of obtain a 3 month refill. It is not clear whether there will be measures in place to prevent this from happening, or if this will be tracked.

Category	Proposed questions	Comments
	3.6 Are local and national contextual factors that could affect the evaluation considered?	None are mentioned.
	3.7 Is the timing of the data collection appropriate given the timing of the intervention?	Endline data collection occurs at the end of a 3 month intervention period. As discussed, this timing may be too soon to capture data that accurately reflects the potential of the intervention to improve outcomes. More time may be necessary to allow the intervention to become established and attain smooth running of operations. The explanation of the timing of data collection is occasionally confusing. Section 6.4.3 mentions a midline assessment, which is not discussed elsewhere in the report. Section 6.1.2 also refers to the baseline component of data collection as being two different stages of the study – we think the first time this is mentioned you mean the 'initial' assessment.
	3.8 Can the findings be expected to have reasonable external validity to inform a wider policy or programmatic decision?	The evaluation has limited external validity, since it only covers facilities in the Lusaka region that are eligible for the intervention under pre-established criteria. As noted by IDinsight recommendations, the findings will only be valid for other urban facilities with a need for the intervention. Initial assessments indicated that even in the Lusaka region there were relatively few facilities that qualified for the intervention. As discussed, this greatly weakens the usefulness of the findings to contribute to a policy decision affecting a broader area.
	3.9 Were there any trade-offs in design due to the relatively short time frame of the evaluation, and if so what were they?	There do not appear to have been any major trade offs in the design of the evaluation due to the short time frame. Although the intervention under test seeks to improve health systems, which we have previously suggested may be better suited to a longer evaluation period, the intervention in this case is not too far-reaching. The basic arm of the intervention reinforces existing systems and guidelines in health facilities, and the innovation provided in the comprehensive arm is to hire a new staff member to oversee supply chain issues and operating procedures. It is not expected that either intervention would require more time to establish routine operations than permitted by the evaluation horizon.
	3.10 Are there other significant methodological limitations (not mentioned above)?	A major limitation with the study is that there are only 8 facilities in each intervention arm. Despite a pair-matched design, with so few clusters it is difficult to be confident that a reasonable degree of internal validity is achieved. This is because systematic differences between the two groups which are unobserved or unmeasured may remain despite matching. We also found the description of how pair matching was done confusing and therefore cannot comment on whether the procedure was suitable. The protocol notes that facilities were pair matched by integration status and the proportion of patients on 3 month refills at baseline. We are firstly not sure what integration status means and how it is defined. Secondly, there is some confusion over the use of the word baseline, which we assume in this context refers to data collected during the initial assessment phase (as described in section 5.4.1). However later on page 15 the term 'baseline' is used to describe a period of data collection which occurs after pair matching, in intervention facilities over a 2 month period. Finally, the protocol does not justify why these two variables were chosen as the basis for pair-matching.

Category	Proposed questions	Comments
		However, we note that a fuller description of the pair-matching process may have been reserved for the final technical report. It is not clear why the baseline assessment that took place after pair matching is not carried out for control group facilities as well as intervention group facilities, since this would have allowed more rich information to be gathered. This decision may have been due to a lack of resources available to roll out the assessment phase to all 16 facilities. The definition of the key outcome variable is problematic. This is discussed in 7.2. No justification is provided for terminating the intervention if there is evidence of a beneficial effect of 20 percentage points or more at midline.
Data	4.1 Were the most suitable data sources selected? If primary data collection was undertaken, were the most suitable data collection methods selected?	The study included several different phases of assessment and data collection. A good description is provided for how source data is physically recorded by interviewers onto an electronic database. The initial assessment phase that took place before facilities were randomised into treatment groups is not well documented. The protocol does not indicate what data was collected, how it was collected or how it was analysed. This is important since the initial assessment forms the basis of the pair matching exercises which determines the treatment groups for the evaluation. More detail is provided to describe data collection for the baseline and endline phases, including which data sources and collection methods were used and how respondents for primary data collection were chosen. The mixture of patient and provider interviews with facility register data is sensible to gather relevant information to inform the intervention design.
	4.2 Have the sampling frame and the sampling populations been correctly defined?	Yes, the sampling frame at the patient level is correctly and clearly defined, to include all ART facilities in Lusaka and HIV positive adult patients on first-line ART who attend those facilities between a given time period. However, the sampling frame of facilities is not clear since the protocol does not state how many eligible facilities there were from which to choose the 16 that made up the sample.
	4.3 Is the sampling procedure rigorous and appropriate? (What is the sample representative of?)	Facilities were selected into the sample randomly from the total population of eligible facilities. The protocol notes that weights will be used in analysis to prevent patients from smaller facilities being overrepresented in the analysis. This would arise otherwise, since facilities were not selected in a proportional manner and a fixed number of patients per facility were included in the analysis.
	4.4 If primary data collection was undertaken, are survey instruments well-constructed (clear, robust skip patterns, relevant answer codes) and are they adequately described?	Data collection forms were not present in the protocol document as provided to us, but appear to have been provided as Annexes in the original version. We have had access to a Clinic Flow Form from the baseline stage only, which is clearly presented. We have not seen Health Facility Provider Interview forms, District pharmacy staff interview forms or medical stores limited staff interview forms.

Category	Proposed questions	Comments
	4.5 Are secondary data sources adequately described and has their quality been checked to determine the data is reliable?	A list of secondary data sources and the outcomes that they will be used to calculate is given. It is not clear whether the data will be validated or quality assured in any way.
	4.6 Were sample sizes adequate?	N/A
	4.7 Were sample size calculations done well and are they presented?	The formula used to calculate sample sizes is clearly presented and is appropriate for comparing two proportions. However not all the parameters used in this formula are defined and there are some issues with the definitions that are given. π_0 is defined twice. The first time it is defined what is meant is π_1, and the definition should specify that the parameter refers to the change in proportion in the intervention sample.
	4.8 If primary data collection was undertaken, are any biases from non-response discussed?	N/A
Data Collection	5.1 If primary data collection was undertaken, were data collected in an appropriate and respectful manner, taking into account cultural, ethical, as determined from the protocols submitted for ethical approval, the field manual and the characteristics of the data collectors?	The study protocol provides a good indication that appropriate ethical practices were followed in data collection, including consent being sought and respected during patient interviews and staff being trained to ensure confidentiality. All questionnaires will be submitted for IRB approval.
	5.2 If primary data collection was undertaken, were the instruments tested and validated (e.g. pre-testing of questionnaires)?	The protocol indicates that a pilot will be undertaken, but does not describe whether this includes pre-testing of surveys, with the option of refining the instruments afterward.
	5.3 If primary data collection was undertaken, were the instruments translated and back translated?	This is not mentioned.
	5.4 Were field teams trained to gather data before the start of the intervention? If primary data collection was undertaken, were the field teams trained by the same people who made and tested the survey instruments?	Yes, training in data collection is planned. The nature of this training is not described.

Category	Proposed questions	Comments
	5.5 Has there been an appropriate level of oversight and data quality assurance in the data collection?	Yes, the protocol describes a good degree of supervision and oversight to data collection by Principle Investigators throughout the study process.
Data entry and cleaning	6.1 If a survey was undertaken on paper, was the data double entered and were discrepancies between the two entries systematically resolved by checking the hard copies?	We understand from our key informant interviews that data was manually recorded into workbooks was double entered.
	6.2 Was the data cleaning done is a robust, clear and transparent way and does it include both range and consistency checks?	Since data has not been collected yet, the protocol doesn't explain what cleaning protocols were followed. It does however indicate that data will be checked before analysis for consistency, logic and range either through electronic tablets or through reviewing the study database. Further details of what the intended consistency checks will be are not given.
Data analysis	7.1 Are primary analysis methods appropriate? If regressions are used, are they correctly specified and are standard errors calculated correctly?	The primary analysis is well described.
	7.2 Are the key indicators clearly defined including how they are calculated, and are they suitable to measure the outcomes of interest?	Yes, the primary outcome variable for analysis is clearly defined as the proportion of stable ART patients on 3 month refill prescriptions. Definitions for stable patients and eligibility for 3 month refills are provided. This is however not likely to accurately capture facility congestion. It is a reasonable proxy only if the stock of patients who visit the facility remains fixed, and the proportion patients receiving a 3 month prescription increases. However even in this case, this measure would only permit comparisons to be made in congestion levels within the same facility over time. It is not possible to conclude that a facility where a higher proportion of eligible patients receive 3 month refills has lower congestion. Furthermore, a high level of 3 month refills may itself by a signal of facility congestion if facility staff are more likely to make these prescriptions during times when congestion is worst. We are unsure why the study doesn't use patient wait time at clinics as the primary indicator of facility congestion. Patient wait times are used to measure 'patient flow' in secondary analysis, but it is not clear how this is different from facility congestion.
	7.3 Have sampling weights been used correctly?	The protocol indicates that weighted values will be used in analysis to account for varying facility sizes.
	7.4 Are departures from the full sample size (non-response) explained and has any non-random attrition been identified and dealt with correctly?	N/A

Category	Proposed questions	Comments
	7.5 Have any differences between treatment and control groups at baseline been accounted for in measures of impact?	N/A. However the study protocol indicates that if facility pairs do happen to be unbalanced, this will be accounted for in analysis.
	7.6 Is the analysis disaggregated to show outcomes and impact on different groups and sexes? Did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?	N/A
Reporting	8.1 Are quantitative results interpreted and presented systematically and logically?	N/A
	8.2 How clear are the links between data, interpretation and conclusions?	N/A
	8.3 Were negative/discrepant results addressed or ignored?	N/A
	8.4 Are the final recommendations and conclusions plausible?	N/A
	8.5 Have alternative explanations been explored and discounted?	N/A

E.5 Family Clinic Day evaluation

Appropriateness of evaluation questions

The need to improve care practices for HIV positive adolescent and paediatric patients is well justified, given the decision taken in Uganda in 2014 to exceed WHO guidelines by recommending ARV treatment for all HIV positive patients under 15 years of age (rather than only under 3). There is also shown to be little documented evidence on the effectiveness of Family Clinic Days, even though they are already practised across Uganda.

The rationale for this kind of intervention could be better motivated, for example by providing evidence to demonstrate that the clinic appointments of adolescent/paediatric patients and their caregivers are in fact currently poorly aligned, and that this contributes to poorer HIV care retention.

Robustness of evaluation design

The following are the main findings of the quality assessment around the robustness of design:

1. The intervention may have been too ambitious for the limited time horizon.

The intervention involved changing the scheduling of clinics in health facilities and the kind of services offered. As with the intervention in the EID evaluation, this kind of health systems reform may require time to become consolidated and may not have therefore been suitable to be evaluated within a short time period.

2. The evaluation design is otherwise well conceived to address the research question under the timeline and available budget.

Overall, this evaluation design was well conceived. The randomisation was conducted at a suitable level given the nature of the intervention (which would not lend itself easily to household level randomisation). Outcome measures were appropriate for the evaluation time frame. A good use was made of facility data to measure primary outcome variables.

Table 10 Detailed quality assessment of the Family Clinic Day evaluation

Category	Proposed questions	Comments
Planning and context	1.1 How relevant are the evaluations questions to the priority questions of the Ministry? (explored as part of validation of the ToC)	The decision to implement ambitious national guidelines on adolescent and paediatric HIV care, which exceed the recommendations made by the WHO, provides a justification for attention on this area of care. The protocol also indicates that the MoH has requested further evidence on the FCD model of care to inform its programming decisions.
Introduction	2.1 Is the evaluation question(s) written simply and clearly?	Yes.
	2.2 Are the evaluation questions suitable given the short duration of the evaluation period?	Yes the evaluation questions are suitable because they focus on medium term outcomes. However it is not clear that this intervention is suitable for a rapid evaluation, since it involves reforms to health facility scheduling and practices which may require some time to bed in.
	2.3 Is there an adequate description of the intervention to be evaluated (this should include detail on the intervention's target groups, timescale, geographical coverage, anticipated impact, outcomes and outputs, intervention logic and/or theory of change)?	The components of the intervention and target groups are well described. More detail could be provided on which districts and areas within the central, eastern and northern regions of Zambia the intervention is rolled out to. The intervention logic could be better described. Although some family-centred issues around HIV care are outlined, a more convincing case could be made to suggest why delivering care in family units is expected to be effective in this context. For example, it would be helpful to provide evidence to demonstrate that appointments for adolescents and their caregivers are in fact poorly aligned currently and that this contributes to appointments being missed. It would also be useful to outline what kind of care is provided by standard ART clinics and in what specific ways the FCD model is different.
	2.4 Is there a discussion of other programmes or interventions that may also affect impact, outcome and output indicators?	None are described.
Method	3.1 Is a RCT the most appropriate method to answer the evaluation question	An RCT is a reasonable choice in this case, since there is a policy relevant question to answer for which there is currently an identified evidence need. There is also no reason to believe that control group individuals will be disadvantaged by not having access to the intervention, since no group in the study is prevented from obtaining appropriate HIV care. The intervention may be too ambitious to evaluate using an RCT in a short time frame, since it attempts bring about a systems change to health facilities rather than a simple delivery of some service. As with some of the other 3DE evaluations, this kind of reform may take time to become established, and a rapid RCT would not therefore capture its full potential.
	3.2 Is the unit of randomisation appropriate?	Yes it is appropriate, since any scale up of the programme would occur at the health facility level too. There are no practical possibilities for rolling out this intervention at any other level.

