METHODS LAB

April 2016

IMPACT EVALUATION FOR PORTFOLIO PROGRAMMES ON POLICY INFLUENCE

REFLECTIONS ON THE INDONESIAN POVERTY REDUCTION SUPPORT FACILITY

Jessica Mackenzie and Simon Hearn

KEY MESSAGES

- Donors are increasingly using portfolio-based programmes that embrace 'good failure' and adaptive, political programming.
- However, measuring the impact of these programmes is challenging, especially for those working on policy influence and building country systems; not only do you need to measure the positive and negative impact of the overall portfolio, but also the different pathways tested.
- Programmes, therefore, need a light-touch monitoring and evaluation system that allows it to remain flexible.
- Good practice examples of portfolio-based programmes present six strategies to evaluate impact: 1. Develop appropriate logic models 2. Collect observational data throughout implementation 3. Develop stories of change or case studies 4. Understand causal relationships without a counterfactual 5. Purposefully select which activities to study 6. Be explicit about how impact will be valued across the portfolio.
- These strategies are only useful if monitoring and evaluation is placed at the centre of programme decision-making.

A METHODS LAB PUBLICATION

ODI.ORG/METHODSLAB

The Methods Lab is an action-learning collaboration between the Overseas Development Institute (ODI), BetterEvaluation (BE) and the Australian Department of Foreign Affairs and Trade (DFAT). The Methods Lab seeks to develop, test, and institutionalise flexible approaches to impact evaluations. It focuses on interventions which are harder to evaluate because of their diversity and complexity or where traditional impact evaluation approaches may not be feasible or appropriate, with the broader aim of identifying lessons with wider application potential.

How to cite this working paper:
Mackenzie, J., and Hearn, S. (2016) 'Impact evaluation for portfolio programmes on policy influence'. A Methods Lab publication. London: Overseas Development Institute.

Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
Tel +44 (0) 20 7922 0300
Fax +44 (0) 20 7922 0399
info@odi.org.uk
www.odi.org

BetterEvaluation
E-mail: bettereval@gmail.com
www.betterevaluation.org

About this paper

This paper was commissioned by the Methods Lab, a collaboration between ODI, the Department of Foreign Affairs and Trade (DFAT) and BetterEvaluation. The Methods Lab seeks to develop, test, and institutionalise flexible approaches to impact evaluations. It focuses on interventions that are harder to evaluate because of their diversity and complexity, or where traditional impact evaluation approaches may not be feasible, with the broader aim of identifying lessons with wider application potential.

The purpose of this paper is to increase the knowledge and understanding of how the impact of certain types of portfolio-based programmes – especially those working on policy influence and building country systems – can be evaluated. This includes increasing knowledge on how most appropriately to define impact and judge success, as well as which methods and approaches are useful in these types of programmes. To do this, we use an evaluation case study from Indonesia: the Poverty Reduction Support Facility (PRSF).

The audiences for this paper include those designing, managing and implementing portfolio programmes, in particular development practitioners working at DFAT. This paper is also intended for those responsible for developing or implementing monitoring and evaluation (M&E) for portfolio programmes.

The methodology for this study on which this paper is based included over 50 interviews (see Annex 2) with practitioners, evaluation experts and Indonesian government staff over a three-month period. It also included a review of more than 40 documents ranging from the programme design, reporting and key outputs, to relevant comparative evaluation systems and emerging impact evaluation literature.

It should be noted that the case paper does not intend to evaluate any of the Indonesian government programmes discussed in this paper. It is not an evaluation; instead, it draws on the evaluation tools, methods and approaches used by PRSF. The National Team for the Acceleration of Poverty Reduction (TNP2K) is discussed in the paper only as far as many of its activities are symbiotic with PRSF – which exists to support TNP2K. It is also important to note that DFAT added a number of activities to PRSF's mandate (known as 'DFAT Windows') but, for simplicity, this report focus only on PRSF's work supporting TNP2K.

Acknowledgements

This working paper is a Methods Lab publication written by Jessica Mackenzie and Simon Hearn (Overseas Development Institute). Peer reviewers include: Patricia Rogers (BetterEvaluation), Stewart Norup (Mampu) and Louise Shaxson (Overseas Development Institute).

The authors would like to thank Bernie Wyler, Scott Guggenheim, Thomas Pratomo, Vincent Ashcroft, Fiona McIver, Francesca Bastagli, Ajoy Datta, Tiina Pasanen, Anne Buffardi, Irene Guijt, Angus Kathage, James O'Brien, Rick Davies, Vera Scholz, Jess Dart and Patrick Sweeting for their input and support in developing the case study. Special thanks to Louise Ball, Hannah Caddick and Steven Dickie for their editorial and design work.

Acronyms

BPS	Indonesian National Bureau of Statistics
BSM	Help for Poor Students Programme
CA	contribution analysis
CDKN	Climate & Development Knowledge Network
CIFOR	Centre for International Forestry Research
COR	collaborative outcomes reporting
DDD	doing development differently
DFAT	Department of Foreign Affairs and Trade
GCS	global comparative study
GEF	Global Environment Facility
GEM	General Elimination Methodology
ICR	Independent Completion Report
IDEAs	International Development Evaluation Association
IPR	Independent Progress Review
JPAL	Jameel Poverty Action Lab
M&E	monitoring and evaluation
MAMPU	Empowering Indonesian Women for Poverty Reduction programme
ODE	Office of Development Effectiveness
ODI	Overseas Development Institute
OECD	Organisation for Economic Co-operation and Development
OH	outcome harvesting
OPM	Oxford Policy Management
PDIA	problem-driven iterative adaptation
PRSF	Poverty Reduction Support Facility
PT	process tracing
QCA	qualitative comparative analysis
RASKIN	Rice for Poor Families Programme
REDD+	Reducing Emission from Deforestation and Forest Degradation
Susenas	Indonesian National Social Economic Survey
TNP2K	The National Team for the Acceleration of Poverty Reduction
TWP	thinking and working politically
UDB	Unified Data Base

Contents

Executive summary

As donors grapple with different mechanisms to help build country systems (Organisation for Economic Co-operation and Development, 2010), they are increasingly turning to portfolio-based programmes that work on trial and error, embracing 'good failure' and adaptive, political programming. But these programmes' outcomes and impacts involve less traditionally defined or measurable concepts, and instead relate more to complicated notions of whether a country system has (in whole or in part) been 'built'. So an important question for the development community, and the focus of this paper, is: how can donors evaluate the support mechanisms that underpin these less traditional programmes in the area of country systems building?

This paper takes as a case study the Australian-funded Poverty Reduction Support Facility (PRSF) and the Indonesian government's National Team for the Acceleration of Poverty Reduction (TNP2K), which the PRSF was set up to support. TNP2K was established in 2010 to help build country systems in Indonesia to improve the rate of poverty reduction, including government coordination and delivery of poverty programmes. It has been regarded as a clear, demonstrable success, and an example of how to conduct such types of complicated programmes sensibly (Ashcroft, 2015). But, while TNP2K was widely acknowledged to have *worked* and to have provided a demonstrably high return on investment (Ashcroft, 2015), *how* this success was achieved is far less clear. This is because the impact evaluation systems were not always in place to capture it.[1]

With portfolio-based programmes assuming greater importance, if we are to replicate the success of initiatives such as PRSF and TNP2K, we need better approaches to capturing data on why and how they succeed or fail.

This paper provides an overview of the PRSF programme, including its ambitious objectives and challenging scope, and why it represents a useful case study for this investigation (Chapter 1). It then addresses the inherent challenge of evaluating support programmes like PRSF and why more traditional approaches (like randomised control trials and other forms of counterfactual analysis) are unable to tell the full story. It summarises why the systems set up by PRSF to capture impact in TNP2K did not always work (Chapter 2), and then recommends strategies for measuring the impact of future policy influence portfolio-based programmes of this kind (Chapter 3). A full in-depth analysis of what PRSF and TNP2K intended to and actually did measure in terms of impact is also provided as a subsidiary case study in the Annex A of this paper.

Given that policy influence programmes need to be opportunistic and nimble, a rigorous measurement system to capture what works and why will never be perfectly comprehensive as it would make the programme too unwieldly or slow. Instead, successive programmes will need to have a light-touch system so they can remain flexible and agile.

In addition to advocating for the use of light-touch systems, this paper recommends the consideration of six strategies (and guidance on how to apply them) to enhance planning, M&E of impact, discussed in Chapter 3. They are to:

1. develop appropriate logic models
2. collect observational data throughout implementation
3. develop stories of change or case studies
4. understand causal relationships without a counterfactual
5. purposefully select which activities to study
6. be explicit about how impacts will be valued across the portfolio.

This paper also considers several programmes of similar scope, funding levels and approaches to policy influence as the PSRF. These could provide helpful real-life examples of how these strategies are applied, and the combinations that might work well for future programming.

The approaches recommended in this paper are only worth applying if M&E is placed next to the centre of senior decision-making on the programme. This will allow politically adaptive programme management to occur in real time. It requires strong communication of findings and results, with accessible sense-making (or synthesis) tools, and also relies on strong programme relationships, with access given to and trust in the M&E team. In short, it is not only the use of different tools and strategies that will help portfolio-based programmes to better assess their impact, but also positioning M&E to be a more integrated and useful part of the programme.

This paper attempts to address some of the key elements of this problem of how to build country systems through portfolio-based programmes and identifies what it hopes are useful strategies to support more effective programmes of this nature.

1 A vast amount of high quality reporting was produced by TNP2K in order to make recommendations to Indonesian government, but this was not focused on measuring their own performance, which was seen as PRSF's role. For an overview of PRSF's systems, see the case study in Annex A of this paper.

1. Introduction

1.1 Why this study is important and interesting

The National Team for the Acceleration of Poverty Reduction (TNP2K) has been acknowledged as 'an unqualified success' (Ashcroft, 2015: 3). It was established by the Indonesian government in 2010, in direct response to the government's commitment to accelerate poverty reduction. Its purpose was incredibly difficult to achieve: to help build country systems in Indonesia to improve the rate of poverty reduction, including government coordination and delivery of poverty programmes.[2] That is, its goal was to build the systems that produced these results, rather than to produce the results themselves. This is not the first programme to address country systems building but, where most programmes have been ineffective (OECD, 2010: 44), TNP2K is regarded as a clear, demonstrable success, and an example of how to conduct these types of complicated programmes sensibly (Ashcroft, 2015: 11).

The difference with TNP2K's approach to others addressing similar issues (as stated by key stakeholders during interview) was that: (i) it was a portfolio-based programme, taking a flexible approach to activity design throughout the life of the programme; (ii) it delivered results almost exclusively through evidence-based policy-making approaches; and (iii) it was designed and implemented in line with progressive principles including problem-driven iterative adaptation (PDIA), 'doing development differently' (DDD) and 'thinking and working politically' (TWP) – even if this was not explicit. Compounding this was the scale of its operations, reaching target populations of tens of millions of people spread over 16,000 islands, in just four years – and in a country recognised for having many weak or corrupt public institutions.[3]

As with many programmes working on country systems building, TNP2K's results frameworks are basic and its theory of change remains implicit in parts. Furthermore, as with many portfolio-based programmes, it conducted a range of pilot activities – some of which produced successes and some of which were failures, with large unintended spill overs that were sometimes uncaptured. Despite these evaluation shortcomings, TNP2K was widely acknowledged to have *worked*, and provided a demonstrably high return on investment (Ashcroft, 2015: 13). Yet *how* that success was generated remains locked in a 'black box'.

With donors grappling with the different mechanisms to build country systems (OECD, 2010: 55; Gillies et al, 2012), the question for development practitioners, and focus of this paper, is: how can donors evaluate the support mechanisms (such as the DFAT-funded PRSF) that underpin country systems building programmes like TNP2K, when the outcomes and impacts are not the more traditionally defined or measurable concepts, but rather whether a country system has – in whole or in part – been 'built'?

It is this question – the unpacking the components of the 'black box' and how to evaluate them – that this paper aims to elucidate. By doing so, the authors hope to improve the replicability, scalability and innovative efforts of future portfolio-based programmes working on building country systems.

1.2 The programme: what is the Poverty Reduction Support Facility?[4]

In 2009 the Indonesian government committed to accelerating poverty reduction, aiming to lower the (stagnating) poverty rate from 14.1% in 2009 to 8-10% in 2014.[5] The government recognised an urgent need to increase efficiency and reduce waste across national social protection programmes (Homes et al, 2011: v-vii). It cited the proliferation of overlapping and sometimes mis-targeted programmes (for example, as many as 90 on community-driven development alone) that each had different planning, oversight and accountability systems (PRSF, 2010: 5). There was an urgent need for high-level coordination and strategy.

TNP2K was established by the Indonesian government in 2010, in direct response to this need. The TNP2K Secretariat[6] has a mandate to accelerate poverty reduction and strengthen social protection systems by: (i) improving the performance of poverty reduction programmes; (ii) improving programme targeting through common methods and better household listing for all social

2 The programme goal specified in both the design and the Monitoring and Evaluation Plan was: 'Increased rate of poverty reduction and reduced impact of shocks and stresses on the poor and vulnerable', through 'improving poverty reduction and social assistance policies based on evidence; improving the delivery of social assistance services and programmes for the poor and that government coordinates better to develop and implement integrated poverty programmes.' (M&E Plan: 10).

3 Transparency International 2014 Corruption Perceptions Index, Indonesia ranks 107 out of 175 (www.transparency.org/cpi2014/results).