Category	Proposed questions	Comments
	3.3 Did the randomisation produce treatment and control groups that were similar at baseline?	N/A
	3.4 Are issues related to spillover effects/externalities (untreated individuals are affected by the treatment) considered and dealt with appropriately?	Spillovers are not discussed. It is plausible that individuals who receive the specialised health education through an FCD share their knowledge with other individuals assigned to the control group, and that this influences control group members to alter their behaviour in some way. The extent of knowledge sharing and influence between individuals would be difficult to measure and account for.
	3.5 Are issues related to imperfect compliance (people in treatment group not being treated, or people in control group being treated) considered and dealt with appropriately?	The risk of non-compliance is not discussed in the study protocol. However it is not clear whether there are any procedures in place to prevent control group individuals or family units from seeking treatment at a FCD. Table 1 indicates that patients not scheduled for an appointment may still be seen (last) if they attend the FCD.
	3.6 Are local and national contextual factors that could affect the evaluation considered?	Yes, there is a good discussion of contextual factors and background to the study.
	3.7 Is the timing of the data collection appropriate given the timing of the intervention?	As discussed, the timing of data collection may be too soon to capture data that accurately reflects the potential of the intervention to improve outcomes. More time may be necessary to allow the intervention to become established and attain smooth running of operations.
	3.8 Can the findings be expected to have reasonable external validity to inform a wider policy or programmatic decision?	The study spans a fairly broad area, covering the Central, Northern and Eastern regions of Uganda. The findings may therefore have reasonable external validity to the potential scale up region. However the Protocol does not indicate how many districts are included in each region so the exact coverage of the evaluation is not known.
	3.9 Were there any trade-offs in design due to the relatively short time frame of the evaluation, and if so what were they?	If the evaluation period is too brief to allow the intervention time to establish routine operations over the study horizon, the findings may underestimate the extent to which Family Clinic Days have the potential to improve outcomes.
	3.10 Are there other significant methodological limitations (not mentioned above)?	The study is generally well designed and does not have significant methodological limitations. As discussed in the Protocol, one of the main outcome variables for the evaluation (adherence) is difficult to measure accurately without taking blood samples, and results may therefore be sensitive to the particular choice of proxy variable used for analysis. It also appears that quality of facility data may be an issue more generally, since the study seems to anticipate a potential loss of data of up to 50%. 23 facilities are included in each treatment arm, which is reasonably small. Care would need to be applied during analysis to account for possible implications of this limited sample (such as the effects of

Category	Proposed questions	Comments
		correlation in outcomes within clusters, and the possibility that the two treatment arms are not fully balanced at baseline).
Data	4.1 Were the most suitable data sources selected? If primary data collection was undertaken, were the most suitable data collection methods selected?	The study proposes a good use of data within the available budget, combining routine patient data from facility registers and HIV care cards with focus group discussions to supplement the quantitative results. Data sources, including qualitative methods, and data collection methods are described very well.
	4.2 Have the sampling frame and the sampling populations been correctly defined?	The sampling frame appears to be correctly defined to answer the research questions on paediatric and adolescent HIV outcomes. However the protocol does not indicate the sample frame of facilities; it is not clear from how many facilities the selected 46 were chosen from.
	4.3 Is the sampling procedure rigorous and appropriate? (What is the sample representative of?)	The terminology used to describe the sampling process is a bit confusing, in particular, a 'cluster randomised sampling process'. We think you mean simple random sampling with stratification. The sampling process is otherwise well described. The sampling process may have suffered from defining too many strata. The protocol notes that stratification is done by implementing partner, region and health facility level. With three regions (Northern, Central and Eastern), three implementing partners (NUHITES, Mildmay Uganda and STAR-E) and three health facility levels (Health Center III, IV and General Hospital) this implies $3^3 = 27$ strata. It would be helpful to explain the distribution of the 46 facilities across this number of strata. A fixed number of patients were chosen from each facility, which means that patients from smaller facilities will be overrepresented in the final sample. This is acceptable if sampling weights are applied in analysis to account for this. However we would recommend that the study could have selected a representative sample to begin with by applying the probability proportional to size method (PPS). This would involve purposefully selecting larger health facilities into the sample with a higher probability to compensate for the oversampling of patients from smaller facilities which occurs when a fixed number of patients are chosen from each facility at the second stage. Using this method would deliver a sample in which each individual in the population has an equal probability of being sampled.
	4.4 If primary data collection was undertaken, are survey instruments well-constructed (clear, robust skip patterns, relevant answer codes) and are they adequately described?	Survey instruments are presented in annexes to the study protocol and appear to be well constructed.
	4.5 Are secondary data sources adequately described and has their quality been checked to determine the data is reliable?	As discussed below, the protocol seems to imply that there may be some data quality concerns with facility register data, since the study is powered to handle a loss of up to 50% data. Data quality checks to verify patient records do not appear to have been planned.

Category	Proposed questions	Comments
	4.6 Were sample sizes adequate?	N/A
	4.7 Were sample size calculations done well and are they presented?	Yes sample size calculations are presented and use an appropriate formula, the reference for which is cited. It is not clear why a different formula should be used in the power calculation to specify the test of outcomes for adult patients. The original specification should be equally appropriate to handle a positive change in outcomes and a negative change, if that is what is anticipated.
	4.8 If primary data collection was undertaken, are any biases from non-response discussed?	N/A since the study is still underway. However it is interesting to note that a possible loss of 50% patient data is anticipated from facilities. It would be interesting if the protocol could describe why such a loss might be expected, and discuss whether this is a possible source of bias. The fact that many health facility records could be potentially missing could also be an indicator of poor quality data records more generally, which raises questions about the suitability of the facility registers as a primary source of data for the evaluation.
Data Collection	5.1 If primary data collection was undertaken, were data collected in an appropriate and respectful manner, taking into account cultural, ethical, as determined from the protocols submitted for ethical approval, the field manual and the characteristics of the data collectors?	Yes, ethical protocols are well described and appear thorough.
	5.2 If primary data collection was undertaken, were the instruments tested and validated (e.g. pre-testing of questionnaires)?	Yes, pre-testing of data collection instruments is discussed.
	5.3 If primary data collection was undertaken, were the instruments translated and back translated?	It is not clear whether back translations were made of the qualitative data collection instruments.
	5.4 Were field teams trained to gather data before the start of the intervention? If primary data collection was undertaken, were the field teams trained by the same people who made and tested the survey instruments?	Training procedures are well documented and seem to be appropriate.
	5.5 Has there been an appropriate level of oversight and data quality assurance in the data collection?	Yes, data consistency checks and back check processes describe seem thorough.

Category	Proposed questions	Comments
Data entry and cleaning	6.1 If a survey was undertaken on paper, was the data double entered and were discrepancies between the two entries systematically resolved by checking the hard copies?	N/A
	6.2 Was the data cleaning done is a robust, clear and transparent way and does it include both range and consistency checks?	Cleaning of final data is not described as the study is still underway. However procedures for checking the consistency and quality of data as it is being collected are described.
Data analysis	7.1 Are primary analysis methods appropriate? If regressions are used, are they correctly specified and are standard errors calculated correctly?	Full regression models are not specified at this stage. The analysis plan seems good and includes a mixture of analysis on the difference in proportions of patients attaining various outcomes between the treatment and control arms of the study, and individual analysis based on logistic regressions.
	7.2 Are the key indicators clearly defined including how they are calculated, and are they suitable to measure the outcomes of interest?	All indicators are clearly defined. A good discussion is also included around the shortcomings of different measures of adherence, to justify measuring this outcome in different ways as a robustness check. The variables used for individual level analysis are not defined, but may be inferred. Difficulties around measuring adherence accurately without using blood samples are acknowledged and discussed.
	7.3 Have sampling weights been used correctly?	The use of sampling weights is not mentioned, but will be necessary given the overrepresentation of patients from smaller facilities in the selected sample.
	7.4 Are departures from the full sample size (non-response) explained and has any non-random attrition been identified and dealt with correctly?	N/A
	7.5 Have any differences between treatment and control groups at baseline been accounted for in measures of impact?	N/A
	7.6 Is the analysis disaggregated to show outcomes and impact on different groups and sexes? Did the 3DE evaluation questions and designs ensure effects on women and girls and the poorest and most vulnerable were included?	N/A

Category	Proposed questions	Comments
Reporting	8.1 Are quantitative results interpreted and presented systematically and logically?	N/A
	8.2 How clear are the links between data, interpretation and conclusions?	N/A
	8.3 Were negative/discrepant results addressed or ignored?	N/A
	8.4 Are the final recommendations an conclusions plausible?	N/A
	8.5 Have alternative explanations been explored and discounted?	N/A

Annex F Political Economy of Ministry of Health and its impact on the 3DE Model

F.1 Introduction

This Annex discusses the political economy of the health sector in Zambia with a view to understanding the contextual factors influencing the outcomes of the 3DE model. More specifically it focuses on two areas outlined in the Inception Report[21]. First, it will explore the political economy of decision-making and resource allocation within the Ministry of Health and the Ministry of Community Development Mother and Child Health. Second, it assess how evidence is used at each step of the policy making cycle. Implications for 3DE and how it operates are given at the end of each of the major sections.

F.2 Country context

Overview

Zambia is a peaceful democratic country with enormous economic potential that is grounded in its rich endowment of natural and mineral resources. Recent macroeconomic trends and developments suggest that Zambia has a robust economy[22]. However, despite its strong economic growth, two-thirds of Zambians still live in abject poverty.[23] In addition, the rural poor persistently lag behind urban dwellers in most measures of social welfare. The contextual factors summarised below provide a synthesis of the political, macroeconomic and the epidemiological factors that influence the outcome of the 3DE model.

The political context

Zambia is a stable constitutional republic that is governed by a democratically elected president and a unicameral national assembly. The country has successfully held national multiparty elections in 1991, 1996, 2001, 2006 and 2011 and has benefited from 24 years of democratic governance. International and domestic election observers have reported that with the exception of the 2001 elections, all the other presidential and parliamentary elections that have been held in Zambia were free of irregularities and fair.

Although political and legal institutions are maturing, there is an emerging political culture that is neither based on rules, nor characterized by unbridled rent-seeking. This culture tends to favour indecision and conservatism rather than radical political or economic departures that could provoke a wider reaction by civil society. As a result, there is limited domestic opposition to maintaining the status quo. In addition, domestic lobbies are not well organised and thus not very visible; and are too weak to effectively pressure the state to improve health services. This promotes inertia in the system.

The macroeconomic context

Zambia's macroeconomic performance is a critical element in the analysis of health outcomes and policy reform. Increased levels of national income per capita allow individuals and households to buy better living and housing conditions and more health care. Similarly, increased economic growth expands the revenue possibilities for government, and thus the opportunities for expenditures to provide sustainable preventative and curative health services.

A review of Zambia's macroeconomic performance indicates that between 2006 and 2010 the country's real Gross Domestic Product (GDP) grew by an average of 6.3 percent per annum. This growth trend continued in 2011 and 2012 when real GDP grew by 6.8 percent and 7.2 percent respectively. Total national income

[21] See the Independent Evaluation of the Demand-Driven Impact Evaluations for Decisions (3DE) Pilot, Final Inception Report of 1st May 2015, p21.
[22] The World Bank, Zambia Economic Brief: Recent Economic Developments and the State of Basic Human Opportunities for Children, Issue No.1, December 2012.
[23] According to the Central Statistical Office's Living Conditions Monitoring Conditions Survey Report 2006 and 2010 the proportion of the population falling below the poverty line reduced from 62.8 percent in 2006 to 60.5 percent in 2010.

has risen by more than 56 percent over the period 2002 to 2010. In 2006 the GDP per capita was US$890. It rose to US$1,221 in 2010 and was US$1,486 at the end of 2012. The GDP per capita was expected to be US$1,622 in 2013 and is projected to reach US$1,858 in 2015.

Official figures suggest that health is a lower priority than education or economic affairs, and that Zambia spends a relatively low proportion of its GDP on health. Expenditure on health in Zambia, at 2.9% of GDP in 2013, is relatively low compared to Mozambique (3.1%), and Uganda (4.3%) though it is higher than Tanzania at 2.7%. In terms of the national budget, functional resource allocation still appears to favour Education and Economic Affairs[24]. As a share of the total government budget, public health spending has typically been below 11 percent (projected 9.7% in 2015). This compares unfavourably with the Education sector, which has received an average of 20 percent of the total government budget over the past four years (projected 21.2% in 2015), and the Economic Affairs function which has, on average, received 22 percent of the total budget (projected 28% in 2015).

The Health context

An overview of Zambia's epidemiological profile reveals the predominance of communicable diseases such as malaria, HIV and AIDS, STIs and TB. In addition, Zambia is faced with a rapidly rising burden of non-communicable diseases such as diabetes mellitus, mental health, sickle cell anaemia, hypertension, cancers, chronic respiratory diseases, blindness and cardio-vesicular diseases[25]. The top ten causes of morbidity and mortality[26] include malaria, respiratory infections (non-pneumonia), diarrhoea (non-blood), trauma (accidents, injuries, wounds and burns), eye infections, skin infections, respiratory infections (pneumonia), ear, nose and throat infections, intestinal worms and anaemia. The most significant epidemiological development in recent years was the advent of HIV and AIDS.

Zambia's disease profile clearly shows that pre-epidemiological transition disorders such as infectious diseases and high infant mortality co-exist alongside risk factors such as unhealthy diets and associated health problems such as cardio-vesicular diseases. Addressing these disorders depends on a host of factors. These include the need to address public health and clinical effectiveness and environmental risk factors such as poor quality and inadequate quantities of water, poor sanitation and poor personal hygiene practices.

Given Zambia's epidemiological profile, the Government and its Co-operating Partners (CPs) have supported a wide range of health sector reforms that are aimed at improving equity and efficiency in healthcare financing and delivery. More recently, Government and the CPs have supported health sector reforms that attempt to meet the health-related Millennium Development Goals (MDGs). These reforms include changes to financing arrangements, resource allocation, public financial management and planning and changes to the organization and management of the health sector.

F.3 The Political Economy of decision-making

The evolution of the political economy of decision-making and resource allocation is deeply embedded in Zambia's political system and processes. The background to the evolution of the political economy of decision-making in Zambia is summarized in Box 6 below.

[24] Government of the Republic of Zambia, Ministry of Finance, Medium Term Expenditure Framework and the 2014 Budget, Lusaka,

[25] For a more detailed analysis of Zambia's epidemiological profile see the Central Statistical Office's recently published Zambia Demographic and Health Survey -2013-2014 of March 2015

[26] Government of the Republic of Zambia, Ministry of Health, National Health Strategic Plan, 2011-2015

Box 6 Evolution of the Political Economy of decision making and resource allocation

To capture the political salience of the current trends in decision-making and resource allocations in the MoH and the MCDMCH, one needs to understand and appreciate the evolution of policy formulation and decision-making in Zambia. Prior to 1990, decision-making and resource allocations to all Ministries, Provinces and Spending Agencies (MPSA) were based on a centralized form of government. Under this system of governance, policy formulation, decision-making and resource allocations were centralized in the Office of the President and at the United National Independence Party (UNIP) headquarters. These decisions were based more on the socialist dogma of the Kaunda administration than on a careful analysis of problems or objectives and possible actions to address them. During the Kaunda administration, there was a notable "disconnect" between policy decisions and outcomes. There was also a marked lack of internal systems for monitoring the results of policy decisions.

In 1990, a group of reformists headed by Frederick Chiluba and the Movement for Multi-Party Democracy (MMD) led the country into a new democratic order that ended 27 years of authoritarian leadership and a state-led and state-controlled economy. The new administration established and maintained a government that was based on Britain's Westminster style of governance and began a process of political and economic transformation that showed strong democratic principles in which sovereign authority and decision-making was increasingly vested in the people.

During the second half of the 1990's, the erstwhile showcase of a smooth democratic transition experienced an authoritarian regression. Chiluba's desire to establish a new and better functioning government apparatus soon collided with the realities of marshalling a government machinery that was steeped in authoritarianism. In addition, Chiluba had to try to energize a demotivated civil service that had long discovered that longevity in government service meant avoiding risk and deferring even the most routine matters upward for decision-making. Democratic accountability began to deteriorate and political and civil rights were increasingly curtailed. The envisaged decision-making process that was based on democratic principles reverted to the centralized form of decision-making

In 2001, civil protests proved effective when civil society garnered massive support to defeat President Frederick Chiluba's unconstitutional bid for a third term in office. After this constitutional victory by civil society, the MMD government regained some democratic legitimacy and went on to win the 2006 elections. Since the 2006 elections, the country has been ranked as an electoral democracy by Freedom House. Despite being ranked as an electoral democracy, the decision-making process is still dominated by the president who is the ultimate dispenser of resources and rents to key supporters.

At the time that DFID agreed to support Clinton Health Access Initiative (CHAI), the *Bertelsmann Transformation Index (BTI) 2012 Zambia Country Report*[27] indicated that the dominance of the executive clearly extended its jurisdiction beyond the stipulations of the constitution. In addition, the report pointed out that the legacy of an authoritarian political culture and ingrained patterns of neopatrimonial governance had persisted.[28] Overall, the BTI report concluded that:

> *"The principles of a democratic government are observable. There are no violent conflicts, no veto actors and there is a high degree of acceptance of the democratic order. The judiciary is relatively independent and there is a functional separation of powers, although the executive is dominant."* [29]

The BTI showed that while Zambia was considered to be an electoral democracy, its citizens were still far from enjoying a satisfactory level of political rights and civil liberties.

The BTI's observations are reflected in the manner in which the Patriotic Front (PF) and previous administrations governed the country. When President Michael Sata was elected into office in 2011, one of his first priorities was to create a more professional civil service that could make a meaningful contribution to the development of national policies. Sata wanted to create a civil service that could be counted on to effectively implement the PF government's policy decisions. This meant that President Sata had to develop a political and administrative atmosphere where politicians and career civil servants, each with a distinct role, would perform their respective tasks as a team. But with an oversized and undertrained bureaucracy, and a fragmented and undisciplined political coalition, Zambia appears to have reverted to a neopatrimonial system of governance.

Implications for 3DE

'Contributing to a policy decision' is a key performance target for the 3DE programme. With limited civil service capacity and a neopatrimonial system of governance, it becomes increasingly unlikely that it will be possible to attribute a change in policy direction to a specific piece of evidence. Care needs to be taken with

[27] The Bertelsmann Transformation Index is a global assessment of a country's transition process in which the state of democracy and market economy as well as the quality of political management in 128 transformation and developing countries are evaluated.
[28] BTI at http:/www.bti-project.org
[29] BTI 2012 Zambia Country Report p.2

what 'a policy decision' means in the Zambian context—and whether to limit this to operational decisions rather than decisions about policy direction, coverage, funding or other issues that are open to political bargaining and rent-seeking.

F.4 Policy formulation processes

Policy-making in Zambia is a complex set of events that determines what actions government will take, what effects those actions will have on social conditions and how those actions can be altered if they are to produce undesirable outcomes. As a process, policy-making in government can be complex and disorderly, without a beginning and an end. The fact that government has adopted a particular policy and produced a policy statement does not mean that the policy-making process is complete as its implementation is not necessarily guaranteed. The policy formulation process in Zambia can be characterised as having four stages: Formulation, adoption, implementation and evaluation. These stages are described further below.

The policy formulation phase consists of problem identification, agenda setting and formulation of policy options. During this stage, policy problems are defined and the policy agenda is set. Cabinet Office normally bases its decisions on the ruling party's *Manifesto,* however policy formulation may also be initiated by the line Ministry, or specific department within or the Ministry of Finance. Once problems are identified and policy agenda defined through consultations, the Permanent Secretary (PS) of lead ministry, presents this to the Minister. If a cabinet decision is required the PS drafts a Cabinet Memorandum (CABMEMO), which is circulated to all ministries following the Minister's approval. This memo is shared to Policy Analysis and Coordination (PAC) Division at Cabinet Office before presentation to the cabinet.

Within each ministry cabinet related issues are discussed through the Cabinet Liaison Committees (CLSs) which is chaired by the Minister and includes deputy minister, PS and Directors of departments. The Secretary of this committee is the Cabinet Liaison Officers who are responsible for coordination of all Cabinet related business within the ministry and the key contact with PAC and other ministries.

Consultation across ministries take place at the Inter-ministerial committees of officials that are constituted ad-hoc by the PAC and in consultation with the ministry initiating policies to ensure consultation and discussions in preparation of the draft memo mentioned above and in order to support the implementation of the decision. The functions of this committee revolve around collation and scrutiny of evidence and ensuring that alternative options are considered. This committee also advises on whether the selected option is mainly related to administrative decisions or whether it needs Cabinet approval.

The adoption of policy takes place at the Cabinet level following the above mentioned consultation processes. Cabinet has a constitutional role to formulate policy and to advise the President on all matters pertaining to policy and governance of the state. Interests pursed include power and control of the political planning process in the sector by favouring actors, interest groups, districts and provinces that have allegiances to the ruling Patriotic Front government. Although Cabinet has potential to wield a great deal of political power over the sector, informal objectives and inter-personal relations with the President and his inner circle of advisors often takes precedence and centre-stage over formal policy objectives. Decisions on implementation are relayed ministries responsible through the Cabinet Secretary and finally the lead ministry, PAC and the Cabinet Office are mandated with monitoring and evaluating the implementation of policies.

The Cooperating Partners (CPs) mainly engage with the policy formulation process through the *Sector Advisory Groups* (SAGs). The SAGs were developed out of the Poverty Reduction Paper (PRP) working groups. SAGs facilitated the structured participation of both state and non-state actors in the formulation, implementation, monitoring and evaluation of development plans. SAGs provided advice on planning principles and development objectives. In addition, SAGs identified critical issues that might impede development. SAGs were critical in coordinating the delivery of services, streamlining decision-making in the sector and ensuring transparency and accountability. SAGs use evidenced-informed research that was principally generated by multilateral and bilateral agencies. The MoH and MCDMCH SAGs were established as a consultative forum for representatives of key stakeholders active in the health and social protection sector and are viewed as one of the more effective SAGs.

The CPs are involved in the Zambian Health Sector Wide Approach (SWAp) is managed through various structures and meetings that are enshrined in the MoU signed in 2012/13 between the Government and

Partners, in addition to the SAGs and TWGs other coordination meetings include the Annual Consultative meeting, MoH/ CP Policy meetings; and Health Sector Joint Annual Review.[30]

Implications for 3DE

The CP can engage with the policy formulation process during the consultation phase of policy formulation process or by earlier lobbying through the technical channels to instigate a policy review and discussions from the within the Ministry. SAG is the main forum for discussion around policy between CPs, NGOs and other non-governmental stakeholders. Technical discussions on specific health issues take place within the TWGs. The TWGs are an appropriate forum for discussion of the 3DE evaluations, given their operational nature.

Sector Advisory Groups are an important vehicle for evidence-informed policymaking but it is not clear that they prioritise evaluation evidence over other forms of evidence. Given the point made above about the need to specify what 'a policy decision' means for 3DE, it is worth considering whether getting 3DE-generated evidence discussed in SAGs is a worthwhile goal of its own.

F.5 Ministry of Health and Ministry of Community Development, Mother and Child Health

Delineation of responsibilities

The MoH has an unabbreviated and detailed organizational chart. This chart is based on the National Establishment Register which details the established positions and salary scales for each government institution. It must be pointed out that staff are not hired or assessed against thoroughly developed job descriptions that define their duties and responsibilities and performance standards. Most of the MoH job descriptions are centrally mandated and are not regularly updated and the delineation of responsibilities are not tailored by the MoH or fully aligned with the Establishment Register. Many job descriptions are minimal in scope and do not include the details of the positions' duties and responsibilities.

There are a number of key decision makers in both the MoH and the MCDMCH. At the policy level, the Cabinet Minister is both the political head of the ministry and the key decision maker. He/she is answerable to the legislature and through the legislature to the public, both for his or for her personal acts and for the acts of their departmental subordinates within the ministry. The Minister has the constitutional role of formulating policy and/or facilitating the formulation of policy. The Cabinet Minister advises the President on all matters pertaining to policy and governance of the ministry and the sector. The Minister's networks and their underlying interests and ideas have a strong bearing on discourse and advocacy for evidence-informed research. Interests pursed include power and control of the political planning process in the sector by favouring actors, interest groups, provinces, districts or geographical areas that have allegiances to the ruling Patriotic Front government.

The Cabinet Minister who is either an elected Member of Parliament or a nominated Member of Parliament is supported by one or two deputy ministers[31]. The Deputy Ministers in the MCDMCH and the MoH are appointed by the president and, in the absence of their respective Cabinet Ministers, play a pivotal role in the policy and decision-making process.

At the executive level, each of the ministries is headed by a Permanent Secretary (PS). The PS is the administrative head of the ministry and is responsible for a number of functions. First the PS acts as the principal advisor to the Cabinet Minister. Under this role, the PS provides objective advice on policy issues and on the government's options in dealing with them and the implications of each option. The advice requires a complete understanding of complex technical, managerial, legal and financing issues. As chief executive of the ministry, the PS must direct and coordinate the activities of the ministry on behalf of the Minister and within the laws and statutes governing the health sector. A key role of the PS is to ensure that the departments respond to ministerial priorities and that the administration of the ministry is carried out in a way that reflects the Minister's direction and interests. Another role that the PS plays is to set policy and to

[30] WHO (2014), Country cooperation strategy brief.
[31] At the time of this evaluation, the MoH had one Deputy Minister, while the MCDMCH had two Deputy Ministers.

ensure that there are effective linkages among and within the functional Heads of Department who are responsible for executing the technical aspects of the ministry's functions.

At the operational level technocrats in the MoH and the MCDMCH have a vast array of skills and knowledge of the health sector. Many started their careers in the civil service as enthusiastic "anglophiles" who had inherited and adapted the British model of public service management to their own ends. In the initial stages of their career, many had been trained to handle politicians. They learned to 'say "no", without saying "no" to politicians by pointing out that "it can be done sir but you may want to consider the following....." They also understood the principle of keeping the civil service separate from politicians.

Under the PF and previous governments, the principle of keeping the civil service separate from politicians has been eroded and the public service is more politicised. Professionalism in the heath sector has given way to a two-track socio-political and economic programme. The first is grounded in the formal policy documents as outlined in the Health Strategic Plan. The second follows a different path, namely a clientist political logic that is aimed at pleasing and appeasing the political leadership and keeping them in power. In such a situation the interests pursued by the civil servants is to pay allegiance to their political masters whilst trying to amass as much wealth as possible during the PF's term of office.

At both the executive and directorate levels, failure to delegate significant powers has not only helped to de-motivate the bureaucracy, but has also helped increase the level of inefficiency in the health sector. With a new PF government in place, vital decisions are being delayed because senior civil servants are afraid to take action in the absence of authority from above. As one key informant noted, failure to question irrational decisions, failure to bypass obsolete rules whose origins and functions are obscure and failure to tackle problems in a proactive manner continue to stall project planning and implementation in the health sector.

Implications for 3DE
This raises the question of what 'demand-driven' really means for 3DE. The extent of political encroachment into the civil service implies that it may be hard to distinguish a political demand for evaluations from an evidence-based demand (for example, one based on a review of the evidence about what is needed to improve coverage, target minorities, increase effectiveness or address other specific priorities).

Decision making process within Ministry of Health

The decisions made by the ministry originate from ideas and suggestions made by stakeholders from either within or outside the ministry. Subject specialists within the ministry subject each idea to a thorough analysis. This includes reviewing existing policies and guidelines to ensure that the idea helps improve the operations of the ministry.

If the idea passes the first level of scrutiny it is sent to the PS who may authorize the Directorate of Planning and any other key players to review the pros and cons of the suggestion. If the idea has some merit, the Policy and Planning Directorate prepares a Position Paper for the PS to review. Once the PS reviews the Position Paper and is happy that the idea has merit, the PS calls for a Senior Management Meeting to discuss the Position Paper and to obtain buy-in from the PS's Senior Management Team[32] consisting of the Directors of Policy and Planning, Human Resource and Administration; Clinical Care; Disease Surveillance, Control and Research; and Technical Support. Staff from Accounts and Audit; Procurement; and Technical Advisors are regularly invited to the meeting.

If the Senior Management team reaches consensus, the PS submits the Position Paper to the Cabinet Minister for approval. The Cabinet Minister will review the Position Paper and ensure that the views contained in the Position Paper are technically and politically correct. Before the minister approves the Position Paper, he/she may call for a Policy Making meeting. The minister chairs this meeting and his/her approval signifies that a Cabinet Memorandum (Cab Memo) can then be prepared for submission to the PS Policy Analysis and Coordination Unit (PAC) which is housed at Cabinet Office. The PS at PAC would ensure that the idea is technically correct and politically acceptable. The Cab Memo would then be sent to the Secretary to the

[32] It was established to oversee all policy and programme implementation and to facilitate all development and monitoring of the various programmes within the ministry.

Cabinet who would check it and place it as an agenda item at the next Cabinet Meeting (See r description of the policy process above).