4 For more information on PRSF and TNP2K see the Independent Completion Report conducted in 2015, which provides a comprehensive overview.

5 Indonesia's Medium Term Development Plan 2010-2014.

6 Both TNP2K and the TNP2K Secretariat will be treated as the same entity for the purposes of simplicity in this paper.

Table 1: PRSF-TNP2K budget over the life of the programme

	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015
BUDGET (AU$)	4 million	8 million	25 million	35 million	30 million

protection programmes; (iii) undertaking monitoring and impact evaluations of the social assistance programmes; (iv) identifying important but troubled social assistance programmes and resolving their implementation issues.

In response to a request in 2009 from Indonesia's Vice-President, the Australian government established the Poverty Reduction Support Facility (PRSF) to support TNP2K (PRSF, 2010). It was created to provide the technical, managerial and financial support services TNP2K needed to fulfil its mandate. This included the provision of basic equipment, staff and premises. Beyond this, PRSF was directed to generate knowledge to inform social protection policies, define policy options, translate policy choices into operational programmes and provide high quality monitoring and evaluation. It would do this by: producing research; designing and managing pilot reform projects; supporting reform initiatives undertaken within relevant ministries and agencies; developing and managing the Unified Data Base (UDB); and other DFAT directed activities.

PRSF began with a budget of AU$15 million over four years, and this increased significantly over time to an operating budget of approximately AU$30 million for 2014 alone. Its total expenditure from 2010 to September 2014 was AU$76.8 million – five times its original budget.

TNP2K and PRSF worked in close coordination, with TNP2K taking the policy and technical lead. PRSF in contrast had limited strategic control and yet was responsible for contracting and administering staff and resources for TNP2K, while remaining accountable to DFAT. For more information about the programme refer to Annex A, which presents the full case study.

1.3 Characterising the programme

The aim of the Methods Lab is to learn about impact evaluation of programmes with complex and complicated aspects. TNP2K can certainly be described in this way. There are a number of characteristics of TNP2K that make it an interesting case to study: (i) the mechanism through which TNP2K was funded was a facility takes a portfolio approach to programme design; (ii) it focused on supporting and influencing policy change by providing an evidence base for policy-makers; and (iii) it worked innovatively through political and adaptive approaches, and at times being considered an extension of Indonesian government.

1.3.1 Implemented through a portfolio approach

In recent years, portfolio-based programmes (also known as facilities)[7] have become an increasingly popular model of aid delivery in the Australian aid programme.[8] Portfolio-based programmes have a defined budget to support the development and implementation of projects and activities to achieve a high-level objective. In the Australian aid programme they exist as a partnership between the Australian government and a partner government in the country where the programme will focus. They have substantial flexibility in choosing which projects to fund – a decision that is usually guided by demand from the partner government (Dawson, 2009).

In portfolio programmes, activities are developed during implementation. A portfolio-based programme has a clear goal but a loose theory of change (perhaps with several hypotheses) and the expectation is that activities will be developed throughout the life of the programme. Essentially, this allows the programme to test different pathways towards that one goal, so it can course-correct and determine how best to spend any marginal funding. In this way, the programme increases its understanding of the context and what works as it goes – rather like a laboratory, the programme is testing different pathways or conditions to achieve the end goal.[9] As typically government-to-government mechanisms, many portfolio-based programmes work towards their objectives though supporting policy change and reform processes through the whole policy cycle: agenda setting, policy development, decision-making, implementation and monitoring and evaluation.

Portfolio-based programmes have several advantages over other aid models, such as projects or traditional programmes, which together suggest they are likely to continue as a popular mechanism for DFAT (DFAT, 2015).[10] They are flexible – allowing the programme to be responsive; they provide simplicity of administration; they provide something of a laboratory effect, allowing the

7 The terms can be used interchangeably, but this report will use the term 'portfolio-based programming' to avoid duplication and confusion.

8 Examples from the early generation include the Philippines-Australia Governance Facility (1999-2004) and the East Timor Capacity Building Facility (2003-2006), whereas more recent additions include the Australia Indonesia Facility for Disaster Reduction, the Papua New Guinea Governance Facility, the Australia Indonesia Partnership for Economic Governance (AIPEG) Facility and the Timor-Leste Justice Sector Support Facility.

9 For more information on portfolio-based programmes, see Annex A.

10 DFAT Strategic Framework 2015-2019 (http://dfat.gov.au/about-us/department/Pages/strategic-framework-2015-2019.aspx).

programme to find out what works as they try multiple approaches at the same time; they are suited to pursuing innovation; they allow programmes to determine the most efficient and cost-effective pathway, and revisit this during implementation (useful in DFAT's restricted budget environment); and they are an effective modality in complex environments where 'course corrections' are often required (Dawson, 2009).

1.3.2 Supporting and influencing policy change

TNP2K works almost exclusively by providing an evidence base for policy change in Indonesian poverty programmes. That policy change is an inherently political and uncertain process compounds the difficulty of TNP2K's work; there is no guarantee that research, even if of highest quality, will lead to the recommended policy change (Young and Court, 2004). If programmes want to understand how research influences policy then we need to know where to look to trace the influence. The policy objective may be a budget change, a new programme, initiation of reform or new legislation. All of these objectives will generally take a long time to come to fruition, with many steps in the process and many people influencing the decisions. Programmes can learn from other models to a certain degree but what works will depend on many local factors – e.g. the type of policy issues addressed, the politics around the issues, the form of influencing, the amount of evidence that exists already, the kinds of findings and recommendations being made and the people involved and their skills. There is no set recipe. All of these factors compound what TNP2K was trying to achieve and complicate efforts to understand how change happens.

1.3.3 Working politically and adaptively to build country systems

TNP2K approached its mandate by working innovatively through a political and adaptive approach. It embodied principles of 'doing development differently' (DDD), 'problem-driven iterative adaptation' (PDIA), as well as 'thinking and working politically' (TWP) – even if this was not explicit in its programming. These approaches are often heralded as good practice in international development programming, particularly for programmes with complex aspects or in complex settings. In reality, they can be hard to apply operationally as government bureaucracies, donors or managing contractors, struggle to provide the flexibility to apply them. TNP2K embodied PDIA in its approach to resolving problems across Indonesia's social protection programmes – identifying duplication or shortcomings, testing a variety of solutions, providing metrics to demonstrate the benefits of any adjustment, and advocating for changes to be applied. Furthermore, as noted by the Executive Secretary of TNP2K, the Australian government's trust to allow Indonesian government representatives to predominantly manage the programme (and high level of risk borne by the managing contractor) was a very important factor in the programme's success and a good example of DDD. Because this, largely donor-funded, programme was widely seen as a part of Indonesian government, it was able to navigate highly political sensitivities as it advocated for and achieved change to multi-million dollar programmes through Cabinet and a polarised multi-party parliament, relatively smoothly. These elements were seen as crucial to its success but, though implicit in the staff/leadership's activities and approach, were not documented in the theory of change.

2. The challenge of evaluating impact in PRSF and similar support programmes

This chapter considers the challenge of evaluating impact in PRSF and similar support programmes. Firstly, it states the elements of impact evaluation as defined in this paper. Secondly, it explains how and why portfolio-based programmes are challenging to evaluate, and how, in the case of PRSF, this is even more difficult because of its complicating factors beyond typical portfolio-based programmes. Thirdly, it considers that, based on the implicit assumptions within the programme and the gaps identified, there are certain components that future evaluation systems could be set up to address (this is the focus of Chapter 3).

PRSF is referred to throughout this chapter. The assessments made are based on a detailed, in-depth analysis of PRSF's intended approach to evaluating impact, and what actually occurred in practice. The full case study is presented in Annex A.

2.1 What is impact evaluation

Within the development community there is much debate about what impact means (Hearn and Buffardi, 2016). This paper uses the OECD-DAC definition (OECD 2002), which explains impact as long-term, positive or negative, direct or indirect, primary or secondary changes. Evaluation helps to judge the merits of a particular intervention – for example, to demonstrate the value of a donor or government investment, or to help a programme improve its effectiveness. All evaluation should consider, from the outset, what its purpose and audience is, and what kinds of evaluation questions it should address. Impact evaluation is distinct from other types of evaluation because it makes a judgement about the *causal relationship* between the observed impact and the programme, i.e. to what extent the programme caused or contributed to the changes (White, 2009; Rogers, 2012).

Impact evaluations generally attempt to answer three kinds of questions: descriptive, causal and evaluative (Rogers, 2014). We summarise these into three basic questions that we will use throughout this paper (adapted from Befani, 2016):

1. *What* has changed and for whom?[11]
2. *How* and *how much* have the programme activities contributed to those changes?[12]
3. *So what* is the overall merit and worth?

Answering the '*what*' questions involves collecting quantitative and qualitative data to describe how things are now, how they've changed, what has happened, what the programme has done, what other related programmes have done and what the context is in which this is happening (Rogers, 2014).

Answering the *how* and *how much* questions in impact evaluation can also be called 'causal inference'. The more complicated a programme is, however, the less likely it is that a measure of *how much* impact was produced can be obtained. For example, it is not possible to answer the question *how much did the research contribute to the policy change*, which infers a quantitative answer. Instead, it is more meaningful to ask how and to what extent did the research contribute, which infers a qualitative answer.

Typically, programmes will use one of the following four views of causation to determine causal inference for impact (as outlined by Stern et al, 2012): (i) counterfactual analysis;[13] (ii) regularity approach; (iii) configurational analysis; and (iv) a generative approach – see Box 1.[14]

Answering the '*so what*' questions involves a different set of methods, which help to weigh up the descriptive and causal findings and apply evaluative criteria to come to an overall conclusion about the success of the

11 Focusing, as defined above, on long-term, positive or negative, direct or indirect, primary or secondary changes.

12 Befani lists a range of tools that measure these different dimensions, and combinations of tools that can be complementary. The '*what*' is drawn from the relevance criteria, the '*how much*' is from the effectiveness criteria, and the '*how*' comes from transferability. Although transferability is not an official evaluation criterion, it is hinted at in the various why questions, according to Befani (2016).

13 Counterfactual analysis can be used for both the how much and for the how to some limited extent on simpler programmes. It is less helpful for measuring how on a complicated programme like PRSF.

14 These techniques are explained in some technical detail, in Befani, B. (2012). Models of causality and causal inference. Department for International Development (DFID) (http://betterevaluation.org/resources/guide/causality_and_causal_inference). For reasons of brevity this report will not go into detail about these four dimensions of causal inference, but will provide an overview of each and apply them to PRSF and TNP2K, as well as discuss the tools that flow from them (http://mande.co.uk/blog/wp-content/uploads/2012/07/2012-Causal-Inference-BB-February-26.pdf).

Box 1: Four traditional views of causation

Counterfactual analysis is where a comparison is made between two nearly identical cases differing only in cause and effect. By comparing one situation (where the programme intervention occurred) with a counterfactual one (where the programme intervention had not occurred) one should be able to imagine the difference between the consequences, in order to estimate the 'impact' of the intervention.[15] Tools that use counterfactual analysis include randomised control trials (including constructed qualitative counterfactuals), quasi-experimental and natural experiments.[16] These tools (and which combinations can be most appropriately applied) are explored in Chapter 3 in greater detail.

A regularity approach can be used to determine impact when a potential cause and a presumed effect are observed often enough, after one another, and it is assumed one causes the other.[17] Tools that use regularity include observational studies, statistical modelling and econometrics.

Configurational analysis is based on the idea that many causes come in 'packages' – where several causes are needed in combination to produce an effect.[18] Different combinations can lead to the same outcome, and similar combinations may lead to different outcomes. Configurational analysis helps to unpick these factors and what combinations they occur in, similar to a recipe's set of ingredients. Tools that use configurational analysis include qualitative comparative analysis (QCA) and RAPID outcome assessment.

A generative approach is different again. If the configurational view sheds some light on the recipe, then generative informs on how the ingredients must be put together, following what order and techniques. It details how things follow one another in a sequence. Tools that use a generative approach include realist evaluations, process tracing, case studies, general elimination theory and contribution analysis.

Source: Stern et al. 2012

programme (Rogers, 2014). This is inherently a political process that draws on a set of values to negotiate between alternative views of success.

2.2 The difficulty of evaluating impact in portfolio-based programmes

Normal portfolio-based programmes are hard to evaluate, more so when they are as ambitious as TNP2K. This is because they trial initiatives and ideas, learning from those that do not work and building on those that do, to increase the chances of success. In addition to measuring the positive and negative impact of the overall programme on beneficiaries (to what extent the programme met its goal), they need to measure the effectiveness of the different pathways to reaching that goal.

Some helpful analysis and guidance has been produced on how to evaluate portfolio-based programmes (such as Dawson 2009, unpublished), but not a great deal, and none that is published. There is significant experience among those who have been managing portfolio-based programmes, though it is not always documented or publicly available. As a result, there is a lack of understanding about how to monitor and evaluate the longer term effects of portfolio-based programmes and how to use that information in decision-making. We highlight challenges relating to the three impact evaluation questions introduced above: *what*, *how* and *so what*.

2.2.1 The problem of defining the *what*

Defining impact in a portfolio-based programme is not straightforward. For example, the Australian government expects that its investment in PRSF will have a beneficial effect on the lives of poor people living in Indonesia, namely that their poverty is reduced. We might then expect that the impact of PRSF is accelerating poverty reduction in Indonesia. However, PRSF is designed to support TNP2K, which also has the mandate (from the Indonesian government) to accelerate poverty reduction.