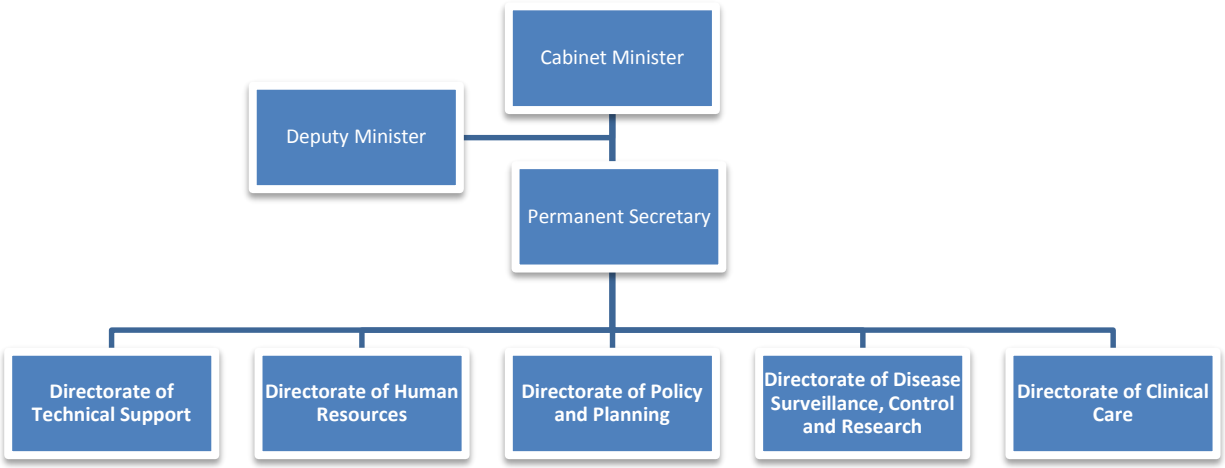
The decision-making model outlined above is an amalgamation of the organizational process model and the political bargaining model. The organizational process model emphasizes the centrality of routines and procedures and reduces the effects of uncertainty. The political bargaining model allows individuals, groups and organizations to have their self-defined interests protected throughout the decision-making process

Organisation and management

The MCDMCH and MOH are both agencies of government. The former was established by a presidential degree in 2011 while the latter was first created by an act of parliament on 11th April 1930[33]. Although the statutory and portfolio functions of each ministry differ, their governance structures are similar.

Figure 10 below summarizes the functional structure and reporting lines of the directorates at the Ministry of Health's headquarters in Lusaka.

Figure 10 Functional and reporting structure of Ministry of Health



Although the functional structure of the MoH gives the image of being a cohesive, clear-cut, centralised, chain of command structure, the actual organisational and managerial structure of the MoH is less clear-cut than it appears at first sight. There are a number of reasons for this. First, the MoH is experiencing a host of staffing challenges. These include the lack of qualified staff, underperforming staff and the chronic problem of significant personnel changes.

Second, a certain "vagueness on questions of competence" is apparent throughout the ministry's administrative system. There is a widespread deployment of staff in positions for which they do not have the experience; partly due to an overall staffing shortage at the MoH. Third, the morale at MoH is low. Consequently, absenteeism and tardiness is high. In addition, numerous vacancies across the system are inhibiting service delivery.[34] Fourth, although attendance logs are kept, the link between these logs and consistent monitoring that is tied to repercussions is not clear. For the most part, attendance is still monitored by direct supervisors who are often out of the office in meetings.

In general, ambiguity concerning responsibilities and powers can lead to a further problem that involves the culture of decision-making. An environment of uncertainty about responsibility and authority to decide can lead to inertia. A pervasive problem in the MoH's administrative structure is the general tendency to stifle initiative and innovation. This, plus the continued fragmentation of institutional responsibilities have perpetuated a culture in which offices are prone to say *"let me speak to my boss and clear this problem for you."*

[33] CAP 535 of the Laws of Zambia provided for the formation of the Public Health Act which regulated all matters connected with public health in Zambia

[34] Government of the Republic of Zambia, Ministry of Health, *Governance Management Capacity Strengthening Plan, 2012.*

Implications for 3DE

The general issues around competence, and the culture of handing responsibility upwards will make the process of establishing the demand for evidence from impact evaluations a challenging one. Absent a strategic approach to all evidence (including evaluation evidence) across the Ministry, the only option is to consult as widely as possible (as 3DE currently does). The time it takes to do this consultation should not be underestimated so, given this, it may be possible to use the consultation process to identify other avenues that could be explored (e.g. different types of evaluation).

Administrative and Policy Implications of the Neopatrimonial system on the MoH

Under the neopatrimonial system of governance, the general political context tends to manifest itself in decisions and health policy terms through the following interlinked mechanisms and tendencies:

- The tendency towards centralization governance by decree;
- The lack of a strong and experienced national policy-making elite;
- The lack of a clear delineation of responsibilities for policy between ministries, agencies and various levels of government;
- The problem of promoting petty and grand corruption; and,
- Limited modernization of the medical profession.

For purposes of this evaluation we will focus on the tendency towards centralization and governance by decree and the delineation of responsibilities with particular focus on its impact on the re-alignment between MOH and MCDMCH.

Centralisation and Government by decree

The trend towards increasing the centralisation of power by the presidency brings with it a tendency toward what might be described as a government by decree. Under this system of governance, decisions and policies are formulated primarily in terms of specific decisions which are passed down for implementation in terms of decrees, orders, or administrative instructions.

The re-alignment of the MoH and MCDMCH which culminated in the hiving-off and transfer of responsibility for primary health care functions of the MoH and grafting them onto a new ministry i.e. the MCDMCH is a typical example of how decisions and policies are formulated by a small cabal of ministers and presidential advisers and passed down for implementation in terms of a presidential decree or order. Here the main drivers of policy and decision-making were a fairly limited group of ministers and presidential advisers. The decision was made without liaising fully with the Ministry of Finance which requires evidence-informed policies so as to formulate detailed instructions on how the management and funding of the two ministries should be handled.

Such centralized decision-making does potentially offer two advantages. First, it has the ability to get rapid results and second it has the ability to coordinate policy between sectors and ministries. However, experience suggests that centralized decision-making has many pitfalls and drawbacks. In the case of the realignment of the MoH and the MCDMCH, the result was what can best be described as a stalled or incomplete transition. The consequences of this incomplete change process included instances where unresolved problems or weaknesses were damaging the ability of the two ministries to deliver services efficiently and effectively where opportunities to improve service delivery were not seized.

Another disadvantage of excessive centralisation of power is that the whole area of policy may simply be overlooked or end up in backwaters if the key issues are not high on the political agenda. A frequent problem observed in centralized decision-making is the difficulty in undertaking rational prioritisation, both in terms of assessing which issues require attention and in terms of setting policies to guide subsequent decisions. Policy makers in the MoH and MCDMCH all too often see their role as making a decision and issuing an instruction to implement that decision as rapidly as possible so as to avoid incurring the wrath of the Minister or the

Permanent Secretary. Policy makers in the MoH and MCDMCH sometimes appear quite reluctant to think through a policy problem before framing possible options for action.

Implications for 3DE

Given the centralised nature of power and deferring of decisions to senior management, the Programme will also need to identify opportunities to engage with or raise awareness with senior management team of the ministry. The Directors have the ability to kick start a decision process and are a good contact point for the programme.

The combination of general inertia across the Ministry with rapid, centrally driven action has implications for how 3DE assesses progress. The political economy of each issue selected for evaluation will affect how progress could be defined – echoing an earlier point, it will not be enough to target 'policy decisions' as the primary measure of success. Where the issue is strongly centrally driven an evaluation may be able to affect policy decisions taken at a senior level. Where it is not, it may be more effective to focus on improving operational decisions and/or improving evaluative thinking and capacity more generally.

Other stakeholders involved in the policy formulation process

There are a number of other actors in the health sector that are involved to different degrees and in various stages of the policy-making process.

At the legislative level parliamentarians are important actors in the health sector. Parliamentarians enact legislation, approve government estimates of revenue and expenditure, scrutinise and oversee government administration and actions. Interests pursued include clientelist relations with various health institutions. The principal actors in the National Assembly are eight members of the Committee on Health, Community Development and Social Services. This committee studies various evidence-informed health and health related reports and makes recommendations to the Government through the House on the mandate, management and operations in the MOH and MCDMCH. The committee carries out detailed scrutiny of certain health activities undertaken by government; make necessary recommendations on the need to review certain policies and existing legislation; examine annual reports of government pertaining to the two ministries; and consider any Bills that may be referred to it by the House. Key incentives include the awarding of lucrative contracts to their friends and relatives. Parliamentarians also receive handsome mid-term and end-of-term gratuities from the state and travel extensively on government assignments.

Doctors and Nursing staff are the frontline providers of health services and are the principal implementers of policy. Due to poor remuneration, key interests are to develop rent-seeking relationships with health providers. Their numerical strength enables political patrons to bestow guaranteed employment opportunities and prospects for advanced training and development.

Health unions members' interests are wide ranging and include lobbying government on social and poverty related issues. Power and resources for influencing policy vary considerably. Overall, informal linkages through patrimonial relationships have often reduced their effectiveness in the health sector.

Cooperating Partners play a significant governance and accountability role in the health sector. Many are highly value-driven experts. Key interests include meeting MDGs and encouraging government to implement the National Health Strategic Plan. Incentives include influencing government policy in line with donor strategies.

Oversight institutions in the prevention chain include the Ministry of Finance, Tender Committees and Audit Committees. Their interests are to carry out investigations without challenging the political elite. Incentives include promotions and guaranteed job opportunities if officers concerned do not challenge elite interests.

Implications for 3DE
The 3DE approach views evaluations as tools for providing answers to the policy process. In this model, engagement with civil society and stakeholders is a method of improving the evidence that informs the answers. Yet civil society in Zambia is generally weak. A complementary model might view evaluations as tools for engaging civil society and stakeholders around an issue—the results of the evaluations being less important in terms of providing the 'right' answers than the processes via which they are conducted. A stronger emphasis on the process of the evaluations, which built evaluative thinking and capacity, might ensure that both approaches are valued and valuable—that one does not dominate at the expense of another.

Role & importance of research in health

The MoH fully acknowledges the fact that evidence-informed decision-making is the most rational and professional approach to attaining positive health outcomes. The MoH also recognises that it is impossible to achieve the national health targets enshrined in the Revised-Sixth National Development Plan (R-SNDP) and in the National Strategic Health Plan (NSHP) without evidence-based research.

The Directorate of Disease Surveillance, Control and Research (DDSCR)
DDSCR is one of five technical directorates that fall under the Ministry of Health. Its functional responsibilities are to diagnose report and notify government on any outbreak of the 11 notifiable diseases[35]; lessen the impact of epidemics in relation to mortality, morbidity, and social disruption; contribute to global and regional needs for disease surveillance; and to generate timely information for evidence-based health service delivery.

Although government has always committed itself to prioritizing evidence-based health research for improving health outcomes, the DDSCR has been a victim of the gap between political policy pronouncements and the political will to effectively fund evidence-informed health research activities. Until the National Health Research Act No. 2 of 2013 was assented to by the President on 21st March 2013, health research activities were carried out without any legal framework or guidelines to regulate and guide various research institutions and researchers in their conduct. Table 11 sums up the trends in approved Estimates of Expenditure for the Disease Surveillance, Control and Research Directorate for the period 2011 to 2015.

Table 11 Trends in approved Estimates of Expenditure for the Disease Surveillance, Control and Research Directorate (2011-2015)

Disease Surveillance, Control And Research Directorate	Approved Estimates of Expenditure ZMW'000				
Department al Programme	**2011**	**2012**	**2013**	**2014**	**2015**
01.Surveillance Control and Research	5,099	9,210	5,902	5,976	17,929
02.Environmental Health	6,267	8,302	1,417	2.009	2,358
03. Malaria Control and Research	1,398	2,014	927	515	3,823
Departmental Totals	**12,764**	**19,526**	**8.246**	**8,500**	**24,110**

Source: Government of the Republic of Zambia, Ministry of Finance, Estimates of Revenue and Expenditure for 2011, 2012, 2013, 2014 and 2015.

The Table shows that the Directorate has three major expenditure budget lines, namely Surveillance Control and Research, Environmental Health and Malaria Control and Research. An analysis of each Sub-Head indicates that despite the provisions for research, the departmental activities under each Sub-Head

[35] The 11 notifiable diseases that are vigilantly reported on are: Acute Flaccid Paralysis; Measles; Neonatal Tetanus; Dysentery; Cholera; Plague; Rabies; Typhoid Fever; Yellow Fever; Tuberculosis; and Human Influenza

do not have a research component. For example under the Surveillance Control and Research Sub-Head, the activities are predominantly of a surveillance and control of communicable and non-communicable nature. Similarly, under the Malaria Control and Research Sub-Head, the principal activities include In-door Residual Spraying, Malaria case Management and Diagnostics, Programme Management and Malaria Survey and Programme Review.

The absence of specific budget lines for evidence-informed research clearly indicates that the funding and the place for research in the health sector is still far from being satisfactory. Furthermore despite there being a Directorate for Disease Surveillance, Control and Research, the new MoH structure provides for only two officers that are charged with the responsibility for research[36] Despite having these two establishments, there is no specific research unit or section in the MoH structure. Other challenges include inadequate human and institutional capacities[37] to conduct research, disseminate results and more importantly translate results into policy and practice.

Despite the challenges listed above, the Directorate for Disease Surveillance, Control and Research has embarked on the implementation of an ambitious National Health Research Strategic Plan that will address institutional, legal, financial and capacity constraints (See Box 7).

Box 7 The National Health Research Authority

The National Health Research Act No. 2 of 2013 provides for the establishment of the National Health Research Authority (NHRA), its functions and powers. The Act also provides for the establishment, functions and powers of the National Health Research Ethics Board. In addition, the Act provides for establishing a regulatory framework for the development regulation, financing and coordination of ethically sound health research.

Government funding to the National Health Research Authority commenced in 2014 when the Ministry of Finance allocated a total of ZMW 2,717,939 for the construction of the National Health Research Authority. In 2015, the NHRA was allocated ZMW1, 500,959. The Research Unit coordinates all the health research activities in Zambia

Implications for 3DE

While research and evidence is noted as a priority of the government, this is not reflected in the budgets allocated to it. This has resulted in limited staffing or budget for conducting any research, and a general lack of evaluative thinking that is needed to take a strategic approach to identifying and managing impact evaluations, and interpreting their results. This lack of evaluative thinking may, in the long term, have a significant effect on whether the 3DE programme has made lasting changes.

[36] The two officers are the Principal Surveillance and Research Officer and the Senior Surveillance and Research Officer.
[37] Under the current institutional arrangements, there is a provision for the National Health Research Advisory Committee (NHRAC) whose overall responsibility is to advise the MoH on all matters related to health research in Zambia. The NHRAC secretariat is considered weak in that it does not have a specific office nor does it have operational funds.

Annex G Literature review

G.1 Introduction

There is a rich literature exploring the role of evidence in shaping policymaking and the factors associated with evidence uptake. Reviewing this literature is crucial for understanding the rationale behind the design of the 3DE pilot and assessing the likely validity of the propositions contained in the Theory of Change (ToC). This review will examine the ways that evidence can be applied to policymaking and the factors that promote or constrain its uptake, taking into account the characteristics of evaluations, evaluators and the political climate. It will then briefly outline some potential lines of enquiry that emerged from this literature in relation to testing the 3DE ToC. The final section will explore the experience of some existing international initiatives aimed at promoting evidence based policymaking and building government capacity to generate and use evidence.

G.2 To what extent and in what ways can evidence influence policymaking?

Evidence can shape policy decisions in a variety of ways. There is often primary interest in understanding whether and under what circumstances evidence directly contributes to a policy decision. Yet as Johnson (2009) describes, the application of evidence to inform concrete policy decisions (such as scaling up, discontinuing or redesigning and particular programme) is only one aspect of its possible influence. In addition to this 'instrumental' role, evidence use in decision making may also be classified as symbolic or conceptual (Johnson et al 2009). Symbolic use is when the mere existence of an evaluation report, rather than its content, is used to justify a policy decision that would have been taken nonetheless. Conceptual use means that the evaluation shifted perceptions or added to knowledge in some way without leading to any definitive policy decisions. These different ways of using evidence may occur at different points in the policy cycle (Johnson et al 2009; Sutcliffe and Court 2005), or depend on the particular kind of evidence that is available.

The literature suggests that the extent of evaluation uptake in policy is often relatively modest. A European Commission study on the effects of knowledge generated by EuropeAid's strategic evaluations found that although there are some notable cases of the information being used to inform distinctive policy choices or raise conceptual understanding, findings do not tend to be incorporated into decision making at an institutional level (Bossuyt et al 2014). This conclusion was echoed by a comprehensive report on relevance, quality and influence of impact evaluations conducted by the World Bank Group. While impact evaluation evidence was observed in some cases to make a positive contribution to development practice and policy debate, the study found that systematic use of evidence was weakened by a number of constraints (Mackay 2007).

Both reviews note that detecting or measuring the extent of evidence uptake is challenging, especially given that some of the broader level influences of knowledge may be somewhat intangible[38]. Yet much can still be said about the factors that may affect the likelihood that relevant evidence is applied to policymaking in a rational way. These factors can be categorised as characteristics of the evaluation, characteristics of the evaluation user and wider contextual factors (Johnson et al 2009).

G.3 Characteristics of the evaluation relevant to evidence uptake

One of the primary findings to emerge in the literature is that evidence produced by evaluations is not always relevant to the practical requirements of policymaking (Oliver et al 2014; Bossuyt et al 2014; Rutter 2012; World Bank 2012). In order to be relevant, information presented to policymakers should address identified policy needs, deliver clear recommendations and pay close attention to political and contextual

[38] See also: http://blogs.worldbank.org/publicsphere/what-evidence-evidence-based-policy-making-pretty-thin-actually

factors. Yet it is consistently noted in many papers that evaluation evidence can appear to cater more to a research audience than the practical needs of policymakers. It may be more focused on what happened in interventions rather than why the results arose, deliver recommendations that have many technical caveats or embed policy relevant content in extensive discussion of methodological points of interest (Bossuyt et al 2014; Rutter 2012). Consideration of the political climate or the likely resource requirements of implementing recommendations may also be overlooked, according to the review of five case studies undertaken on behalf of the Centres for Learning on Evaluation and Results initiative (CLEAR) in 2014. The outcome is that policymakers often struggle to draw lessons from existing evaluation work or to locate evidence that meets their information needs.

Secondly, evaluation evidence appears often not to be made available at the time when policy decisions need to be made (Bossuyt et al 2014; Oliver et al 2014; Rutter 2012). The rapid decision making that may be required by political calendars is incompatible with the kind of in depth evaluation recommendations that may take several months or years to produce (Sutcliffe and Court 2005). This means that the window of opportunity to influence policy is often effectively missed, and by the time results are available evaluations may have lost much of their relevance to current policy issues.

A third salient factor identified in the literature is that empirical evidence is not well communicated to policymakers. This problem has two dimensions. On the one hand reports may not be adequately disseminated to policy audiences. The EuropeAid report notes that key stakeholders are sometimes not even aware of the existence of evidence that is relevant to their policy area (Bossuyt et al 2014). Yet even if results are made available, the way that they are communicated in reports may be excessively technical and therefore not useful to key stakeholders (Work Bank 2012; Hyder et al 2010). This is linked to the finding that the kind of evidence generated by evaluations is not sufficiently relevant to policymakers.

Other factors related to evidence uptake include the quality of the evaluation in terms of how robust and credible its findings are. As noted in the seminar series held by the Institute For Government in 2012 on issues surrounding evidence-based policymaking, some of the policy questions which governments are concerned with are not well suited to the most rigorous evaluation techniques (Rutter 2012). This complicates the design of suitable evaluation methods and means that attaining credible results in policy evaluations may be challenging. Credibility is also undermined if government ministries are not able to provide usable data relating to particular interventions.

The failure of evaluators to deliver high quality, timely, policy-relevant and appropriately communicated findings to policymakers points to a wider concern that the priorities of the producers and users of evaluation are not closely aligned. This is in part due to the fact that evaluations in lower and middle income countries tend to be conducted by development organisations (CLEAR 2013 and 2014). There is some evidence that this is changing in line with the recommendations of the Paris Declarations, which call for greater in-country ownership of M&E efforts. Yet although there are some notable and important examples of evaluations being managed internally by government ministries with designated M&E oversight, the study on supply and demand for evaluation evidence undertaken on behalf of CLEAR finds that in country M&E work remains mostly limited to performance monitoring rather than evaluation (CLEAR 2014). As a result evaluation evidence may often be generated by independent researchers who do not have policy concerns at the forefront of their agenda, and instead prioritise research objectives such as obtaining publication in peer reviewed journals (Oliver et al 2014). The perception that evaluations are not produced with the priorities of policymakers in mind appears to be widespread and is frequently cited in the literature.

G.4 Characteristics of evaluation users relevant to evidence uptake

Guidance produced by the World Bank Group on strengthening government capacity in generating and using M&E evidence argues that the importance of the evaluation supply issues described above is far

outweighed by the extent of evaluation demand (Mackay 2007). A prerequisite for the actual use of evaluations in policy is that key political actors demand evidence to be made available and are receptive to the findings.

There are mixed findings in the literature on whether this demand exists in different contexts. Hyder et al. find that policymakers in their sample actively value the evidence produced by evaluations (Hyder et al 2010). CLEAR also note in their study of M&E systems in nine Sub-Saharan African countries that there are promising and increasing indications of evaluation demand by governments and civil society (CLEAR 2013). Yet many other papers suggest that the limited relevance of some evaluations to policymakers appears to have reduced faith that this type of evidence can be worthwhile at all. The study of EuropeAid strategic evaluations found that policymakers often did not read reports, perceiving evaluation evidence not to be useful or connected to their work (Bossuyt et al 2014). This is confirmed by the World Bank guidance paper, which reports that low demand has been a commonly encountered obstacle across the body of the Impact Evaluation Group's work (Mackay 2007).

Weak demand for evidence is partly to do with issues around limited relevance, timeliness and poor communication of evaluations outlined above. But there are also important causes unrelated to the supply of impact evaluations. In the first place, evaluation evidence may simply be less useful to policymakers under some circumstances than other kinds of knowledge. Additional evidence sources that policymakers may demand include the accumulated experience of stakeholders and institutions, and the knowledge of citizens about their own policy needs (World Bank 2012; Jones et al 2009). A balance of different sources of evidence may be required to develop good policy, and it is not the case that research evidence is self-evidently superior.

Another potential cause of limited demand for evaluation evidence is that policymakers themselves are not sufficiently skilled in evaluation methods. The effect of technical skill in raising demand is twofold. Firstly it can help address the low demand for evidence that is caused by limited awareness of the potential value of evidence. Mackay (2007) describes the problem of weak demand as having a recursive aspect: it is linked to policymakers having weak understanding of evaluation methods, caused in turn by little previous experience with M&E, which itself is an independent source of low demand (Mackay 2007). The potential benefits of raising awareness of M&E is confirmed by a CLEAR report finding that demand for evaluations by policymakers is often ‘latent’. Latent demand means that policymakers do want information to support their decisions but don’t recognise that evaluations can be a source of this evidence (CLEAR 2014). Raising awareness can therefore help to break the cycle of little understanding and experience of M&E that reinforces low demand.

Improving technical capacity in evaluation tools may also raise demand for evidence by improving the ability of policymakers to understand and interpret evaluation findings. Given that evaluation findings are often presented in an academic way, some papers argue that policymakers may require a degree of technical skill to help them understand, interpret and ultimately apply empirical information. Oliver et al. report that policymakers themselves expressed a need for support in building their own knowledge to help them make use of evaluation evidence (Oliver et al 2014).

Policymaking that is not evidence based may instead be driven by the values and beliefs of individual decision makers (Sutcliffe and Court 2005). Their assumptions can be difficult to overturn, particularly where new information contradicts a strongly held ideology (Mackay 2007). The propensity of policymakers to rationally apply evidence to policy issues also depends crucially on political calculations and contextual factors. Evidence may be disregarded or even concealed if it is not consistent with a particular political calculation or threatens the interests of powerful groups (Jones et al 2009). The political determinants of evidence uptake are described in further detail in the following section.

G.5 Contextual factors relevant to evidence uptake

Beyond the immediate characteristics of evaluations and evaluators, the literature also emphasises the role of political and institutional factors in shaping the way that evidence is used in policymaking. The central argument of the CLEAR study on supply and demand for evaluations is that if evaluations explicitly engage with the political economy and respond to an active demand for information, the evidence will be used (CLEAR 2014). The importance of engaging with characteristics of the wider political environment is echoed throughout the literature. There are some acknowledged challenges associated with assessing political economy factors. Political systems are complex and difficult to characterise, and Liverani et al. (2012) note that there are several gaps in current understanding of the implications of different political systems for evidence uptake. However there have been some recent advances in developing knowledge in this area. The same review synthesises the findings of recent work on the use of evidence in public health policy to draw some overarching conclusions. Studies undertaken on behalf of CLEAR and the ODI which emphasise political economy considerations provide some clarity in ways to map different political systems, taking into consideration where the balance of power in the system rests, which checks and balances on political actors exist, what incentives and constraints shape the behaviour of key actors and what the nature of formal and informal participation in politics is (CLEAR 2014;Jones et al 2009).

Among the findings of this developing part of the literature is that the distribution of decision making power across the political system has a crucial effect on the opportunities for evidence uptake. Decentralised systems in which many actors have a stake in guiding policy are viewed in some studies to be associated with greater use of evidence to support processes of policy contestation. The ability to marshal evidence becomes important as a way to secure support for particular policy positions or undermine competing views (Liverani et al 2012). In a related point, some studies argue that higher levels of government accountability observed in mature democracies can lead to increased evidence use since policymakers face pressure to demonstrate and justify the basis on which decisions are taken (Nabyonga-Orem et al 2012).

Yet although the existence of political accountability and platforms for policy debate may create potential for evidence uptake, the quality of that evidence use can still be poor. Instead of applying evidence to policy problems in an impartial way, political actors may behave opportunistically by selecting evidence purposefully to back up pre-existing policy positions or presenting findings in a misleading way (Liverani et al 2012; Jones et al 2009). One of the issues discussed in the IFG seminar series on evidence-based policymaking was that where accountability to electorates and civil society is strong, the motivation to demonstrate the evidence basis for policy choices may be outweighed by the pressure to meet public perceptions that go counter to new knowledge, or fulfil election promises (Rutter 2012). Overall the literature implies that there is no system of government that reliably ensures rational evidence-based policy and that policymaking is inherently political (Sutcliffe and Court 2005). Yet the nature of politicisation varies between systems, and therefore understanding the incentives and constraints facing policymakers in different contexts requires an appreciation for the power and decision making structures that make up the political environment .

Several papers also identify features of individual government ministries that are relevant to evidence use. Liverani et al. (2013) report that high divisions of responsibility within individual bureaucracies can reduce the ability of ministers to engage with evidence that falls outside their immediate area of work. A high rate of staff turnover is also found to lower the potential for critical engagement with new evidence by shortening the ‘institutional memory’ of the department, causing existing practices to appear novel.

As described above, the demand for evidence and ability to apply it to relevant policy problems may be associated with policymaker knowledge of and experience with evaluation techniques. Capacity to undertake M&E is in part determined by the provisions made for it in the political system, including the budget allocated for evaluation, whether there are designated ministries with responsibility for M&E activities and the existence or functionality of relevant guidelines or national plans. The governments of Zambia and Uganda have both made provisions for internal capacity building and investment in M&E, through a series of National Development Plans and the former Poverty Eradication Action Plan (1995-

2010) in Uganda. In Zambia the M&E system is relatively mature and considered by at least one study to be among the most comprehensive in Sub-Saharan Africa (CLEAR 2013). But in both countries inefficiencies, shortages of skills and lack of clear mandates for different bodies affect the progress of M&E capacity building (CLEAR 2013).

G.6 What are the implications of the literature for testing the ToC?

The central ideas articulated by the ToC receive support from the literature. There is general consensus that evidence uptake can be raised by greater integration between the research and policy worlds and increased engagement with features of the political and institutional climate. However the literature also makes clear that there are some substantial barriers to the use of evidence in policymaking. This suggests some aspects of the ToC that might warrant elaboration and testing to ensure that the sequence of events expected as a result of the 3DE Pilot is accurately described and the underlying assumptions are valid.

Assumption 1: Impact evaluation evidence is useful for policymakers in the sense that if recommendations are implemented, there will be improved development results.

The 3DE ToC assumes that the evidence generated by the impact evaluations will, if implemented, lead to improved outcomes. However one lesson from the literature is that research evidence is only one of the kinds of evidence that policymakers may incorporate into their decisions, and it is not self-evident that research based knowledge is superior. Policymaking that is highly responsive to technical evidence and does not respond to informal sources of knowledge or citizen knowledge may risk becoming bureaucratic and unresponsive to citizen demands. Even within the category of research based knowledge, it is not clear that randomised controlled trials as planned under the 3DE pilot are necessarily the most valuable from a policy perspective in every case. This suggests that the 3DE pilot may be best viewed as a complement to other kinds of evidence that are relevant to policymaking.

Assumption 2: Is there demand from policymakers for evidence?