Furthermore, TNP2K supports and enhances a number of other programmes across Indonesia that are themselves intended to reduce poverty. This is visualised in the basic five-step logic model in Figure 1 below. The nature of portfolio-based programming is that there is inherent tension between pursuing its own strategic direction and being demand led – responding to various demands from government and demonstrating value, particularly in early stages of implementation. This can undermine evaluability efforts as coherence between activities is gradually lost.

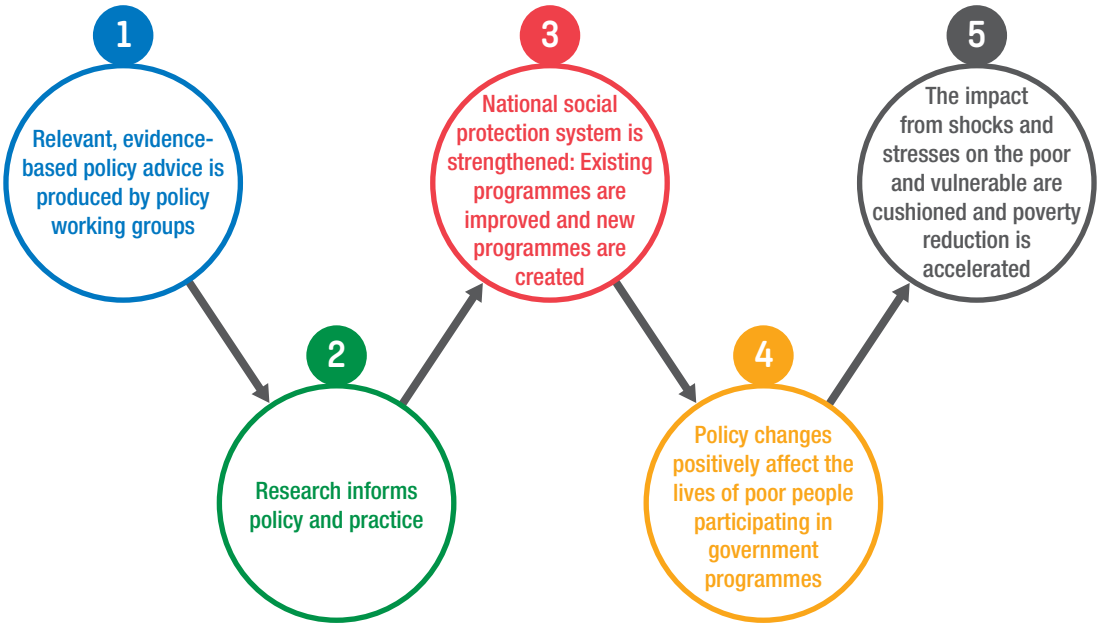
15 For more on counterfactuals – see the BetterEvaluation website: http://betterevaluation.org/search/site/counterfactual.

16 This also draws upon the analysis and findings of White and Phillips (2012), particularly with regard to small and large N studies. This will be explored further in Chapter 3 (www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf).

17 Befani (2012): 3 (http://mande.co.uk/blog/wp-content/uploads/2012/07/2012-Causal-Inference-BB-February-26.pdf).

18 Befani (2012): 14 (http://mande.co.uk/blog/wp-content/uploads/2012/07/2012-Causal-Inference-BB-February-26.pdf).

Figure 1: The PRSF five-step logic model



The question, then, is how do we define the impact of PRSF? Where do we look for results that will tell us whether, and to what extent, it has been successful? Dawson (2009) recommends that the underlying reason for selecting a facility approach should form the basis for its judgment. In the case of PRSF, this is to build the national social protection system. The impact of PRSF, then, should be defined as the third circle in Figure 1, and our question becomes: *what has changed in the national social protection system?*

This is a crucial shift in thinking about impact, from changes in the lives of poor people to changes in national systems. This is not to say that boxes 4 and 5 in Figure 1 are not important; clearly, a national system which is working well will be able to demonstrate that its programmes and services are having an effect on the people being served, but this is the responsibility of the Indonesian government in this case. If PRSF does the best job possible then the government will be able to demonstrate its impact: that poverty reduction is accelerating. Changes in poverty reduction alone are insufficient to assess the impact of PRSF; the important piece is to understand how the system has changed in a way that makes it possible to affect poverty reduction.

The case study shows that PRSF was designed with this definition of impact. Success was to be defined as both 'outcomes… through the programmes they support and by the number of policy proposals acted upon by the Government as a whole,' and 'increasing appetite for evidence based policy making' (PRSF, 2010: 12). The kinds of outcomes it describes are very much about strengthening the government systems.

In implementation, however, the focus of M&E was more mixed. For example, there were a lot of data collected about the uptake of research but not on what effect that was having on government systems, which is what an impact evaluation would have wanted to address. When analysing the value for money of PRSF, the approach used focused not on outcomes relating to the country system but on measurable changes to benefits delivered to poor people. This confusion as to what, exactly, the programme was to be measured against diluted its ability to be clear about its overall success or failure.

2.2.2 The problem of getting to the how

Many programmes focus on trying to calculate *what* and *how much* impact was delivered, in order to demonstrate value of investment to donors. However, very few reach the stage of understanding how an intervention made a difference.[19]

Answering the *how* question involves an investigation of the intervention's interplay and causal relationship with other factors. In a portfolio-based programme, where innovation is key, understanding the *how* is crucial. Essentially, to replicate a programme, or to try and deliver a better programme, you need to know why it worked at least as much as *how much* impact was delivered. Because different pathways to achieving impact are tested concurrently in a portfolio programme, understanding how a pathway is delivering impact allows the programme to adjust and to know where to spend marginal money for increased impact.

19 Which might use tools like case studies, Qualitative Comparative Analysis, or RAPID outcome mapping.

The case study shows that there is a great deal of understanding about how to measure what changed in the lives of poor people in Indonesia. It also shows that PRSF has developed effective ways to measure *how much* contribution it had to particular policy reforms. It shows that some studies have been able to understand how particular policy changes have contributed to change in the lives of poor people but this only gets at half the question.

To understand *how* PRSF contributes to changes in the lives of poor people, we also have to examine the nature of the policy changes themselves and how PRSF and TNP2K activities contribute to those changes. PRSF demonstrated a good understanding (in the 2012 evaluability assessment) of the importance of the relationship between research activities and policy change, but were unable to make progress in monitoring and evaluating these relationships. In essence, PRSF did not capture information on how research informs policy and helps to build the country system.

2.2.3 The problem of weighing results across a portfolio – the 'so what' question

Very rarely will any kind of programme be completely successful, and the more complex the programme the less likely this will be the case. There will almost certainly be a mix of things that worked well and things that did not – or the strategies worked, but only in certain settings, with certain people or at certain times, and not universally.

This is certainly the case for portfolio-based programmes, and it is often explicitly built in to the design that a certain proportion of the portfolio will often be for 'high risk' activities that have a high chance of failing. The main difference with portfolio-based programmes is that, whereas in a conventional programme all the activities are, for the most part, planned ahead of time and linked through a strategic framework, in a portfolio the activities emerge over time. Although they share common high-level objectives, activities can often have very different immediate outcomes.

This makes it more difficult to weigh up the results of the individual activities to make an overall judgement; it is not simply a case of adding all the net-effects of activities together. A high-risk activity may have failed but in the context of the portfolio that may have been a good thing because it leads to learning that enhances other activities. But it can be difficult to distinguish between failures that were 'worth it' and failures that were due to bad planning or implementation.

PRSF's quality assurance system provided, to some extent, the data and decision-making spaces to make sense of failures but it was not implemented from the start, so is insufficient for the purpose of evaluating impact. There was little attempt to synthesise the data from individual activities, beyond the value for money assessment, which focused on net-benefits of four of the major programmes that TNP2K supported.

2.3 The added complexity of building country systems

Evaluating impact in the case of TNP2K is even more difficult because it has additional complicating factors that go beyond typical portfolio-based programmes. As discussed in Chapter 1, TNP2K's ambition was to help to build country systems in Indonesia, an emerging middle-income country with institutional weaknesses, corruption and a transitioning democracy.

As discussed by Gillies and Alvarado (2012), country system building has been an objective of development interventions for decades but the approaches and underlying theories have changed over time, from focusing on individuals to institutions to whole systems. They define a 'strengthened' system as: 'one that is robust, coherent, integrated, self-organizing, self-driven and resilient' and suggest that strengthening a system implies 'improving its characteristics and increasing its ability to address challenges and solve problems' (Gillies et al, 2012). PRSF is an example of the latest wave of these kinds of initiatives that takes a systemic approach. This means it doesn't just deal with individual parts of the system in isolation but recognises the interrelationships and boundaries between these parts, and the different perspectives they have (Williams and Hummelbrunner, 2010).

Taking a systemic approach requires a change from 'conventional' development interventions. One of the most recent manifestations of the kind of difference needed is captured in the DDD manifesto (see Box 2).

In retrospect it is possible to see that PRSF demonstrates many of these principles in the way it worked, yet this way of working wasn't explicit in its planning, monitoring or evaluation. This is understandable given that practice of this kind is at the cutting edge, as Gillies and Alvarado (2012) show, our conventional planning, M&E systems are not appropriate for system building work. Although there is guidance available to better understand how to incorporate a systems and complexity approach to planning M&E (Williams and Hummelbrunner 2010; Britt 2013) in practice, there are few examples that address impact at the level of country systems.

2.4 How does PRSF measure up?

This chapter has explained the elements of the challenge of evaluating impact in PRSF and similar support programmes. From how impact evaluation is defined and why portfolio-based programmes are challenging to evaluate, to how the systems that PRSF set up did not capture the means by which impact was produced or sufficient valuation of the impact/s.

Box 2: Doing Development Differently manifesto

Initiatives that are able to address complex challenges and foster impact reflect these common principles:

- They focus on solving local problems that are debated, defined and refined by local people in an ongoing process.
- They are legitimised at all levels (political, managerial and social), building ownership and momentum throughout the process to be 'locally owned' in reality (not just on paper).
- They work through local conveners who mobilise all those with a stake in progress (in both formal and informal coalitions and teams) to tackle common problems and introduce relevant change.
- They blend design and implementation through rapid cycles of planning, action, reflection and revision (drawing on local knowledge, feedback and energy) to foster learning from both success and failure.
- They manage risks by making 'small bets': pursuing activities with promise and dropping others.
- They foster real results – real solutions to real problems that have real impact: they build trust, empower people and promote sustainability.

Source: http://doingdevelopmentdifferently.com

There are several conclusions in the case study (Annex 1) that are important to note about the extent to which PRSF and TNP2K's approach to evaluating impact addressed these challenges. These are derived from the case study's assessment of what was intended to be measured and evaluated in terms of impact (from the design and other formative documents) and what was actually measured and evaluated in practice.

The first relies on an acknowledgement that more reporting was done on this programme than almost any other DFAT programme, as perceived by key informants at interview. The significant funding given to TNP2K to produce research relevant to policy was a huge investment by the Australian government that we are perhaps unlikely to see again (Ashcroft, 2015), so it is unhelpful for a report to simply recommend 'more investment'. Instead this paper suggests an alternative balance in future programming (with a focus that includes evaluating how things worked) and appropriate allocation of funding to reflect this. The research that was produced by TNP2K was considered high quality (produced in conjunction with world leaders like the Abdul Latif Jameel Poverty Action Lab (JPAL) and Oxford Policy Management (OPM)). However, it had limited focus on the programme's own performance and, rather, focused on how to improve Indonesian government programming. In future, there should be very clear delineation between what was funding for research (such as appraisals and randomised control trials of social protection programmes in Indonesia) and what funding was to be spent on evaluating the programme's own impact (actually very little in PRSF and TNP2K's case).

The second conclusion was that PRSF at some point deviated from the design's intention to focus on uptake and influence. The reporting and synthesising (or sense-making) for this important aspect of the programming was lost. This was pointed out in one of the reviews during implementation but was never really addressed and brought back on track.

Third, TNP2K and PRSF had a helpful focus on *how much* impact had been generated by the programme, particularly towards the end of its mandate in 2015 (for example this was when PRSF produced the informative value for money exercise). However, what would have been helpful is if the programme had been able to follow through on their own impact into Indonesian programmes more, and apply metrics to measure that impact better. For example, they could follow through on TNP2K's recommendations to programmes like the Rice for Poor Families Programme (RASKIN) and the Help for Poor Students Programme (BSM),[20] and seek information on what their recommendations achieved in practice once applied. In addition to this, their focus on how much should not distract from the need to set up a system to capture how activities worked, and the focus on where to look for their impact (within the complex objective of country systems building).

Fourth, the important role played by informal M&E systems should not be forgotten. Having a DFAT technical specialist based in-house in TNP2K for two days per week led to real changes that vastly improved the programme's capacity to evaluate impact (not least the quality assurance system that was established in 2012). These remained largely undocumented and it was only through interview that such mechanisms were illuminated.

Fifth, PRSF's focus was largely on monitoring rather than an evaluative role. This was in large part due to structural incentives for reporting on the programme and relationships. PRSF found it difficult to gauge the uptake of recommendations made by TNP2K to Indonesian

20 Detailed in the case study at Annex 1.

government, and TNP2K had few incentives to report to PRSF. The reporting culture needs to be geared towards evaluative roles, as well as monitoring, and the support mechanism empowered to do so.

Chapter 3 examines these gaps and makes suggestions for how portfolio programmes can monitor and evaluate intermediate effects on policy and practice, in order to understand how they contribute to measured or observed impacts. By using a combination of tools from different causal inference views (regularity, configurational and generative approaches, for example) programmes such as PRSF can ensure they are measuring not just the *what* and *how much*, but also the *how* in order to determine the best pathways. This goes to the heart of the programme aims and is one of the main purposes of adopting a portfolio modality (in policy influence programmes): that is, to experiment and learn about what works.

3. Strategies for evaluating impact of portfolio programmes

Chapter 2 has highlighted the challenges of evaluating impact in portfolio-based programmes and Annex 1 details the approach to impact evaluation taken by PRSF, outlining what was intended and what actually happened. Chapter 2 concluded that there are gaps in the approach taken relating to the three core impact evaluation questions: *what*, *how* and *so what*. This chapter will explain what tools and strategies there are to fill these gaps, and how a future programme might set up systems to manage them from the outset. Six strategies are recommended. These are to:

1. develop appropriate logic models to define the *what* and hypothesise about the *how*
2. collect observational data throughout implementation to understand *what* is changing
3. develop stories of change or case studies to develop narratives about *how* things are changing
4. understand causal relationships without a counterfactual to draw reliable conclusions about the impact of the portfolio
5. purposefully select which activities to study to support answering all three questions
6. be explicit about how impacts will be valued across the portfolio to answer the *so what*.

These strategies have been selected for their abilities to help evaluate the impact of portfolio-based programmes as described in this paper. That is not to say that they are all *impact evaluation* strategies per se (many of them relate to monitoring and planning as much as impact evaluation). Across the Methods Lab case studies, one of the key findings has been that evaluating impact requires more than smart impact evaluation designs; it requires programmes to be oriented towards impact assessment (Peersman et al, 2016). This finding is consistent with other work around evaluation of adaptive programming – for example, the USAID Discussion Note on Complexity-Aware Monitoring (Britt, 2013), which describes a number of approaches for monitoring complex aspects of development assistance.

3.1 Develop appropriate logic models

Policy and system change is a political and uncertain process. If we want to understand how portfolio activities influence policy then we need to know where to look to trace the influence; we need hypotheses that can be tested. A good theory of change or logic model will make explicit the uncertainties, assumptions and risks that need examining, and will also help clarify the expected impacts (both intended and unintended) and how the intervention is thought to contribute to these. The problem with developing a theory of change is that in a portfolio the specific activities are not known ahead of time; there can't be a specific plan for how the activities will influence policy, and usually in a portfolio the activities emerge and need to be developed quickly to be responsive to a current window of opportunity and so there is little time to think about developing complicated models.

There are several tips that can be useful in this context, without having to develop detailed programme theories.[21]

3.1.1 Tips for keeping it simple

Start with generic testable hypotheses

The aim here is to think about the kinds of activities that may make up the portfolio and the characteristics that might be common across them, particularly the characteristics which may affect their ability to influence change. For system or policy change work, the kinds of activities are not unlimited: technical advice, policy briefings, training, networking, research, evaluation, advocacy, public affairs. It should be possible to develop a list of all the measurable or observable characteristics of these kinds of activities and the kinds of effects they might have. Box 3 presents an example of this demonstrating characteristics of research and policy outcomes which might be expected.

From these characteristics, hypotheses can be constructed – for example, about what type of research uptake approach could succeed, where and when. This can then be tested through the collection and analysis of data – see sections 3.3 and 3.5 for more guidance on this. This list of conditions can be used across the portfolio for any activity implemented, as a guide to inform data collection.

21 Another paper in this series, by Rick Davies, focuses on developing 'loose' theories of change for flexible development interventions (Davies, 2016).

Box 3: Examples of characteristics and outcomes

Characteristics of research that may affect its ability to influence policy[22]

- **Characteristics of the policy issue itself.** Is there strong demand for change or evidence? How relevant is the policy issue to intended users or decision-makers? What is its profile in the policy agenda or media? What is the breadth of agreement between stakeholder groups? What coverage does the issue have in terms of population of potential beneficiaries?
- **Characteristics of the social protection programme under scrutiny.** Which ministry it is implemented by? Which cluster or sector does it relate to? What geographical coverage does it have, or number of existing beneficiaries? What is the public interest in the programme?
- **Characteristics of the engagement strategy.** Is there a formal or informal strategy? What is the timing of start of engagement, duration, frequency, directionality (i.e. user-led? Researcher-led?), form (i.e. in person, telephone, letter), typology of the relationship (shared decision-making? Advice)? Is there any budget allocated? Are there person-days specifically assigned to dissemination and communication efforts? Is there a use of different channels for different groups targeted?
- **Dissemination of findings.** What is the length of product or advice? Are there any audio-visuals used – videos, blogs, websites, infographics, charts, figures? Has there been translation into different languages? Has there been follow up investigation arranged to determine impact within ministry?
- **Characteristics of the activity team.** Is there a notable presence of champions? Does the team have sufficient understanding of the relevant context? Is there an ability to anticipate policy influence opportunities? What is the extent of the embeddedness of team's contacts network into the context (previous research experience in the country/geographical area, reputation/credibility of the team)?

Possible policy outcomes as a result of the activities:

- Change in funding (significant or minor?)
- Change in the number of beneficiaries reached (significant or minor?)
- Change policy or programme design (type of change, significant or minor)
- Change policy or programme implementation (type of change, significant or minor)
- Change in culture of evidence-informed policy-making (among intended users, donors, civil society, academia – in particular changing framings/discourse/beliefs/attitudes/debates about evaluation and evidence-informed policy-making).

Take an actor-centred approach

Another approach to developing a theory of change that doesn't rely on detailed knowledge about programme activities is to focus on the actors involved and the interactions between them. Outcome mapping is a methodology based on this approach (Jones and Hearn, 2009). It starts by asking which actors in the system can have a demonstrable effect on its transformation, and then asks what would have to change in the behaviour, activities or relationships of those actors for them to be able to support transformation of the system. Stakeholder analysis of this kind is commonplace in the 'problem-analysis' stage of programme design but it rarely translates into programme theory, strategy and monitoring. In early stages of portfolio management, where outcomes are unknown or uncertain, it can be hugely helpful to have a 'map' of actors to understand who your partners are and who they are engaging with. By focusing solely on these actors, the realm of possible outcomes is drastically reduced and it is manageable to collect data.

Balance up-front planning and adaptive management

Uncertainty is a major barrier to detailed planning. As Snowden and Boone (2007) recommend, in complex settings, it isn't possible to rely on 'best practices' or expertise to plan how to act and instead the manager has to act first and see what happens and then respond. This has big implication for managing portfolios, especially those, like PRSF, which are focused on system building. Instead of implementing programme controls and tight procedures to manage risk up front, the approach has to be more adaptive (Hummelbrunner and Jones, 2013). This is not to say that all aspects of portfolio management will be like this; parts of

22 These characteristics have been adapted from other QCA efforts in development programming that the authors are aware of.

the portfolio may well be more predictable and low-risk where up front planning and application of 'best practices' is possible, but when the focus is on building country systems this is likely to be a small component. Managers should consider the level of complexity across their portfolio and develop appropriate approaches accordingly (Rogers, 2008).

Think about multiple theories of change at multiple levels

For a portfolio, a single theory of change may not make sense. As previously discussed, portfolios are often chosen as a funding mechanism for the explicit reason to test different, and sometimes competing, hypotheses. Each activity under the portfolio could have its own theory of change, which may be tested. In some cases this may be where attention needs to be focused instead of working out a logic for the whole portfolio. For example, PRSF funded several randomised control trials of pilot programmes (e.g. introducing identity cards to the RASKIN rice distribution programme). There will be a theory behind the pilot programme (e.g. how will introducing identity cards improve rice distribution?) A well-executed randomised control trial, on a well-executed pilot, will be able to test whether the theory holds up or not by measuring the impact of the pilot. Getting from pilot results to scaled-up impact, however, is not a straight forward leap and needs careful thinking. A specific theory of change on this part of the portfolio activity is also needed.

In addition to theories about individual activities, there could also be theories about how the portfolio as a whole is developed and managed will also affect its outcomes. For example, the kind of donor assistance modality, the level of partnership between donor and recipient, the level of expertise and background of staff hired, the management approach (as previously discussed) and learning across the portfolio will all have an effect on the overall effectiveness of the portfolio.[23] Many of these decisions will be tacit, based on experience, but it can help to make them explicit and have a plan to test them (Buffardi and Hearn, 2015).

For example, there are several factors that were acknowledged in TNP2K (and in many programmes) as being crucial to its success.[24] If future programmes are interested in more comprehensive replicability, then the nuancing of how programmes were managed should also be captured throughout the life of the programme. These categories include: strong political will (which translates into real demand for the programme's activities); how much risk appetite there is for managing failure; the way that donors support these programmes (type of technical assistance, level of partnership and ability to hire qualified staff); and the leadership's ability to establish metrics to test and demonstrate the progress of the programme.

3.1.2 Application for the case study: why, when and who to do it

The case study found that PRSF had a fixed theory of change, which remained loosely developed and largely unused in the strategic operations of the programme. One of the main recommendations of this paper is to use testable hypotheses instead of a single logic model. This means the programme is: (i) choosing a discreet, manageable number of pathways to test; (ii) really engaging with the assumptions the programme is built on; and (iii) bringing coherence of purpose to the team's work. Testable hypotheses can be created in the first three months of the programme (developed by the whole programme team, but led by the M&E team or senior management, possibly with input from the Indonesian government). These hypotheses can then be revisited and checked, using information from activities across the programme, every second quarter (by the M&E team). These checks would include ensuring that multiple levels across the programme are represented, including hypothesis testing at the activity level and the portfolio level. Half-way through the programme it would be helpful for the whole group (led by the M&E team) to revisit the hypotheses and discuss whether any changes should be made (discard the tested or add new hypotheses).

The case study found that PRSF took an approach that saw surveys or studies produced by researchers, which management then (largely implicitly, in an undocumented way) adapted into appropriate products based on their understanding of the needs of the research users. An actor-centred approach (thinking of the users' needs from the outset, including the research approach) would not only help design, package and support use of research but recording the rationale behind these decisions would make the paper trail more explicit for the M&E team. This is helpful when monitoring research uptake and policy influence, and enables the broader team to be more conscious of working politically, or gearing operations towards research uptake. The actor-centred approach would be applied at the beginning of new activities, as part of the quality assurance process when activities are selected (by programme officers, or researchers, with oversight from the quality assurance process lead or management).

23 Information provided from interviews conducted with TNP2K staff, DFAT staff and Scott Guggenheim.

24 Key stakeholder interviews, as well as Ashcroft (2015).

3.2 Collect observational data throughout implementation

The business of engaging with change processes, be it social, political, environmental, is not often logical, pre-planned or predictable. It is more often spontaneous, emergent and somewhat haphazard. For this reason it can be difficult to look back and be certain of what happened, let alone what might have been influential in bringing about a particular change. Keeping an ongoing record of what happened and what was observed at the time can be a vital way of piecing together how things are working and whether things are going to plan.

One of the most basic ways of capturing information is by keeping a record of observations, trends quotes, reflections and other information that can be recorded as and when they arise, and then forgotten about until analysis at a later date. Our recommendation is to use a log or journal in conjunction with the conditions and hypotheses developed above.

If, for each activity in the portfolio, there is a log of which activity characteristics are present and what types of outcome have been observed then this will enable comparative analysis later on. This can be integrated with the activity selection process and the activity quality assurance or reporting processes. This, combined with a qualitative journal recording contextual information, observations and 'reflections in the moment' (for example, by writing a brief memorandum after an event or significant meeting) can provide rich data for understanding and explaining the particular ways in which research is influencing policy.

3.2.1 Application for the case study: why, when and who to do it

The case study found that PRSF made some efforts to collect data on research uptake (for example, the number of policy briefs produced) for its quarterly reporting but then stopped. They were not reporting significantly on behavioural changes or policy influence. The main recommendation is for the programme to collect observational data throughout implementation. This will help the programme managers to test hypotheses by understanding what is changing, create a paper trail for monitoring and evaluation, and to provide evidence of policy influence. This can be done throughout implementation of activities by someone within each activity, plus someone at programme-level providing a common framework and aggregating the information.

3.3 Develop stories of change or case studies

Stories of change and case studies help to document a single change process and provide analysis or reflection on the causes and effects of the change and *how* the activities contribute to that particular change. The process of developing a story is valuable as a monitoring exercise as it forces programme teams to collect and analyse data and reflect of how things are changing. The story is also useful for later analysis, particularly if a common framework or template is used (such as testing or developing QCA hypotheses). Three variations are described below: stories of change, episode studies and outcome harvesting (OH).

3.3.1 Three approaches to a programme story

3.3.1.1 Stories of change

Stories of change is an inductive case-study method to investigate and report on the contribution of an intervention to specific outcomes. The story does not report the activities and outputs of the intervention but rather the mechanisms and pathways by which it was able to influence a particular change that has been observed. The change being described in the story can be an expected change that the intervention was targeting or it can be an unexpected change that was observed but was a surprise – which itself can be positive or negative with respect to the original objective. Stories could also describe how an intervention failed to influence an expected change, in which case they analyse the possible reasons why.

There are three major steps to writing a story of change.