The literature produced mixed results on the question of whether policymakers demand evidence from evaluations to assist them in decision making, but confirmed that this is an essential prerequisite for new knowledge to be translated into policy. It appears that demand for evidence varies substantially depending on individual policymaker attitudes, perceptions about the usefulness of evaluation evidence and credibility of the evaluator, awareness of evaluation benefits, technical skill in evaluation methods and the nature of the political system. Certainly the presence of demand cannot be taken for granted and would need to be assessed on a case by case basis. It may sometimes be necessary to motivate demand for evaluation evidence through various strategies, such as the carrots, sticks and sermons described by Mackay (2007) or the capacity building approach of the CLEAR initiative.

G.7 Summary of international, government initiatives on capacity building around use and demand for evidence.

Much of the literature described above argues that increased integration of policymakers into the evaluation process can help to promote evidence uptake. Bringing policymakers into the evaluation process is a way of aligning the incentives of policymakers and researchers, which should in turn create evaluations that are more relevant to policymakers, better communicated and produced in time for results to be incorporated into decisions. In addition, policymaker involvement in evaluation may increase the ownership that stakeholders feel over the results, which itself can raise the incidence of evidence uptake.

There are a number of existing initiatives which aim to strengthen the use of M&E information in policy by applying these lessons. These are outlined in detail in the following section.

G.8 International initiatives

Regional Centre for Learning on Evaluations and Results (CLEAR)

Description of the initiative

CLEAR is a collaborative, global partnership, established in 2010, that works to strengthen partner countries' capacities and improve systems for monitoring and evaluation (M&E) and performance management (PM), to guide evidence-based development decisions (CLEAR Mid-Term Evaluation Inception Report, February 2014).

CLEAR is expected to promote replication of high-quality locally or regionally delivered capacity development services involving government agencies as well as civil society, and inspire such efforts globally. The programme's goal is to be achieved by

- stimulating demand for M&E capacity, through outreach and awareness building and developing and delivering innovative, responsive, contextually relevant, and cost-effective services, and
- learning from, documenting, and sharing experiences and knowledge gained from the development and delivery process.

Activities to date

CLEAR states that it brings together selected and recognized academic institutions or think tanks with other organisations, such as foundations and multilateral and bilateral organisations, in a global knowledge and monitoring and evaluation (M&E) capacity development delivery partnership. The academic institutions and think tanks house the Clear Centres, while the Independent Evaluation Group (IEG) of the World Bank group hosts the programme's global hub. (www.theclearinititive.org)

Typical stakeholders involved with CLEAR are parliament, ministries, government agencies, civil society groups, NGOs as well as academic and other research institutions. These include individuals and bodies/teams on the executive, managerial and technical/professional level (Report on Building Blocks of CLEAR's Capacity Development Strategy, February 2013).

Examples of international/supranational partners include the African Development Bank (AfDB), the Australian Department of Foreign Affairs and Trade (DFAT), the Asian Development Bank (ADB), the Belgian Federal Public Service for Foreign Affairs, the International Trade and Development Cooperation, the Inter-American Development Bank (IDB), the Rockefeller Foundation, the Swedish International Development Cooperation Agency (SIDA), the Department for International Development (DfID), the Swiss Agency for Development and Cooperation (SDC), the World Bank and the Independent Evaluation Group (IEG).

How it works

Six regional centres in Africa, East and South Asia and Latin America make up the backbone of the CLEAR programme. Each centre is hosted by a competitively selected academic institution, and provides innovative monitoring and evaluation (M&E) and performance management (PM) capacity-building services across each region. CLEAR's governance structure encompasses the Board, the Secretariat, and Regional Advisory Committees (RAC) established by each of the Regional Centres. The Secretariat is housed in the Independent Evaluation Group (IEG) of the World Bank in Washington DC (www.theclearinitiative.org).

The CLEAR centres work to enhance and foster demand for M&E and PM, strengthen organisational capacity to produce and use evidence, build critical professional expertise and lead innovation in M&E and

PM. CLEAR's global learning objective is anchored by the regional centres, which generate innovative knowledge of and approaches to capacity development, and facilitate peer learning, and mentoring across regions on what works, what doesn't, and why. Approximately 80 percent of the budget is devoted to Regional Approach and 10 percent to Global Learning. The remaining 10 percent is used for programme governance and management, including regular monitoring and reporting and mid-term and final evaluations (CLEAR Strategy (2013-2018), September 2013).

The CLEAR Strategy (2013-2018) further highlights the following factors conducive to the development of solid M&E approaches:

Ownership by clients. CLEAR represents a gravity shift toward partner country ownership by enabling the Centres to drive capacity development in partnership with their clients and customized for relevance to country contexts.

Branding through excellence. By ensuring high quality and standards, CLEAR allows institutions not only to build the CLEAR brand but also to establish their own reputations for excellence in M&E capacity development.

Sustainability. By strengthening the institutional capacity of competitively selected and well regarded academic institutions, CLEAR focuses on building sustainable in-region capacity to build capacity. CLEAR requires the Centres to develop a business model that is ultimately self-financing.

Partnerships to reduce fragmentation. CLEAR changes the way in which countries and donors work together. It catalyses collaboration across the globe and reduces costly fragmentation of support for evaluation capacity development by working in partnership with ten donors and five institutions with their partner academic institutions and in-country financial supporters.

Activities to date

Numerous activities have been pursued since the programmes inception in 2010 and are planned to be done by its end in 2018 in line with the three main phases of development (CLEAR Strategy (2013-2018), September 2013). The first phase comprised the centre selection (2010-2013). This included the selection of regional institutions to house the CLEAR centres, establish a functioning governance structure and operational procedures. The second phase focused on strengthening regional and global capacity (2012-2014). This included building demand for their services, refining their regional strategies and further establishing technical and management capacities. The third phase emphasises the creation of regional and global sustainability (2013 and beyond). This includes the development of ongoing engagement with clients and constituents in key government agencies and civil society organisations and networks. The Centres draw on local, regional, and global innovations through peer-learning in their network to better meet the needs of developing country constituents, while combining quality, depth, and practicality in their work programmes. The emphasis of the programme will gradually shift from an "incubation" and "seed investment" to a sustainable decentralized capacity development model.

Various documents have been written in support of the activities in the phases of development. The most notable ones are mentioned below.

The building blocks of CLEAR's Capacity Development Strategy have been defined focusing on change agents, capacity outcomes as well as M&E Capacity Development Activities (Report on Building Blocks of CLEAR's Capacity Development Strategy, February 2013).

CLEAR's overall strategy for the period 2013-2018 has been formulated which addresses challenges, vision & mission, phases of development, governance and other items (CLEAR Strategy (2013-2018), September 2013). This document gives guidance to CLEAR's ongoing activities.

An Inception Report has been written outlining the stakeholder expectations, highlighting the implications for the evaluation methodology, describing methodology and how the evaluation will be managed (CLEAR Mid-Term Evaluation Inception Report, February 2014).

CLEAR's Theory of Change has been developed which is based on a dynamic learning-by-doing model where regional and global knowledge is shared among the network to enhance overall learning regarding M&E (CLEAR Theory of Change, June 2013). Necessary capacity is built through strategically engaging in mentoring, training, leadership development, advocacy, grants/awards, collaboration and well as technical and managerial assistance. This strengthens M&E systems and practices which in turn assist stakeholders making decisions conducive to improved development results.

Outputs produced/Results achieved

Results achieved are summarised by distinguishing between the overall programme level, the regional level as well as the global level. First, the following results at the overall programme level are discussed in the CLEAR Midterm Evaluation report (October 2014):

The CLEAR Theory of Change and Results Framework provided guidance for the establishment of regional centres but were less useful for testing key assumptions, promoting learning within and across CLEAR units, and for assessing progress towards envisaged development results. To date, neither CLEAR overall, nor each of the regional centres has defined what "success" in development terms would look like at global or regional/national levels.

The CLEAR Secretariat has effectively fulfilled its assigned roles, has provided administrative support to the functioning of the initiative, and has provided leadership and guidance for the regional centres. The location of the Secretariat in the World Bank's IEG has both advantages and disadvantages; relocating the Secretariat during the current phase of transition would likely pose more challenges than potential benefits.

The CLEAR Board has fulfilled its three assigned roles with varying degrees of success. It provided effective leadership on operational matters but less guidance on the questions and issues emerging as a consequence of CLEAR's experimental design, or on longer-term strategic decisions on the future of CLEAR. The current Board composition lacks diversity in regional representation, experience and expertise, which limits its legitimacy in the eyes of stakeholders. Making changes to the composition of the Board (or the addition of a Steering Committee with diverse membership) could address some of these issues, but would not automatically solve the noted gaps in leadership for guiding an experimental initiative.

Second, the following results at the overall programme level are discussed in the CLEAR Midterm Evaluation report (October 2014):

In terms of design, the internal and external contexts of the five reviewed regional centres varied considerably. This was not sufficiently accommodated in the programme design and resulted in lost learning opportunities. The CLEAR regional centres are in relatively early stages of developing their own strategies and do not yet have a clear, appropriate basis for measuring "success" in terms of development results.

In terms of capacities, regional centres have varying levels of institutional capacity, which in some cases limits their potential to make the kinds of contributions envisaged in the CLEAR design. With the exception of the Latin America centre, CLEAR regional centres have to date established relatively few strategic, longer-term linkages with regional partners and other like-minded institutions. Affiliations with their respective host institutions have affected regional centres in different ways, due to structure and administrative requirements, but overall these relationships have enhanced the credibility of regional centres and have provided access to potential clients and partners. With the exception of the centre in South Asia, progress towards establishing Regional Advisory Committees has been slow, depriving most centres of relevant and regionally grounded strategic advice.

In terms of performance, CLEAR objectives and activities are considered relevant to the M&E needs of government and non-government stakeholders. All five centres have met most of the midterm targets, which focused on the establishment of centres and their ability to provide a variety of capacity building services for M&E and RBM. Almost all centre achievements to date relate to creating favourable conditions that – in the longer term – have the potential to contribute to individual actors or organisations producing (and eventually using) more or better evidence, but in keeping with the programme's mid-term status, there is limited evidence of their contribution to these higher level envisaged results. The likelihood that regional centres and their services will continue without CLEAR funding varies – from low in Anglophone Africa to very strong in Latin America and South Asia.

Third, the following results at the overall programme level are discussed in the CLEAR Midterm Evaluation report (October 2014):

CLEAR stakeholders and beneficiaries value many elements of the global learning component, such as the Global Forums, CLEAR training modules, and the Secretariat's support to regional centres. However, it is difficult to assess the effectiveness of this component as CLEAR has not yet articulated a global component strategy nor its desired results. At midterm, CLEAR units are still experimenting with ways and areas of collaboration, and the regional centres have shown varying degrees of interest in and capacity to engage in mutual knowledge exchange and related efforts. Overall, the global learning component has not yet realised its potential as CLEAR has not harvested the knowledge, lessons and evidence emerging from the CLEAR experiment. This is a missed opportunity.

Challenges faced

Several challenges are mentioned in the CLEAR Midterm Evaluation document. First, the original evaluation TOR (check details) and methodology did not adequately address the experimental nature of CLEAR which led the evaluation team to review data and reformulate findings to better reflect the experimental nature of CLEAR. Second, the organisational network analysis (ONA) was removed as a line of evidence due to the limited and inconsistent survey responses. Third, the evaluation TOR indicated an interest in how the cost of CLEAR centre services compared to those of other capacity building providers in the respective regions. Despite its efforts, the evaluation team found insufficient data to pursue this line of evidence.

Further to that, CLEARS continues to address the challenges as noted in the CLEAR Strategy document. For instance, the capacity for developing and implementing contextually appropriate monitoring and evaluation (M&E) and performance management (PM) approaches varies across countries and remains weak in many. Further, at the national level, concerns regarding equity and effectiveness of development programmes have fuelled citizens' and civil society's demand for transparency, access to information, and accountability for results. In addition, a wide range of international and national nongovernmental organisations have strengthened their own M&E/PM, learning and accountability capacities, among them Oxfam, PACT, and BRAC. But their reach is limited to their own constituencies, and they are often not connected with the larger national and regional institutions to scale up their activities. Further, some countries are not implementing M&E/PM well enough to produce systematic and robust evidence and some and have not advanced toward linking evidence to decisions. Moreover, thoughtful and knowledgeable professionals and an appropriate range of services to build government, civil society, and philanthropic capacity to monitor and evaluate is still relatively limited.

Building Capacity to Use Research Evidence (BCURE)

Description of the initiative

Building Capacity to Use Research Evidence (BCURE) is a programme that aims to build the skills, knowledge and systems that will allow policy makers and practitioners in low and lower middle income countries to access, appraise and use rigorous evidence. Starting in 2013, the BCURE programme brings

together a set of six strategically linked projects[39], spanning across Sub-Saharan Africa and South Asia to improve development interventions through better decision making processes (https://bcureglobal.wordpress.com).

Who is involved

Building Capacity to Use Research Evidence (BCURE) is a programme of work funded by the UK Department for International Development (DFID). The following partners are involved with the BCURE programme: Adam Smith International (ASI), African Institute for Development Policy, ECORYS, Harvard University, University of Johannesburg and INASP who oversee the VakaYiko Consortium (https://bcureglobal.wordpress.com).

The VakaYiko Consortium (http://www.inasp.info) is a three-year project involving five organizations. The project starts with the understanding that the routine use of research to inform policy requires at least three factors to be in place:

- Individuals with the skills to access, evaluate and use research evidence
- Processes for handling research evidence in policy making departments
- A wider enabling environment of engaged citizens, media and civil society

The consortium works to build capacity at all three levels. Consortium members are Ghana Information Network for Knowledge Sharing (GINKS), Zimbabwe Evidence Informed Policy Making Network (ZEIPNET), the Human Sciences Research Council (HSRC), the Overseas Development Institute (ODI) and INASP. This project is funded by DFID under the Building Capacity for Use of research Evidence (BCURE) programme.

DFID has made significant investments in building research capacity in low and middle income countries through other programmes that BCURE stakeholders should be aware of the following initiatives (https://bcureglobal.wordpress.com):

- Development Research Uptake in Sub-Saharan Africa (DRUSSA): The DRUSSA programme aims to improve the accessibility, uptake and utilisation of locally contextualised development research evidence on climate change and environment, health, information, education, governance, food security, and livelihoods to inform sub-Saharan and global development policy and practice. Policies underpinned by sound research, systematic evaluation and impact assessment, and demonstrable Research Uptake can lead to scientifically based interventions and programmes for poverty reduction and improved quality of life for Africa's children, women and men (http://www.drussa.net).
- Global Development Network (GDNet): The GDNet program was GDN's knowledge service which supported Southern researchers to contribute and debate ideas in development for over a decade. The GDNet program formally closed in June, 2014 (http://www.gdn.int)
- Synthesising Research and Knowledge Systems (SRKS): SRKS works with an international network of researchers, editors, publishers, librarians, ICT professionals and policy-makers to ensure that the research communication cycle works effectively. The programme builds on the Programme for the Enhancement of Research Information (PERI) model by building stronger, higher quality and durable research and knowledge systems. This involves activities such as librarian skills training, supporting emerging researchers in preparing research for publication, collaborative licensing and purchasing of digital library resources using locally sourced funds, and provision of cost-effective

[39] No specific information publically available on these projects.

national and regional Journal Online networks (EVIDENCE INTO ACTION TEAM PROGRAMME GUIDE: A guide to programmes funded by the Evidence into Action Team (2014)).