1. **Choose the story.** The emergence of a success (or failure) usually prompts a story – this may become evident through any of the data collection methods described above (e.g. through a journal or impact log) – so there is already a sense, or hunch, that the intervention has made a significant enough contribution to make an interesting story.
2. **Gather the evidence.** To understand the contribution of the intervention and provide a plausible argument you will most likely have to search for additional information. This will involve interviewing key stakeholders and programme staff to trace the influence of your work and identify the mechanisms that led to the change. This should involve an element of substantiation of claims that the intervention has had an influence through, for example, consulting experts in the field or those close the change at hand.
3. **Write the story.** Stories should be relatively short – two to four pages – and written as a narrative that is easy to read and leaves an impression. It should make a clear case for the intervention, describing the situation or challenge it was responding to and how it intended to engage; it should focus on who was doing what when and what effect did that have; and it should discuss the success or failure factors and any lessons to take forward to future interventions.

3.3.1.2 Episode studies

Another case study method similar to stories of change is episode studies, which rather than starting from the perspective of an intervention, starts from the perspective of the change and tracks back. The steps would be the same as those for stories of change except that the evidence gathering stage investigates any and all factors that influenced the change, including but not limited to the intervention. This can be quite an involving task and generally requires access to those close to the decision-making around the change in question. The advantage of this approach is that it can give rise to the relative contribution of the intervention to the change in relation to other influencing factors and actors – not in a quantitative sense but through the perspectives of those close to the decisions. An episode study is an account of the different mechanisms that led to a particular change. It is not a systematic assessment of the level of contribution of each factor that influenced the change, but is still very labour and evidence intensive.

3.3.1.3 Outcome harvesting

As discussed, it is not always possible to develop clear indicators or specific intended outcomes for policy influencing as the effects of such work are uncertain and emergent. This means we don't always have a clear plan of what to look for and measure for regular monitoring. Outcome harvesting (Wilson-Grau and Britt, 2012) was developed for situations such as this and doesn't rely on a pre-existing logic model, theory of change or results framework; it is an example of what is referred to as 'objective free' evaluation.

The premise is quite simple: start with what you do know – you know who you have been working with and what you've been doing – and conduct a 'harvest' of outcomes to record what has changed. The output of the harvesting process is a set of short narratives about changes that have occurred in the organisations, institutions, people or groups a programme has been working with. The narratives have a particular form, which keep the process systematic and replicable. There are three paragraphs: description, significance and contribution. The description of what has changed: who has changed; when and where did this change take place; what is the nature of the change, which could be behaviour, actions, activities, relationships or policy change – but has to be something observable. 'Significance' discusses the relevance of the change, why it is important, what it will mean for other people, how it relates to the programme's goals. 'Contribution' describes how the programme is thought to have contributed to this change, using the best evidence available. A key part of the process is to verify, with external informants, whether the claims are reasonable; we come to this part of the process again in the next section.

Because of the brief nature of the outcome narratives it can be possible to collect quite a number over the course of a programme and develop a database. For example, the World Bank Institute complied a set of case studies of several institutional strengthening programmes that used the outcome harvesting approach. Each case study was typically based on 20-30 outcome narratives. Evaluations of larger programmes can document hundreds of outcomes. In this way, a collection of outcomes can help to re-construct the change process retrospectively by focusing on the relationships between intermediate changes.

3.3.2 Application for the case study: why, when and who to do it

Stories are a natural way of engaging that reaches diverse audiences, putting the 'flesh on the bones' of monitoring and evaluation. The three examples shared have increasing levels of involvement, with stories of change the simplest to produce and outcome harvesting the most complicated and probably requiring external advice. The case study found that PRSF did not produce stories of change or case studies of policy change, possibly due to the time and costs involved. Using stories could have helped the programme communicate more meaningful outcomes for the successes of the programme and, through harvesting, could have shown interesting aggregate lessons for the sector, raising the profile of the programme. In the first six months of the programme, the monitoring and team could have developed a timeframe and decided on the number of stories or studies to be produced, including rough criteria for selection. This would need to be presented to the whole team for approval. Several stories or studies would be produced, throughout the year as decided, by researchers or programme officers and checked by the M&E team with possible assistance from an advisor or evaluation consultant. A process of synthesising these to draw broader conclusions (for example via outcome harvesting) and triangulation with stakeholders and the Indonesian government, would be conducted by the M&E team (with possible help from an advisor or evaluation consultant) on towards the end of the programme.

Box 4: Developmental evaluation

One approach to evaluation that embodies many of the strategies suggested in this section is developmental evaluation (DE). DE describes a different way of doing evaluation: as a long-term partnership between an evaluator and a programme, specifically complex and uncertain programmes. It aims to generate understanding about the programme, its environment and its effects, and to support innovation and further development of the programme. This is in contrast to standard evaluation approaches, which often start with the assumption that the programme has already been developed and implemented and is ready to be assessed. DE is not a method as such but an approach to conducting evaluation that does not rely on or advocate any particular evaluation method, design, tool or inquiry framework (Quinn Patton et al, 2015). Instead, DE offers eight essential principles (Quinn Patton et al, 2015):

1. Developmental purpose: the primary role of the evaluator is to support the development of the innovation
2. Evaluation rigour: gather, interpret and report data using appropriate methods and standards
3. Utilisation focus: decision and actions should be made with respect to intended uses by intended users from beginning to end of the process
4. Innovation niche: processes of innovation and adaptation should be at the core of the enquiry
5. Complexity perspective: development and change should be interpreted through the lens of complexity theory
6. Systems thinking: evaluators should think wider than the initiative in question and consider about the interrelationships, perspectives and boundaries within the wider system
7. Co-creation: the evaluation and innovation should be developed together such that the evaluation becomes part of the change process
8. Timely feedback: findings and insights should be shared when they are needed, not at pre-planned intervals.

Most of these principles are well-suited to the evaluation of portfolio-based programmes that are working at country-system level. Such funding modalities are chosen because of the need for innovation in a context of uncertainty. However, putting these principles into practice could be a challenge for large, high-profile donor programmes – particularly the co-creation principle, which requires a closer relationship between programme management and evaluation than is normal in these programmes and would require a relaxing of independence requirements. While elements of DE may work in specific parts of a portfolio (those that are high-risk, high gain), it may not be suitable as a replacement for standard evaluation approaches completely.

3.4 Understand causal relationships without a counterfactual

Chapter 2 discussed how, using the examples from PRSF and TNP2K, you can measure to a high degree of accuracy the contribution of specific policy changes (e.g. introducing identity cards to RASKIN) on specific beneficiary outcomes (e.g. household income). These kinds of studies rely on counterfactual analysis, which, although there are significant methodological and practical considerations, can work well in controlled and discrete environments. In this section we are concerned with understanding the causal relationships between portfolio activities and policy or system-level changes where counterfactual analysis is not feasible. Fortunately there are alternative strategies which don't rely on counterfactuals. This section suggests four strategies that go beyond the case study and story approaches in the previous section (see also BetterEvaluation for more detail).

3.4.1 Going beyond case study and story approaches

3.4.1.1 Compare activity characteristics across the portfolio

Portfolios present an ideal opportunity for conducting comparative analysis. The activities within a portfolio will have natural variation in that they are implemented under different conditions with different strategies by different people in different places, but are all working towards the same goal and seeking similar kinds of outcomes. There are techniques available which, given the right quality of data, can determine which activity characteristics or contextual factors are important for producing certain outcomes and, more importantly, which combinations of characteristics or factors are effective. This is key for policy influencing activities because there will never be a single strategy that will always work and there will most likely be multiple ways of achieving the same outcomes. Methods such as QCA (Befani, 2012) and Decision Tree modelling (Davies, 2012) are beginning to be used to evaluate programmes, particularly programmes where there is not a single theory of change. Both methods rely on a similar approach but use different algorithms to find the optimum solution and present the results in different ways.

At the core of both methods is the construction of a simple table with cases (in this instance, portfolio activities) listed row by row, against their characteristics, contextual factors and outcomes of interest, which are listed column by column. For this purpose, the characteristics and outcome types listed in Box 3 above present a helpful set of conditions. There is not a minimum number of cases required for this kind of analysis but it is most useful if there enough cases to potentially exhibit all possible combinations present (BetterEvaluation).[25] A recent survey of QCA users found that the median number of cases was 22 and the median number of characteristics was 6. The data for the table would ideally come from a log such as that suggested in Chapter 3.2 (collecting observational data), and is usually presented as 0 or 1 indicating the presence or absence of a characteristic or outcome. Software algorithms are then used to compare the different combinations and develop a model for the conditions that are most likely to produce each of the desired outcomes.

There are few publicly available publications on the application of QCA in development programming but there is much ongoing work.[26] The development community will likely be in a very different situation in late 2016. There are two notable exceptions that are relevant to TNP2K's work and future programmes of a similar nature. Firstly, the Climate and Development Knowledge Network (CDKN) is trialling QCA in their 'negotiations support' component of the programme to test long-standing assumptions about what combination of factors leads to uptake of CDKN commissioned research by policy-makers and development practitioners. A forthcoming paper on how this approach has worked to monitor and evaluate policy influence in a large scale programme will be relevant for future programming in this area. Secondly, the Global Environment Facility (GEF) published a report in October 2015 that assesses the impact of their investments in non-marine protected areas. This report applies QCA and analyses the 'extent to which the management and governance approaches supported by GEF have led to the achievement of GEF objectives.' (GEF, 2015: vii). QCA is also increasingly becoming a topic of discussion at evaluation forums like the International Development Evaluation Association (IDEAs) conference in Bangkok in October 2015.

QCA is technically demanding and requires the use of specialist software and so future programmers may also choose to analyse their data using simpler tools such as EvalC3.[27] EvalC3 is an Excel application that enables users to identify one or more sets of project attributes, which are good predictors of the achievement of an outcome of interest. It is expected to be made freely available in 2016 under a Creative Commons licence. The important point to note is that understanding causal relationships without a counterfactual is possible and is a rapidly developing field with interesting ideas emerging for future programme application.

25 http://betterevaluation.org/evaluation-options/qualitative_comparative_analysis.

26 One of the few publications is DFID's review of evaluation approaches and methods for interventions related to violence against women and girls http://r4d.dfid.gov.uk/pdf/outputs/misc_gov/61259-Raab_Stuppert_Report_VAWG_Evaluations_Review_DFID_20140626.pdf.

27 See website: http://evalc3.net/ (managed by Independent Consultant Rick Davies).

3.4.1.2 Test activities against a constructed theory

Many qualitative causal analysis methods rely on comparing what actually happens in a programme with what was expected to happen. This is usually described in a programme theory, theory of change or logic model. For portfolios, it isn't usually possible to have a detailed theory of change in the planning stages, for reasons already discussed. It is possible, however, to develop a theory of change in retrospect. This is unlikely to be useful at the portfolio level but it can be useful when examining individual activities within the portfolio. Two approaches that use this strategy are contribution analysis (CA) (Mayne, 2008)[28] and process tracing (PT) (Collier, 2011;[29] Bennett, 2010).[30]

Both approaches begin by examining programme documentation and other relevant literature (e.g. sectoral analyses, prior evaluations, theoretical frameworks), as well as asking those involved in the activity, in order to develop a model or hypotheses of how the activities were thought to lead to the intended outcomes. In PT the theory of change is in the form of competing hypotheses describing cause and effect relationships, known as causal-process-observations. In CA the theory of change is in the form of a results chain which elaborates the assumptions and risks inherent in the causal relationships.

In both cases the aim is to test whether the constructed theories of change hold true in practice and whether what was predicted by the theories has in fact occurred. If the theories are sufficiently grounded and plausible then all that is needed are observations or non-observations of predicted outcomes or events. In CA this involves developing a 'performance story' (similar to the stories of change already described). A performance story documents the extent to which the activities were implemented in line with what was expected in the theory, whether the expected results occurred and what other factors were present which may have affected outcomes. In PT a chronology of events is constructed in the form of a narrative, detailing the order of events that have occurred, and then applying a set of standard tests to compare the actual events with the hypotheses.

3.4.1.3 Triangulate with key actors

One of the simplest ways to understand causal relationships is to ask people: those involved in the activities, on the receiving end of activities, or people on the outside looking in, such as experts or commentators. The story of change methods described in previous sections will rely on a range of data to understand the outcomes of portfolio activities and may even be able to propose the ways in which the activities have contributed to the outcome. But to be taken seriously these relationships should be triangulated with other sources: does anyone contest the facts being reported, are they consistent with results from other similar programmes, does anyone have any reason to doubt the causal claims being made?

The OH process, which was introduced in section 3.3.1.3, includes an explicit step to 'substantiate' the outcome statements that have been developed. This involved asking for feedback from key informants to check the validity of the description, significance and contribution statements. This is predominantly done through interviews with identified people, both internal and external to the programme.

3.4.1.4 Investigate possible alternative explanations

Even when we have good evidence that our activities have contributed to particular outcomes, we still have to consider the possibility that external factors may have a greater, and maybe overriding contribution – such as another programme operating in the same space but with a much larger budget, or a political change in the country that creates space for transformative change. It is good practice, therefore, to rule out or document alternative explanations for observed changes. Many of the methods discussed in previous sections of this paper inherently include an element of this – for example, by outlining competing hypotheses, PT involves ruling out alternative explanatory variables throughout the process. General Elimination Methodology (GEM) (Scriven, 2008: 11-24) is an approach that specifically attempts to do this as part of an evaluation process. For a given outcome, a list of possible causes or competing explanations is developed, each with a unique footprint that can be observed. The process is then to establish the 'facts of the case' so as to determine, for each cause, whether there is evidence that it exists or not. If the footprints cannot be found then it is possible to eliminate that cause from the list, leaving just those that have a plausible causal link.