- Global Open Knowledge Hub (GOKH): Three key services - British Library for Development Studies, Eldis online portal and BRIDGE gender services – which make research more available, accessible and re-usable. There is particular emphasis on technical tools and innovations that enable southern partners to source and upload open access content into the Hub through use of open source technology and to draw out relevant content and present it on their own websites in forms suited to their own audiences and contexts (EVIDENCE INTO ACTION TEAM PROGRAMME GUIDE: A guide to programmes funded by the Evidence into Action Team (2014)).

- Services available Eldis internet-based information service: filtering, structuring and presenting development information primarily via the web and email.

- Eldis Communities brings together development professionals to augment their networks, and strengthen their thinking and practice through exchange and dialogue online.

- British Library for Development Studies (BLDS) is the largest collection of materials on social and economic development in Europe, with over 200,000 print titles on a comprehensive range of development themes, many of which are unavailable in either European or US libraries.

- BRIDGE supports gender advocacy and mainstreaming with print and online information services, sharing development research, policy and practice.

In addition to the in-country focus BCURE also supports two pan-Africa networks. The Africa Cabinet Government Network is supported by ASI and works with Cabinet Secretaries in 12 different countries to improve the use of evidence in government decision making. The Africa Cabinet Government Network (ACGN) has been established to provide formal and informal opportunities for collaboration and mutual support between Cabinet Secretaries and others involved in managing Cabinet processes in Africa. Among the reasons is a desire to share knowledge and experience on steps taken to improve the procedures and capabilities of Cabinet Secretariats, especially to improve evidence-based policy-making (http://www.cabinetgovernment.net).

The Africa Evidence Network is a community of people who work in Africa and have an interest in evidence, its production and use in decision-making. The Africa Evidence Network is a community of people who work in Africa and have an interest in evidence, its production (in particular but not exclusively through systematic reviews) and use in decision-making. We include researchers, practitioners and policy-makers from universities, NGOs and governments. Our members include those who work with the Joanna Brigg's Institute, the Campbell Collaboration, the Cochrane Collaboration, the EPPI-Centre, the Collaboration for Environmental Evidence and others (http://www.africaevidencenetwork.org).

How it works

As stated in the BCURE Newsletter November 2014, BCURE is being delivered through a consortium of different partners (see above), with a specific focus on building the capacity and skills of locally based organisations in the countries where the programme is currently operating.

Each of the six BCURE projects is being delivered by a partner / primary provider who will oversee the development of skills and capacity in key decision making institutions which are central to policy and practice in that country.

BCURE is being delivered with a specific focus on building the capacity of locally based organisations in the low and middle income countries where projects are operating. Each BCURE project has a primary provider, who will oversee the development of organisational systems and incentives and the skills of individuals in key decision making institutions which are central to policy and practice in that country.

Activities to date

Several activities are mentioned on the BCURE website, including the following examples:

- BCURE Commences Work with Malawian Local Government: During a recent visit to Malawi, members of the University of Johannesburg (UJ)-BCURE team took part in a workshop with representatives from the Ministry of Local Government and Rural Development (MLGRD). The workshop was organised by the Parent and Child Health Initiative (PACHI), UJ-BCURE's implementing partner in Malawi. Key government officials such as the Director of Planning in the MLGRD, the Head of the Planning, Monitoring and Evaluation Directorate and representatives from various districts attended. The workshop was part of UJ-BCURE's needs assessment process, following a landscape review. The aim of the workshop was to develop a programme of work that will focus on strengthening evidence-informed decision-making in four districts in Malawi in 2015 and 2016.

- Evidence-Informed Implementation workshop held in Pretoria: The first UJ-BCURE workshop was attended by senior colleagues from the Departments of Science and Technology, Basic Education, Planning Monitoring and Evaluation and the Strategic Relations department at the University of Johannesburg. The allowed participants to apply the frameworks for evidence-informed decision-making in the Implementation Plan for the e-Education White Paper for the 2015-2016 period. A situational analysis was done, and the Implementation Plan's logical framework approach re-worked. UJ-BCURE's continued support to the Department of Basic Education will see the completion of the exercise.

- BCURE short films to explain the work of BCURE: The University of Johannesburg have recently launched three short films that give an outline of their projects work in Malawi and South Africa. This helps to provide an understanding of the long term goals of evidence informed policy making. All three of the films can be found on YouTube, each taking a slightly different perspective on the work of BCURE (Evidence Into Policy, Bringing Evidence and Policy-Makers Together and Bringing Communities of Practice Together).

Outputs produced/Results achieved

The following outcomes are in the process of being delivered (BCURE Newsletter, November 2014)

- Development of the African Evidence Network, a sustainable and engaging community for policy makers and practitioners to discuss and share lessons on evidence use

- Focusing on the high level decision making process to improve evidence use by Cabinet Ministers

- Working with civil service training programmes to incorporate the use of evidence into the curriculum

- Developing research and evidence frameworks to encourage rigorous use of evidence in policy areas

- Establishing open policy dialogues between government and the research community to promote the use of evidence in decision making

Overall, DFID believes that project teams have understood the various policy making landscapes and what is meant by policy making and evidence in these different contexts (BCURE Annual Meeting: Getting to the Heart of the BCURE Programme, December 2014).

Challenges faced

Several challenges have been discussed (BCURE Annual Meeting: Getting to the Heart of the BCURE Programme, December 2014). It is commonly believed that reforming any political institution (in the case in terms of M&E) is challenging and that experience shows that many reforms will lose momentum eventually. Further to that, questions about longevity and sustainability are emerging in connection with the impact of current BCURE programmes and how the generated momentum can be maintained.

Critical buy-in from senior leaders and the necessary level of ambition for reform among staff in political institutions sometimes appear to be hard to obtain. This is connected with the issue of incentive setting in favour of the objectives of the programme. BCURE partners actively try to think of realistic and influential incentive structures in the context they are working in. This is mainly done through concrete expectations such as formal procedures around evidence use and/or through building a conducive organisational culture and leadership.

DFID hopes that by the end of 2015, the project teams will know in more detail what is working well, what isn't and why (BCURE Annual Meeting: Getting to the Heart of the BCURE Programme, December 2014).

Regarding the above-mentioned organisational culture and leadership, the (EU paper…) provides more input. It states that organisational characteristics indicate the weight the rest of the organisation gives to the evaluation evidence. This is linked to the existence of the learning culture of the organisation. The location of the Evaluation Unit and its reporting lines to the rest of the organisation partly determines the strength of the messages from the Unit. So do its functions - whether it is mandated to conduct strategic or project-level evaluations, and whether it is focused on accountability or learning. The organisational leadership sets the tone for how knowledge is used and transferred within the organisation. Finally, organisational links between results based management (RBM) processes and evaluations help define relationships between project and strategic evaluations.

In addition, the paper mentions that the nature of evaluation policies sends signals about the types of evaluation activities that are favoured and how they harmonise with other policies and programming priorities. Other incentives include the levels of resources allocated to the evaluation function; the commitment to management responses to evaluations and ongoing follow-up. A low profile response from management to following up the recommendations from evaluations will ripple through the organisation in diverse ways. These include the wider influences on an organisation which affect its decisions about how to scope, source and apply the evidence from evaluations. This may include wider political aspects, the country context within which an evaluation is being conducted, or broad debates around key issues such as methodology. These influences are likely to come from a variety of different government and non-government sources.

Moreover, the paper also discusses the sense of ownership as an important factor. Individuals and teams who feel a sense of ownership over the results of the evaluations will be more likely to take up and use the evidence they produce. Ownership comes from having a real interest in what the evaluation determines. This is more likely if the teams or individuals have been involved in setting the questions which need answering. Developing a sense of ownership therefore begins at the design phase.

International Initiative for Impact Evaluation (3ie)

Description of the initiative

3ie is an international grant-making NGO promoting evidence-informed development policies and programmes. We are the global leader in funding and producing high-quality evidence of what works, how, why and at what cost in international development. We believe that better and policy-relevant evidence will make development more effective and improve people's lives. Since its founding in 2008, 3ie has awarded

over 200 grants (146 impact evaluations, 33 systematic reviews and 38 other studies) in over 50 countries, with a total value of US$84,225,205 million (http://www.3ieimpact.org).

Who is involved

The three main funders of 3ie are the Bill & Melinda Gates Foundation, UKaid through the Department for International Development and the William and Flora Hewlett Foundation. 3ie has a long list of affiliates including members, associate members and partners which ca be accessed on its website.

How

3ie's work focuses on generating high quality evidence that contributes to effective policies for the poor. It plays a dual role as funding agency and knowledge broker and carries out several activities and offers multiple services such as the following:

- Impact Evaluation Programme: offers support and resources for researchers in international development.
- Impact Evaluation Services: offers programmes, products and services that improve the quality and transparency of impact evaluations.
- Synthesis and Reviews Programme: offers support and resources for researchers in international development.
- Policy influencing activities: helps researchers to better communicate the findings of their studies to influence policy.

Activities to date

Among the many 3ie's activities, the following are highlighted in 3ie's Annual Report (2014):

- Funded 26 impact evaluations, 3 systematic reviews and 32 proposal preparation grants
- Launched the first rolling replication window with a focus on HIV prevention
- Launched the Philippines Policy Window, commissioned by Australian Department for Foreign affairs and Trade and the National Economic and Development Authority, Government of the Philippines
- Published 21impact evaluation reports; three systematic review reports; two working papers; three replication papers; and one scoping paper
- Reached 100peer-reviewed publications with 3ie-funded research
- Produced the first 3ie video lecture series of15 videos covering introductions to impact evaluation, systematic reviews and policy engagement
- Sponsored the third 3ieinternational conference on impact evaluation and the first one in Asia with the Asian Development Bank in Manila
- Awarded 70 bursaries to build researcher capacity through training, conferences and meetings
- Contributed to methods briefs on the building blocks of impact evaluation for the new web-based UNICEF impact evaluation series

Outputs produced/Results achieved

The Annual Report (2014) mentions several examples of recent results achieved which are discussed in more detail below:

- Redesigning the safety net in Ethiopia
- Enhancing learning outcomes in India
- Putting targeting outcomes in context in Zimbabwe
- Improving tax collection in Pakistan
- Informing global policy on water supply and sanitation
- Influencing public debate

Redesigning the safety net in Ethiopia: The Ethiopian government viewed the Productive

Safety Net Programme – one of the largest social protection programmes in Sub-Saharan Africa – as a key tool in its fight against malnutrition. Yet a 3ie-funded study conducted by the International Food Policy Research Institute (IFPRI) found the programme has had no impact on nutrition. The government has asked the study team to advise on how to redesign the programme so it is effective in bringing down malnutrition.

Enhancing learning outcomes in India: The Indian Central Board of Secondary Education has introduced the Continuous and Comprehensive Evaluation (CCE) system to tackle dismal learning outcomes in much of the country. A 3ie-funded study showed that CCE had no impact on learning outcomes, but a Learning Enhancement Programme (LEP), developed by the Indian NGO Pratham, had a significant effect on students' Hindi language skills. The state government in Haryana, where the study took place, has commissioned a detailed review of CCE. Meanwhile, based on these findings, Pratham has expanded LEP to over 2,000 villages in the states of Jharkhand, West Bengal and Uttar Pradesh.

Putting targeting outcomes in context in Zimbabwe: 3ie-funded research by the University of North Carolina showed that the Government of Zimbabwe's Harmonised Social Cash Transfer programme had high inclusion and exclusion errors. However, the study team showed the programme's main donor–the UK Department for International Development (DFID)- that the Zimbabwean programme's targeting performance was similar to that in other cash transfer programmes, such as the Livelihood Empowerment against Poverty programme in Ghana and Progresa in Mexico. DFID decided to continue its support for the programme.

Improving tax collection in Pakistan: A 3ie-funded randomised controlled trial in Pakistan showed that better incentives for tax collectors resulted in higher tax collection with no damage to public perceptions of the Excise and Taxation Department. Encouraged by these results, the department has asked the researchers for a follow-up study to assess the impact of non-monetary incentives such as merit-based transfer and posting in improving performance.

Informing global policy on water supply and Sanitation: During 2014 3ie started tracking how its systematic reviews are being used to inform global policy. The first-ever systematic review, on water supply and sanitation, is listed on the World Health Organization's (WHO) website as a source of evidence, and specialist publications by DFID, the Australian Agency for International Development (AAID) the World Bank, the Organisation for Economic Co-operation and Development (OECD), InterAction and World Vision. As these examples show, 3ie continues to be successful in informing policy, with evidence being

used to: take successful programmes to scale; close those that do not work; inform the redesign of programmes or policy discussions, including the design of other programmes; and improve the culture of the use of evidence. To date 3ie has documented a total of 48 cases of such uses of evidence from 3ie-funded studies, of which 10 were in 2014. Through 2014, almost half of completed or nearly completed 3ie impact evaluations have had policy impact.

Influencing public debate: 3ie works to ensure that evidence from 3iefunded studies enters public debate. 3ie does this through presentations and press coverage. Grantees presented 3ie-funded studies at over 500 events during 2014, and 3ie staff participated in over 130 events. Together they reached over 1,200 policymakers in 2014. Press coverage extends 3ie's reach still further. 3ie has recorded over 77 media citations of 3ie during 2014, including in The New York Times, The Economist, The Guardian, and The Hindu.

Challenges faced

A major challenge arises from the development community's focus on policy recommendations. As discussed in 3ie's Annual Report (2014) and 3ie's Replication Paper 1 (2014), the concerted push for the statement of policy recommendations, particularly from research in international development, can create perverse incentives for researchers in the analysis and reporting of their research. Research sponsors such as 3ie, have explicit objectives to translate research into policy. 3ie publicly states its preference for greater policy influence and policy relevance in its selection criteria for impact evaluation awards. Journals also emphasise the importance of policy recommendations, particularly those journals designed to publish applied research. A review of the submission criteria for the websites of the top 15 journals in international development reveals varied emphasis on providing policy recommendations for submitting authors. More than half of development journals mention the promotion of policy relevance such as the Journal of Development Effectiveness and the Development Policy Review.