28 http://betterevaluation.org/resources/guides/contribution_analysis/ilac_brief

29 www.ukcds.org.uk/sites/default/files/uploads/Understanding-Process-Tracing.pdf

30 http://philsci-archive.pitt.edu/8872/1/Bennett_Chapter_in_Brady_and_Collier_Second_Edition.pdf

3.4.2 Application for the case study: why, when and who to do it

The case study found that PRSF did not have a systematic approach to comparing characteristics or testing the theory of change throughout the programme. They approached this task implicitly, adapting as they went to the politically changing environment. Explicit causal analysis is important in order to test your theory of change and hypotheses, to reorient the programme adaptively and be able to publish findings to inform the broader discourse at the end of the programme. The recommendation of this paper is to implement a QCA approach, which relies on comparing characteristics of activities and occurrence of outcomes across the portfolio. Characteristics (and a common data collection approach across the portfolio) can be developed at the same time as the hypotheses creation, in the first three months of the programme. This can be done by the whole group, led by the M&E team, with strong input from senior management and possibly the donor. The data is analysed on a regular basis determined by the chosen analytical tool. For example, a light-touch tool could be run every few months, whereas QCA is more complex and so analysis would be run less frequently – likely towards the end of the programme.[31] The data analysis would be led by the M&E team (including triangulation and testing alternative explanations), possibly with assistance from a technical expert. Testing the theory of change and hypotheses would occur at the end of individual activities (especially those that demonstrated interesting results, were of significant size or tested complicated pathways), by the M&E team, with input and advice from researchers or programme officers. Results would be fed to senior management and the donor.

3.5 Purposefully select which activities to study

The strategies described above, particularly the more intensive ones such as stories of change, PT, CA, OH and GEM, needn't be applied to every activity in the portfolio; activities can be selected as cases for in-depth investigation. This is where triangulation of methods is useful. For example, if data is being collected in a regular and systematic way as described in section 3.3, it may be possible to identify specific cases that can yield useful information from in-depth analysis. Depending on the evaluation questions, different criteria can be used to select cases. For example, Gerring (2007)[32] suggests the following to identify cases:

1. **Typical**. Choose activities that look representative of all other activities
2. **Diverse**. Choose activities that demonstrate the full spectrum of activities
3. **Extreme**. Choose cases which are at the extreme end of the spectrum of activities
4. **Deviant**. Choose cases which look different from the rest; a deviant case aims to better understand and develop a new model of how change takes place
5. **Influential**. Choose cases with an influential combination of factors; an influential case aims to highlight factors which greatly affect the outcomes
6. **Crucial**. Choose cases which are least or most likely to exhibit a given outcome.

Case studies can also be selected on the basis of QCA or Decision Tree modelling to qualitatively investigate claims which are suggested by these techniques. A discussion by Rick Davies on the stages of case study selection, particularly for QCA, is available on the EvalC3 website.[33]

3.5.2 Application for the case study: why, when and who to do it

The case study found that PRSF had an ad hoc approach to selection of which activities to study, and did not select many. An explicit, organised approach to activity selection allows you to answer questions of what, how and so what. The selection process should be developed in the first six months of the programme, after the hypotheses and characteristics have been developed by the team. It might include either selection criteria or an identification of which future activities are likely to be complicated, interesting or high profile in nature. In a portfolio programme it might be more appropriate to develop criteria and watch as (previously unplanned) activities emerge, to select them. This would be driven by the M&E team with input from senior management and researchers or programme officers providing for developing the criteria at the outset.

31 Programmes like CDKN may have advice on suitable intervals.

32 Gerring, J. (2007). *Case Study Research: Principles and Practice*. Cambridge University Press.

33 http://evalc3.net/how-it-works/selecting-cases.

Table 2: Impact possibility continuum

	Intended	Positive unintended	Negative unintended
Foreseen	Planned programme goals	Predicted spill-over effects	Predicted risks or side-effects
Unforeseen	Emergent programme goals	Nice surprise	Calamity, mishap or backlash

Source: Hearn and Buffardi, 2016, adapted from Ling, 2014

3.6 Be explicit about how impacts will be valued across the portfolio

As discussed in section 2.1, for evaluations to be evaluative they have to go beyond describing impacts and analysing their causes. They must make an overall judgement about the quality or value of the impacts and whether the portfolio as a whole is a success and worthwhile investment of resources, which we have called the '*so what*' question. This involves evaluative reasoning to synthesise findings from across a portfolio in order to directly answer high level questions about the portfolio as a whole (Davidson, 2014). This requires two elements in place: specific *key evaluation questions* to direct the process, which should be in place for any kind of evaluation system; and *evaluative criteria*, which define what 'success', 'good performance' or 'high quality' are.

Developing the evaluation criteria is an important step and will require the input of stakeholders to ensure that different perspectives are taken into account. In diverse programmes working on policy and systems change there will be different views of what success looks like and it will be insufficient to develop criteria without identifying these. Having a shared rubric, which has been developed through a collaborative process involving partners, can help to avoid misunderstandings that often arise when dealing with potentially subjective judgments.

One particular challenge is that the impacts emerging from portfolio-based programmes will be diverse and will include positive, negative, intended and unintended changes. Table 2 presents different six different kinds of impact that might be covered in an impact evaluation. Exploring different kinds of impact is crucial to forming an overall judgement and looking at looking at impact in this way can help to *weigh* the different kinds according to their importance in the overall success of the portfolio.

Addressing negative impacts raises the question of how to weigh success and failure of portfolio activities. As discussed in section 2.1, failure can still be valuable as, in the case of some activities, it may contribute to learning, which in turn increases the chance of success of subsequent activities. It can be useful to think about different kinds of failure. Stame (2010) describes three kinds of failure: theory failure (the understanding of the problem and how it could be addressed was wrong), implementation failure (the intervention wasn't implemented correctly) and methodological failure (the assessment approach didn't give an accurate answer). Theory failure has the potential to contribute significantly to learning because it highlights false assumptions or gaps in knowledge, which can transform the ability of other initiatives to affect change. Implementation failure has limited value for learning other than to identify the specific reason why an intervention failed and then to fix it. Methodological failure is much harder to spot but if it is identified it can contribute to more appropriate and accurate assessment methods.

Valuing impacts across the portfolio needn't be an onerous task. Regular 'sense-making' among key programme staff and partners can highlight important trends. If data is being gathered throughout implementation (section 3.3), and cases are selected (section 3.5) and analysed (sections 3.3 and 3.4) then there will be plenty of data to use as a basis for discussions among the senior team. Informal reflection of these data with respect to an appropriate logic model (section 3.1) can provide quick feedback to either confirm or challenge initial thinking or it might throw up surprises, which can prompt creative thinking and new avenues.

3.6.1 Application for the case study: why, when and who to do it

The case study found that PRSF had key evaluation questions in the design document but did not revisit them or use them in the M&E reporting significantly. Evaluative criteria and sense-making across the programme allows impacts to be valued at an aggregate level, to ensure the programme is on track and to identify when things are not working so the programme can readjust. This would be developed in the first three months of the programme, in tandem with the hypotheses creation (by the whole team, with discussion led by the M&E team, and input from senior management, Indonesian government if necessary and the programme design team). The M&E team would regularly monitor and report on this throughout the life of the programme, including to the donor, with sense-making occurring as often as practical.[34]

34 Among DFAT funded programmes, the MAMPU M&E advisor role is an example of how this can work well.

Box 5: Using the collaborative outcomes reporting (COR) approach to evaluate the impact of CIFOR research on policy

The following example describes how many of the strategies recommended in this paper can be applied to a retrospective impact evaluation.

From 2009 to 2015, The Centre for International Forestry Research (CIFOR) conducted a global comparative study (GCS) on Reducing Emission from Deforestation and Forest Degradation (REDD+) with the aim of supporting country and global level strategies to reduce carbon emission in an effective, efficient and equitable way. The GCS is organised around four modules focusing on governance of national climate change policy; sub-national REDD+ projects; emission measurement, reporting and verification systems; and carbon management at the landscape scale, with a cross-cutting module dedicated to the sharing and dissemination of knowledge.

In mid-2014, CIFOR established a team, led by ODI and involving CIFOR staff and other experts, to evaluate the outcomes and impact of the GSC and its influence on policy. The assessment used a modified collaborative outcomes reporting (COR) approach to describe how GCS research outputs and engagements have contributed to expected and unexpected outcomes (Young and Bird, 2015).

Collaborative outcomes reporting (COR) is a participatory approach to impact evaluation based around a 'performance story', which describes how a programme has contributed to outcomes and impacts (Dart and Roberts, 2014). As Dart, the originator of the approach, describes:

> *'COR starts by developing a theory of change about the program or policy. It makes maximum use of existing data ("data trawl") before focusing on additional data collection to fill gaps. It uses the rigorous non-experimental techniques of contribution analysis and multiple lines and levels of evidence to make causal inferences without a counterfactual. It uses both an expert panel ("outcomes panel") and a stakeholder summit workshop to review and synthesise data into an overall evaluative judgement. And it produces reports that are brief, but with links to detailed evidence.' (Dart, 2013)*

Applying this approach, the evaluation team developed a methodology that was highly collaborative and multi-faceted. It followed four stages, the first of which was to plan the evaluation: select the cases to be studied, design the methods and collectively develop the theories of change to be used for the evaluation. The second was the research stage, which included six individual studies: two in depth case studies on two major GCS work streams; a set of light country case studies in countries where the GSC operated; a set of episode studies in countries where the GCS was not operating to assess spill over effects; ten 'stories of change to document particular events or changes which were identified as being influenced by CIFOR; and a communications review examining reach and uptake of GCS research products.

The third stage was a data integration workshop to review the emerging results with senior programme staff, identify further evidence needs and develop the framework for analysis. This included the development of results tables' which presented the evidence from across the six studies according to the key evaluation questions and the elements in the global theory of change. The fourth stage was a sense-making workshop, bringing together the programme staff and other high-level stakeholders (internal and external to the programme) to review the results tables and the tentative conclusions the team had drawn from them and to generate the recommendations. The final report, based on the findings from these two stages was presented and discussed at the annual CIFOR staff meeting.

The team found a number of strengths with this approach, including the strong focus on developing theories of change that could be tested and articulating 'end-of-program' outcomes at the level of changes in discourse and actions in both policy and practice domains (Belcher et al, 2016). They also found the manner in which the approach synthesised and clearly presented evidence from a range of sources particularly helpful for assessing the theory of change. CIFOR has already started to implement several of the recommendations.

There were three main challenges, however. Consistency across the various data collection activities was a challenge because of the number of people involved and the diversity of approaches used. They were not entirely satisfied with the level of confidence in causal relationships which the largely qualitative methods helped establish – but they recognised that getting a stronger confidence was exceedingly difficult in this context. Lastly, the approach relied heavily on detailed theories of change that did not exist, and the theories developed during the evaluation were somewhat simplistic, limiting the analysis which could be done. Overall, the participatory nature of the approach significantly contributed to the learning objective of the evaluation. (Belcher et al, 2016).

Conclusions

This paper has addressed the nature of the programme and its characteristics (Chapter 1), the challenge of evaluating support programmes like PRSF (Chapter 2) and recommended strategies for measuring impact of policy influence portfolio programmes (Chapter 3), which might be applied in future programming. The full in-depth analysis of what PRSF and TNP2K intended to and actually measured in terms of impact is also provided in the form of a subsidiary case study (Annex 1).

The key recommendations for future programmes are to consider six strategies to enhance planning, M&E of impact, as discussed in Chapter 3. These are to:

1. develop appropriate logic models
2. collect observational data throughout implementation
3. develop stories of change or case studies
4. understand causal relationships without a counterfactual
5. purposefully select which activities to study
6. be explicit about how impacts will be valued across the portfolio.

To remain opportunistic and flexible, policy influence programmes need a light-touch system monitoring system. Both MAMPU and CDKN have struck a good balance on how to implement this.

The authors acknowledge that programme management and implementation is determined by much more than theoretical approaches and research. We recognise that there are other factors outside of those considered in this paper that will impact upon future programming decisions. Firstly, there are budget constraints, given there is less money available in current Australian aid programme budget. However, the authors and many of those interviewed do not think reduced budgets should mean that funds are not spent on discovering what works and what is replicable; by contrast, this will in fact reduce costs in future programming and improve the transferability of programmes.

Secondly, important consideration needs to be given to how agile and opportunistic a policy influence programme needs to be. This is directly affected by the cadre of the staff on the programme. Several interviewees commented that success in programmes will rely on the political agility of staff and their ability to 'sense' when uptake and influence are occurring. What was noteworthy about TNP2K is that while there was foreign technical assistance, it was largely supporting or technical inputs (such as JPAL or OPM involvement). The managerial roles were all held by senior and experienced Indonesian staff, who stayed at the helm. Having a team that is structured around key programme relationships is essential to track influence. This means hiring a cadre of staff whose role includes significant liaison and relationship management work. The hiring systems of PRSF were flexible enough to provide this, and they were able to strike the delicate balance of recruiting people with technical expertise as well as with liaison and facilitation skills. Being able to recruit strong staff for a programme of this nature will rely on credibility, prestige, budget and timeframe.

Thirdly, how programming works in reality will depend upon the ability to create appetite to capture this kind of learning within DFAT. There will be a strong need to create a 'culture of enquiry' within the department, to ensure that the management and less formalised reporting incentives are adjusted from more typical programme management.

The approaches recommended in Chapter 3 are only worth applying if the positioning of monitoring within the programme is reconsidered and placed at the centre of senior decision- making on the programme. If the strategies this paper recommends are applied for future programmes but the equivalent support mechanism (in the case of TNP2K, the PRSF) remains outside of major decision-making, then the strategies will have little effect. This will also require better communication of findings and results than was evidenced in PRSF, with more synthesis (the creation of a management dashboard would be useful) and other types of accessible sense-making tools. MAMPU has developed a successful way of packaging their performance story, which provides all the necessary data, in an accessible way, and could be a useful guide. It also relies on strong programme relationships, with access given to and trust in the monitoring and evaluation team. M&E needs to be housed close to the heart of programme decision making for any politically adaptive programme management to occur in real time. In short, what really needs to change is not only the use of different tools and strategies, but how M&E is positioned within a programme.

This paper has attempted to address some of the key elements of this problem, and identified what it hopes are useful strategies to support more effective programmes of this nature. We know that TNP2K was considered a success, but how that success was generated largely remains inside a black box. Unpacking the components of that 'black box' and how to evaluate them is a crucial challenge for future efforts to elucidate. This paper hopes to have taken initial steps towards helping them do so. If we can achieve this, then we improve the replicability, scalability and future innovative efforts of ambitious, complicated yet successful programmes like TNP2K.

Annex A: The PRSF case study

In 2009, the Indonesian government committed itself to *accelerate* poverty reduction, aiming to lower the (stagnating) poverty rate from 14.1% in 2009 to 8-10% in 2014.[35] The government recognised an urgent need to increase efficiency and reduce waste across national social protection programmes (reference). It cited the proliferation of overlapping and sometimes mis-targeted programmes (as many as 90 on community driven development alone) that each had different planning, oversight and accountability systems.[36] There was an urgent need for high-level coordination and strategy.

The National Team for the Acceleration of Poverty Reduction (TNP2K) was established by the Indonesian government in 2010, in direct response to this priority. The TNP2K Secretariat[37] has a mandate to accelerate poverty reduction and strengthen social protection systems by: (i) improving the performance of poverty reduction programmes; (ii) improving programme targeting through common methods and better household listing for all social protection programmes; (iii) undertaking monitoring and impact evaluations of the social assistance programmes; (iv) identifying important but troubled social assistance programmes and resolving their implementation issues.

To meet this mandate, TNP2K undertook a suite of activities to determine policy recommendations[38] to Indonesian social protection programmes.[39] For coherence, these social protection programmes were organised thematically into several clusters (see Figure 1).

These activities included producing relevant desk and field research, conducting pilots, hosting conferences and running workshops. TNP2K also received funding to assist certain line ministries to implement the changes they recommended. Between, 2010-2015, the four key social protection programmes that TNP2K researched and made recommendations for are:

- **Rice for Poor Families** (RASKIN) aims to subsidise rice provision to 15.5 million poor households (with monthly distribution) to combat malnutrition and improve food security.[40]
- **Help for Poor Students** (BSM) is a conditional cash transfer programme to assist students from 15.5 million poor households to meet their basic education costs.[41]
- **The Family Hope Programme** (PKH) is a conditional cash transfer programme for 2.8 million very poor households.[42]
- **Community Health Insurance** ('Jamkesmas') aims to provide free basic health services to 86.4 million poor and near poor individuals.[43]

For more information on how this was structured, see the Independent Completion Review that was conducted in 2015.

In response to a request in 2009 from Indonesia's Vice President, the Australian Government established the Poverty Reduction Support Facility (PRSF) to support TNP2K.[44] It was created to afford TNP2K the technical, managerial and financial support services it needed to fulfil its mandate quickly. This included the provision of basic equipment, staff and premises. Beyond this, PRSF was directed to generate knowledge to inform social protection policies, define policy options, translate policy choices into operational programmes and provide high quality monitoring and evaluation. It would do this by producing research; designing and managing pilot reform projects; supporting reform initiatives undertaken within relevant ministries and agencies; developing and managing the Unified Data Base (UDB); and other DFAT directed activities.

PRSF began with a budget of AU$15 million over four years, but this increased significantly over time to an operating budget of approximately AU$30 million for 2014 alone. Its total expenditure from 2010 to September 2014 was AU$76.8 million. This is five times its originally envisaged budget.

TNP2K and PRSF work in close coordination, with TNP2K taking the policy and technical lead. PRSF in contrast has little effective control and yet is responsible for contracting and administering staff and resources for TNP2K while remaining accountable to DFAT (these issues were discussed in the Independent Completion Report).

35 Indonesia's Medium Term Development Plan (RPJMN) 2010-2014.

36 DFAT Design Document for PRSF, p.5. (http://dfat.gov.au/about-us/grants-tenders-funding/tenders/business-notifications/Documents/prsf-to-end-2014.pdf)

37 Both TNP2K and the TNP2K Secretariat will be treated as the same entity for the purposes of simplicity in this paper.

38 Often in the form of PowerPoint presentations or policy briefs. One example is the 'Grievance mechanism for Scholarship for poor students (BSM) Programme Policy Brief'.

39 There were initially plans to establish new programmes also but given the time frame in which they wanted to demonstrate impact, a strategic decision was made to focus energy on improving existing ones.

40 Rice for Poor Families (Beras untuk Keluarga Miskin).

41 Help for Poor Students (Bantuan untuk siswa miskin).

42 Family Hope Programme (Program Keluarga Harapan).

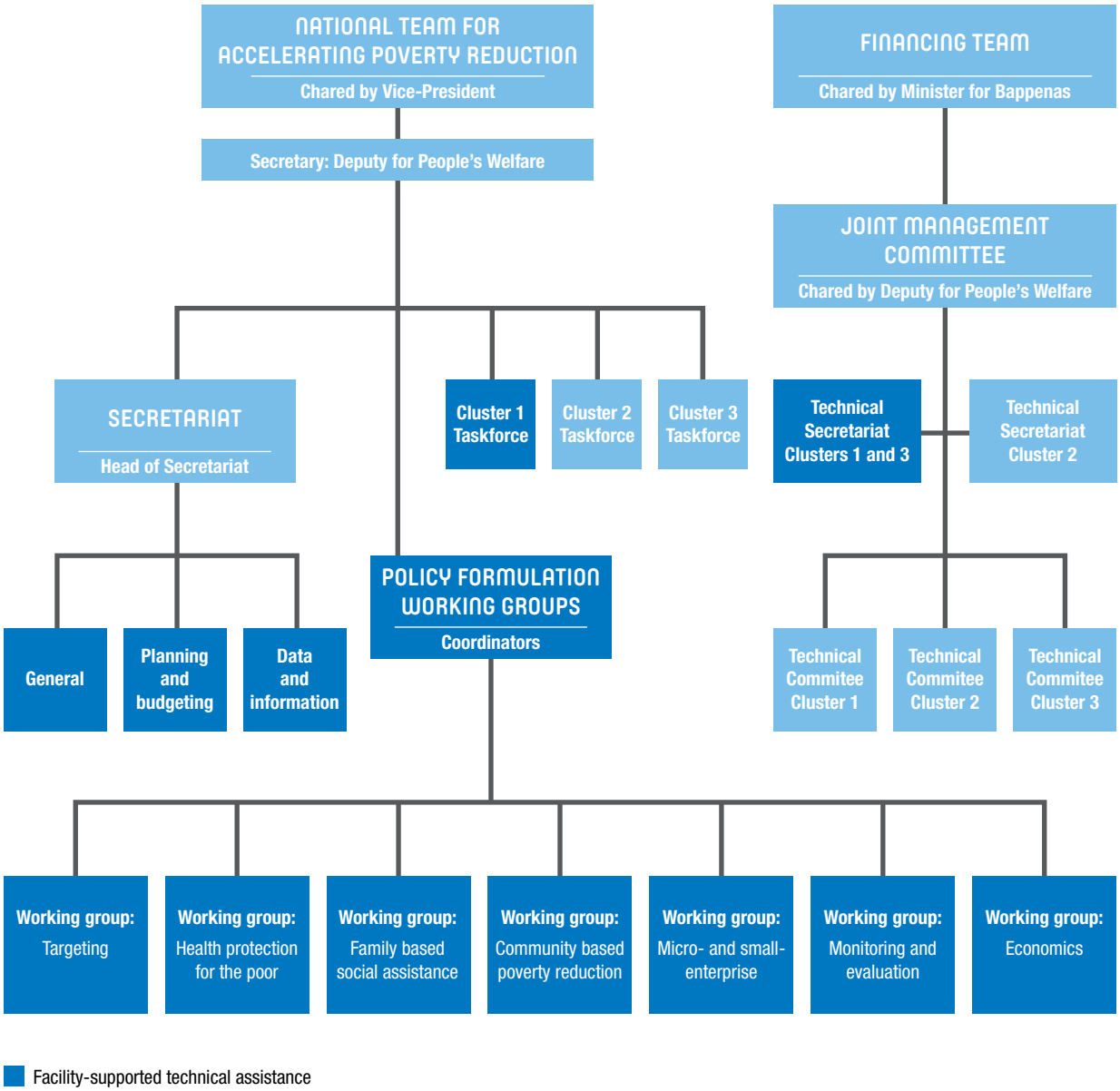
43 Community Health Insurance (Jaminan Kesehatan Masyarakat).

44 DFAT Design Document for PRSF (http://dfat.gov.au/about-us/publications/Pages/poverty-reduction-support-facility-design-document.aspx).

Table 1: PRSF-TNP2K budget over the life of the programme

	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015
BUDGET (AU$)	4 million	8 million	25 million	35 million	30 million

Figure 1: The PRSF technical assistance model



A.I What makes TNP2K a portfolio-based programme

There are several characteristics that TNP2K displays that are typical of portfolio-based programmes:

1. TNP2K (and PRSF in supporting it) has an overarching mandate to accelerate poverty reduction and strengthen social protection systems via four key social protection programmes. To achieve this goal, they conduct a series of activities that are diverse in approach.
2. PRSF is a separate entity that sits under TNP2K and so there is indirect delivery through intermediary agents. In this case, it takes the typical form of the managing contractor working alongside technical and government staff, with the contractor responsible for the bulk of reporting and M&E activities.
3. PRSF's initial theory of change included two macro steps, highlighting the centrality of policy engagement: (i) funded activities by PRSF and TNP2K will support improvements in Indonesian government policies and programmes; (ii) improved policies and programmes will lead to acceleration of poverty reduction.
4. PRSF has dual reporting lines: to DFAT and to the Vice-President's Office of Indonesia.
5. Like many portfolio-based programmes, they were established quickly, in response to a high-level request. Somewhat unusually, they were very well resourced – though what is more common is that their funding flows changed over time. With the broader Australian scale up of international development funding prior to 2014, PRSF and TNP2K's staff and work grew in size, as well as absorbing additional activities (such as the DFAT windows).[45]
6. Political changes impacted upon the programme over its lifespan, with changing government administrations in both Australia and Indonesia – resulting in a shift of policy priorities.
7. Like many portfolio-based programmes, PRSF had a theory of change that was based on the design document, with articulated goals and activities that seemed sensible. However much of the programme logic was not fleshed out beyond this. The question of how and why these activities might contribute to goals was omitted and several assumptions left unexplored.
8. A typical, yet often unacknowledged feature, was PRSF's dual accountability to deliver on the Indonesian government's priorities while delivering value for money to the government of Australia.
9. There was uncertainty across the range of activities about which strategy would prove to be most effective for policy reform, and why or how certain activities deliver better benefits.

A.II How PRSF intended to measure impact

PRSF and TNP2K's initial design suggested that it would measure *what* impact, *how much* impact and, although not explicitly, *how* to learn what worked and did not work to improve programming as implementation progressed.[46] PRSF did not set up the systems to capture these and the intended focus on uptake and policy influence (and particularly how to learn what worked and did not) from the design was lost in implementation. It is important to understand what was intended to be measured in terms of impact from the outset.

Table 2: What PRSF intended to and actually measured in terms of impact

PRSF	How much	How
What did PRSF (and others) **say they would do** for evaluating impact?	Yes – explicitly[47]	Yes – by implication[48]
What did PRSF (and others) **actually do** for evaluating impact?	Yes – to some extent[49]	Not really[50]

45 PRSF ICR, p.8.

46 Constructed of the concept note, the actual design document, the evaluability assessment, the M&E framework, the 2013 Inception Design Team Implementation Planning report.

47 DFAT Design Document for PRSF, pp.6, 8, 13.

48 The use of portfolio-based programming by its nature tests approaches. Plus the key evaluation questions (in design) are a set of hypotheses to test for DFAT – part of the *how* or *why* dimension. Furthermore as a high risk programme which needs regular monitoring and potential course correction, understanding how and why is important. And finally some of the interim objectives (e.g. 'policy advice is realistic and implementable').

49 TNP2K, Bah et al. (2014) *An evaluation of the use of the UBD for social protection programs by local governments in Indonesia*. And the BSM update as well as other studies undertaken by TNP2K explored below.