The emphasis on policy recommendations is laudable in the quest to improve evidence-based policymaking. Ex ante, policy relevance considerations should lead to better designed studies, which is why research sponsors emphasise policy relevance in their funding competitions. Ex post, however, particularly in the absence of ex ante publication of comprehensive analysis plans, the push for policy recommendations may lead researchers to draw policy conclusions consistent with, but not proven by, their study's findings. Even when researchers are careful not to overstate their policy, others can be quick to make policy recommendations based on the tested (or implied) theory of change without asking whether alternative theories, or different causal mechanisms, were also tested. Replication can provide the opportunity to further explore the causal chain using the article's own data and perhaps adding data and information from other sources. A replication study can be used to conduct sensitivity analysis on the policy recommendations in much the same way as it can be used to conduct sensitivity analysis on the primary estimates.

G.9 Government led initiatives

Department: Planning, Monitoring and Evaluation, The Presidency, Republic of South Africa

Description and Mandate

The Department for Planning, Monitoring and Evaluation (DPME) is tasked with the continuous improvement in service delivery through performance monitoring and evaluation. The DPME works with partners to improve government performance in achieving desired outcomes and to improve service delivery through changing the way government works. This is done through coherent priority setting, robust monitoring and evaluation related to the achievement of outcomes, institutional performance monitoring, monitoring of frontline service delivery and supporting change and transformation through innovative and appropriate solutions and interventions.

As the custodian of M&E in government, DPME coordinates the Government-Wide M&E System. The Policy Framework on the GWMES is supported by three other frameworks, namely: The National Evaluation Policy Framework (NEPF) under DPME, the Framework for Managing Programme Performance Information (FMPPI) under the National Treasury and South Africa's Statistical Quality Assessment Framework (SASQAF) under Stats SA (http://www.thepresidency-dpme.gov.za/). DPME is also the custodian of work to strengthen frontline service delivery monitoring, management performance information, Outcomes reporting, citizen-based monitoring and other programmes designed to collect evidence on performance at outcome level (e.g. Operation Phakisa, which has a specific focus on delivery around priority issues such as the oceans economy).

Policies and strategies

The National Evaluation Policy Framework (NEPF) (November 2011) provides the basis for a minimum system of evaluation across government. Its main purpose is to promote quality evaluations which can be used for learning to improve the effectiveness and impact of government, by reflecting on what is working and what is not working and revising interventions accordingly. It seeks to ensure that credible and objective evidence from evaluation is used in planning, budgeting, organisational improvement, policy review, as well as ongoing programme and project management, to improve performance. It provides a common language for evaluation in the public service. The key elements of the framework are:

- Large or strategic programmes, or those of significant public interest or of concern must be evaluated at least every 5 years. The focus will be on government's priority areas including the 5 key areas of health, crime, jobs, rural development and education.
- Rolling three year and annual national and provincial evaluation plans must be developed and approved by Cabinet and Provincial Executive Councils. These plans will identify the minimum evaluations to be carried out.
- The results of all evaluations in the evaluation plan must be in the public domain, on departmental and DPME websites.
- Improvement plans to address the recommendations from the evaluations must be produced by departments and their implementation must then be monitored.
- Departments will be responsible for carrying out evaluations. DPME will provide technical support and quality control for evaluations.
- Appropriate training courses will be provided by PALAMA, universities and the private sector to build evaluation capacity in the country. The University of Cape Town and DPME jointly put on a course on evidence-informed policymaking for Directors General and DDGs across South African Departments, which is very well subscribed.
- DPME will produce a series of guidelines and practice notes on the detailed implementation of the policy framework, to elaborate various aspects of the system, and to set quality standards for evaluations.

Activities to date

DPME has established the National M&E Forum and the Forum of Heads of M&E from the Offices of the Premier. These stakeholder forums as well as the M&E learning network of government officials enhance the sharing of knowledge and good practices on M&E.

Outputs/results

The DPME Annual Report 2011/12 states three areas of outcomes, namely M&E, data systems and public sector oversight. In terms of M&E, the following outcomes have been achieved with the goal to advance the development and implementation of the outcomes approach, monitoring and reporting on progress and evaluating impact:

Institutionalised quarterly monitoring of the delivery agreements by Cabinet focusing on key areas of progress and challenges requiring unblocking.

Initiated reviews of the delivery agreements

Carried out a Mid-Term Review that provided an assessment of progress towards meeting government priorities

Produced the National Evaluation Policy Framework

Assisted the political principals in the Presidency with technical support for their hands-on monitoring visits.

In terms of data systems, the following outcomes have been achieved with the goal to promote monitoring and evaluation practice through a coordinated policy platform, quality capacity building and credible data systems:

- Managing national and provincial monitoring and evaluation forums
- Managing learning networks and developing training courses for officials
- Managing data forums linked to improving data
- Developing guidelines on various aspects of M&E
- Providing the Programme of Action platform for the outcomes
- Production of the 2011 Development Indicators
- In terms of Public Sector Oversight (PSO), the following outcomes have been achieved with the goal to conduct institutional performance monitoring and front line service delivery monitoring:
- Developed a Management Performance Assessment Tool
- Completed assessments of 27 national departments and 60 provincial departments by the end of March 2012
- Developed high-level proposals for linking results of assessments of departments to individual assessments of Heads of Department
- Instituted monitoring of a range of indicators of the performance of national and provincial departments
- Assessed draft Annual Performance Plans of 33 national departments
- Developed and implemented a Frontline Service Delivery Monitoring Programme and conducted more than 100 unannounced monitoring visits

Phillips et al (2014) argue that there are increasing examples in South Africa of where M&E information is used to inform policy- and decision-making. These include quarterly reports o progress in implementing priority outcomes, briefings to the President on the performance of ministers and briefings to Cabinet and Parliament on management performance. In case of departments and individuals performing poorly, corrective actions is taken to address identified problems. With the evidence base is increasing, a growing number of changes have been made to evaluated programmes.

Challenges faced

Actual challenges can be derived from the guidelines for improving the operation of M&E. (DPME Guidelines No 3.1.4: Improving the Operation of M&E in Offices of the Premier, March 2013)

In general, more demand for M&E needs to be created by changing and promoting the M&E culture, based on continuous learning and improvement, a common understanding of what M&E entails and what it intends to achieve.

Data quality needs to be improved that emanates from systems within individual provincial departments and within municipalities. These systems need to be made robust and credible to ensure that information can be gathered and aggregated in a correct and timely manner. More emphasis needs to be put on the verification of data.

Duplication of data needs to be avoided. Provincial departments sometimes report the same information multiple times to various offices.

Moving away from pure data gathering to analysis and communication is a crucial step to advance M&E practices. Acquired data must be analysed in order to give rise to M&E insight and implications for improving performance.

Necessary M&E capacities need to be built and technical support given. More attention needs to be given to capacitating officials in local and provincial government on the technical and managerial dimensions of M&E and indicator development. Better training should be tailored to various M&E practitioners (local versus national government officials, specialist users versus managers).

In addition, Phillips et al (2014) identify a number of hindering factors to M&E. The following factors rank highest:

- Problems are not seen as an opportunity for learning
- Management does not fully buy into M&E
- M&E is seen as the M&E team’s job only
- Weak M&E culture
- M&E is primarily regarded as a tool for controlling staff rather than an improvement tool
- M&E team’s influence limited

An interesting observation by Phillips et al (2014) regards incentives. Compliance is used as the primary incentive which is an important driver for change including the precision of indicator definition, target setting and improving data collection and quality. However, it also limits the potential impact due to a perverse incentive for departments to set targets low so that they are easier to achieve. This suggests that compliance should be complemented with other incentives to achieve better results.

G.10 Initiatives, policies, strategies or statement in relation to use of evidence/ monitoring and evaluations in Zambia and Uganda

Zambia

Initiatives

The CLEAR Midterm Report states that Zambia has taken steps to underline the importance of (and build its own government's capacities) in generating information about and reporting on its government's performance, or engaging in capacity building activities to enhance the use of evidence in decision making by its government cabinets.

The CLEAR's AA centre's strategy (2013) indicates that the centre's future work will focus on four countries (South Africa, Zambia, Ethiopia and Ghana), and that in certain other countries (namely Kenya, Rwanda, Nigeria, Botswana, Uganda and Tanzania) it will limit its engagement to outreach and awareness raising activities and selected demand-driven interventions with individual client organisations.

The AA centre has been effective in generating resources for M&E capacity building initiatives in Anglophone Africa. Nevertheless, the prospects for the centre's viability are modest. However, some potential government partners in the region who were interviewed (e.g., the Government of Zambia) indicated that they would be willing and prepared to share costs of CLEAR services in the future.

Further to that, the Ministries of Finance and Health have some M&E policies in place, as discussed below. In addition, the Zambia Monitoring and Evaluation Association (ZaMEA), a local affiliate of the African Evaluation Association (AFREA), aims at publicising and marketing the profession through training in M&E and related skills, providing a platform for exchange of ideas, setting and ensuring adherence to high level of professional standards of practice and delivery of good and services as well partnering with training institutions to raise the standard of theory and practice of M&E (http://www.afrea.org).

Policies/Strategies

Monitoring and evaluation is stated on the Ministry of Finance website (http://www.mofnp.gov.zm), however, no further information is given. Looking at the Millennium Development Goals Report of the Ministry of Finance provides some information. The report states that if Zambia improved its collection of timely data along with the consistency of data collection and compilation methodologies, it would enable policy design and implementation as well as the accurate monitoring of millennium development goals progress. For many indicators, there is a lack of consensus amongst state actors on the methodologies used. For instance, there are three different data sets on the production of copper, provided by the Central Statistical Office, the Bank of Zambia, and the Ministry of Mines, Energy and Water Development. In addition, old data does not facilitate timely and relevant policy-making and implementation, and can provide an outdated and at times irrelevant story. It therefore advocates for timely, publicly available and robust data, disaggregated to the extent possible. This data must be made available in the public domain for more people to access and engage on it. In doing so, the public would be allowed to contribute their voice to the policy choices made to accelerate progress on the millennium development goals for Zambia.

The Ministry of Health's National Health Strategic Plan 2011-2015 states that monitoring and evaluation of the implementation of the plan will be conducted through appropriate existing and new systems, procedures and mechanisms. The Monitoring and Evaluation Sub-Committee will be responsible for providing advice on all matters concerning monitoring and evaluation.

The following describe the main tools and approaches that will be applied in the monitoring and evaluation of the implementation of the plan.

The Ministry of Health and the sector partners will harmonise sector performance indicators, and use these as the basis for monitoring and joint reviews. Indicators will include: sector performance benchmarks and

triggers for sector budget support, output and process indicators to assess service delivery (quality, access, efficiency) and indicators of health status (impact).

The Ministry of Health will be responsible for coordinating health sector monitoring and reviews and the common routine systems will be the major tools for data collection. This data and its analyses should then be used by various agencies for decision making. It will also plan and lead the Joint Annual Reviews (JAR) every year, together with appropriate involvement and support of other Government ministries and key stakeholders.

There will be two evaluations during the duration of each National Health Strategic Plan developed under this plan, a mid-term review, after the first 2.5 years of implementation, and a final review at the end of the duration. Stakeholders will jointly agree on the timing, terms of reference and composition of these two review missions.

Progress to date

Reviewed documents are mainly forward looking as the approaches to M&E are still in their early stages.

Barriers to use of evidence and evaluation

The Study on the Demand for and Supply of Evaluation in Zambia (December 2013) discusses several challenges. For instance, it notes that the supply of evaluation expertise needs further development before evidence can be used and evaluation done. Consultancy firms and individuals have arisen with specific areas of strengths in response to demand form funding partners but it is observed that qualified staff is leaving for better paid positions elsewhere. The study also notes that there is very little actual demand from stakeholders outside funding / development partners. It is also regarded unclear how the gathered information from monitoring requested by the presidency actually feeds back into accountability and ultimately performance. In this respect it is mentioned that evaluations touching on sensitive areas such as resource allocation or infrastructure need to be taken with “consideration” suggesting that there might be potential conflicts of interests at play which represents a major barrier. As noted earlier, the Ministry of Finance demonstrates an actual demand for evaluation and is in the process of setting up an evaluation function, however, it depends on outside financial and technical assistance which slows the process significantly.

Uganda

Initiatives

As in the case with Zambia, the CLEAR Midterm Report states that Uganda has taken steps to underline the importance of (and build its own government’s capacities) in generating information about and reporting on its government’s performance, or engaging in capacity building activities to enhance the use of evidence in decision making by its government cabinets.

The CLEAR’s AA centre’s strategy (2013) indicates that the centre’s future work will focus on four countries (South Africa, Zambia, Ethiopia and Ghana), and that in certain other countries (namely Kenya, Rwanda, Nigeria, Botswana, Uganda and Tanzania) it will limit its engagement to outreach and awareness raising activities and selected demand-driven interventions with individual client organisations.

In addition, Uganda’s Office of the Prime Minister and the Ministry of Health have some M&E policies in place that are worth mentioning.

Policies/Strategies

Uganda's Office of the Prime Minister dedicates a section on its website to Monitoring and Evaluation, however, information is limited to the progress made in 2013 (see below) and planned activities (http://opm.go.ug/departments/PolicyCoordinationMonitoringandEvaluation/).

Progress to date

Uganda's Office of the Prime Minister (http://opm.go.ug) summarises the key achievements as follows:

- Launched and operationalized the National the M&E Policy.
- Produced the Government Annual Performance Report FY2012/13 and the Government Half-Annual Performance Report 2013/14
- Coordinated & conducted Sub-county Accountability Meetings in 28 districts
- Evaluation Standards and guidelines for Public Sector were developed in collaboration with Uganda Evaluation Association
- Conducted 2 Evaluation studies: Summative evaluation of the Effectiveness of the Avian & Human Influenza Project (AHIP) and Public Procurement & Disposal of Assets (PPDA)'s development impact and its role in ensuring efficiency and effectiveness in public procurement

Barriers to use of evidence and evaluation

An important challenge concerns how donors make use of available M&E systems. While donors often channel supported projects through the M&E systems of partner governments, they are concerned that these systems might not work as well as intended. Donors are faced with the risk that they may not be able to gain sufficient information to assess whether their projects are achieving their goals, if these M&E systems perform poorly (http://devpolicy.org).