50 See detailed explanation below of reporting analysis.

A.II.i Measuring the *what* and *how much*

The PRSF evaluation explicitly aimed to measure *what* and *how much* impact was achieved. There are two main records of M&E design decisions in PRSF. First, the PRSF design document (2010), prepared by AusAID, which outlines the broad priorities for M&E and suggests an M&E framework. Second, in 2012, once the programme was underway, the implementation team developed an M&E plan, which documents detailed steps in developing the M&E approach.[51]

A.II.i.i The design document (2010)

As a portfolio programme, the design document does not provide a detailed logic model, since the activities that will make up the portfolio programme are not known in advance. The document instead outlines a broad five-step logic model:[52]

There were several ways that the *what* and *how much* impact questions were addressed. In terms of *what success might look like*, the design document specified that the Indonesian government had already determined several indicators of success, which would be added to as the programme was implemented. Those specified indicators of success were:[53]

- Indonesian government agencies use a unified, standardised database and methodology for poverty targeting
- Indonesian government develops a high-level system for monitoring and evaluating progress in poverty reduction
- Indonesia introduces a single social security card that entitles eligible holders access to services
- Ministry of Health restructures and improves health insurance for poor people
- Ministry of National Education scholarships and other support programmes ensure full K-9 school completion by the poor
- Indonesia's programme of conditional cash transfers is improved and scaled up to approximately 3,000,000 households (currently at 720,000)
- microfinance programmes are consolidated and follow global best practice principles for outreach and sustainability.

The design document specified actual outcomes across two time horizons: longer term and interim outcomes.[54] These all predominantly focus on the *what*.

Longer term outcomes (three years or more):

- Implementation of policy advice improves the effectiveness of social assistance and poverty reduction programme.
- Social assistance programmes are better targeted.
- Poor families eligible for social assistance programmes have reliable access to these programmes.
- Greater appetite within the diverse implementing agencies of Indonesian government to implement integrated poverty programmes.

Interim outcomes:

- Policy working groups produce policy advice that directly influences programme decisions.
- Policy advice is realistic and implementable.
- Evaluations and pilot programmes provide evidence base for policy formulation.
- Gaps in social assistance coverage are identified and actioned.
- AusAID participates in key policy discussions, in technical committees and influences decision making.

Success was to be defined as both 'outcomes… through the programmes they support and by the number of policy proposals acted upon by the Government as a whole,' (*what* and *how much*) as well as 'increasing appetite for evidence based policy making' (a different *what* to measure).[55]

Beyond these indicators there was also the expectation that End of Programme Outcomes would be developed by the PRSF M&E team during the initial months of implementation once tendered. The M&E overview in the design states that PRSF will 'develop a comprehensive monitoring framework to measure the impact and results of the Facility's work' and 'undertake regular monitoring activities… [to] ensure that results are available in a concise and usable form by all participants… to learn from lessons drawn from all poverty reduction activities', which places emphasis on understanding the how questions.[56]

51 Furthermore, half way through implementation, in 2013, it was proposed to scale up PRSF to a programme of nearly AU$300 million. A lot of work was done at that time to clarify how it would work, including what kind of M&E system would be needed. since this scaled version of the programme was never implemented we won't use those plans here, but since they represent sound thinking on how large facilities can be monitored and evaluated we will refer to them in later sections.

52 While these three stages are detailed, the arrows between the boxes and theory of action is not explicit.

53 DFAT Design Document for PRSF, p.6. (http://dfat.gov.au/about-us/publications/Pages/poverty-reduction-support-facility-design-document.aspx).

54 DFAT Design Document for PRSF, p.13.

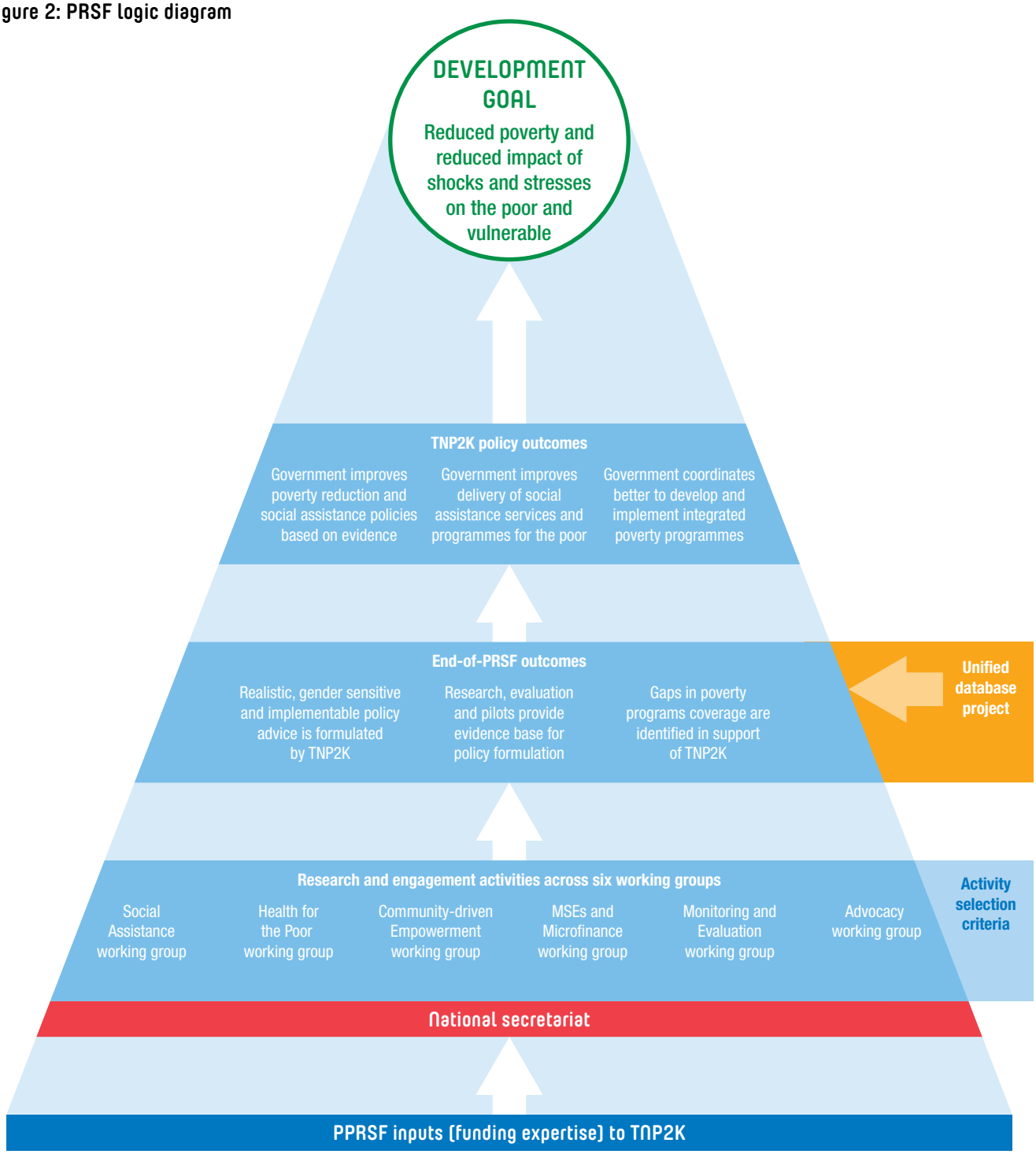
55 DFAT Design Document for PRSF, p.12.

56 DFAT Design Document for PRSF, p.20.

A.II.i.ii The monitoring and evaluation plan 2012

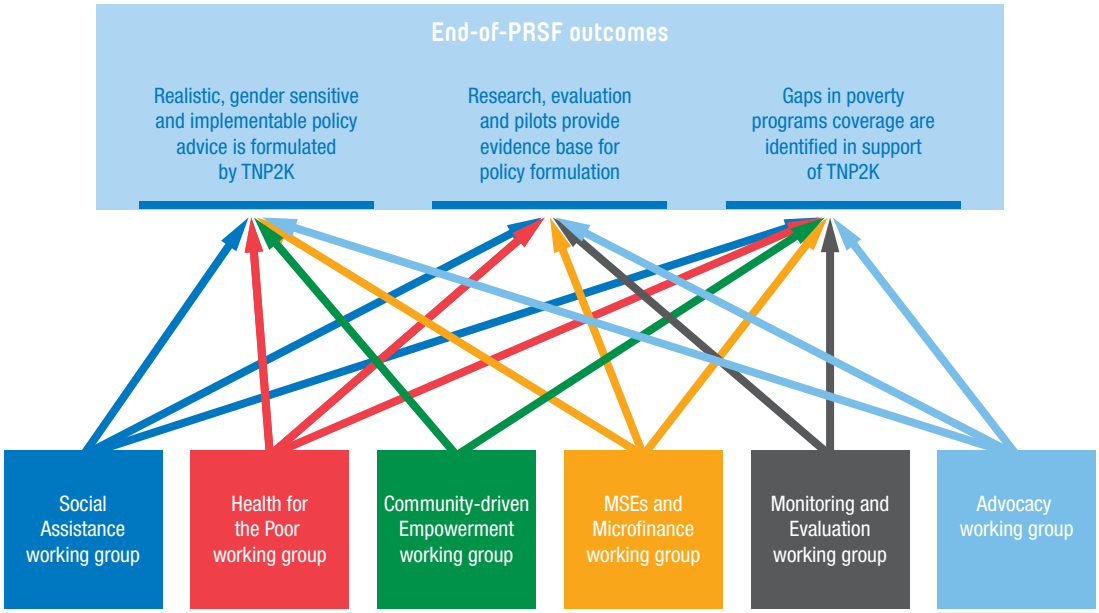
The PRSF M&E team was staffed by a number of GRM International technical specialists, and the M&E Plan (like many programmes) was set up quickly, as there was a rush to get the programme moving in a short timeframe. An evaluability assessment was conducted in early 2012, using goals and objectives outlined in the DFAT design document, which formed a basis for the M&E Plan finalised in 2012.[57] The M&E Plan was a standard M&E framework, using a logframe approach, with somewhat limited analysis of *how the contribution of activities towards goals* would operate (often the arrows in a logframe diagram). The assumptions and theory of change were not explicitly detailed.

Figure 2: PRSF logic diagram



57 PRSF Evaluability Assessment (2012), pp.4-6.

Figure 3: PRSF end of portfolio programme outcomes



The M&E plan purpose was framed around determining policy influence, and to 'track whether evidence based policy advice leads to expected policy outcomes.'[58] The M&E system was aimed at three levels: activity level monitoring, end of facility monitoring and TNP2K policy outcomes monitoring.[59] It contained performance questions of interest, which were on very relevant topics like '% of evidence-based policy research leading directly to recommendations for implementation'.[60] The focus was on contribution rather than attribution,[61] with clear indicators and users outlined.[62] PRSF described the way the activities would be evaluated, with activities such as field visits, key informant interviews, interviews and an information system.[63] All of these examples pertain to the *what* and *how much* questions.

PRSF also proposed to attempt to understand causes of impact. It described how it would collect and analyse the data to answer causal questions about impacts observed (the *how*). This could be seen in the M&E Plan's case study approach.[64] PRSF, like many programmes, did not have an explicit plan in place for how to synthesise learning about causality across the programme (at least no written record was clearly observable). PRSF did plan how it would report and use the information though which is part of the way to synthesis: reporting and dissemination and use.[65]

Thus, in the design document and the M&E plan, PRSF had some explicit aims to address the *what* and *how much* questions of impact for the programme.

A.II.ii Measuring the how

PRSF planning documents do not explicitly state that they would measure how activities work. However, it is implied in the stated M&E purpose, as well as the nature of several activities listed. It is implied through (i) the nature of the modality, (ii) the fact that it was rated as a high risk programme, (iii) the key evaluation questions in the design (which were a set of testable hypotheses) and finally (iv) in 2013, the inception design team noted that measuring *how* was an explicit gap in the necessary operations of PRSF's M&E system. Thus, DFAT and PRSF came towards *how* questions and their importance in what TNP2K was trying to achieve, though this was late and became a sort of retrofit priority.

As discussed above, the very nature of using a portfolio-based modality has implications for evaluating impact. Portfolio-based programmes trial different interventions towards a singular goal, which are largely unknown from the outset, so their M&E needs to determine which pathways are most effective and efficient in reaching that goal. In terms of evaluating impact this

58 PRSF M&E Plan (2012), pp.5, 12, 18, 22.
59 PRSF M&E Plan (2012), p.13.
60 PRSF M&E Plan (2012), pp.16-19.
61 PRSF M&E Plan (2012), pp.22-23.
62 PRSF M&E Plan (2012), pp.16-19 (indicators) and pp.24-25 (users).
63 PRSF M&E Plan (2012), pp.14, 18, 19, 23.
64 PRSF M&E Plan (2012), pp.19-20.
65 PRSF M&E Plan (2012), pp.21-22, 24.

requires understanding *what* each intervention or activity achieves for who, *how much* is achieves, as well as *how* (to answer efficiency and longer term effectiveness). It also requires a meta-level comparison across the different interventions, once these other questions of impact have been determined. This can be called the 'BEST PATHWAY' question.[66] It compares across the different *how much* and *how* answers to see which of the pathways delivered the best comparative benefit under the circumstances. This is implicit in the approach selected, but there is also more detail in the design itself and surrounding documentation.

A.II.ii.i The design document 2010

The design categorised PRSF-TNP2K as overall as a high risk programme,[67] and discussed that it would need to be closely monitored throughout implementation to avoid any negative impacts. Furthermore, this suggests that it would need to be able it to adjust activities throughout implementation if required. This is in fact the benefit of facility programming.[68] This also goes to the *how* things were working.

Beyond this, the design stated that the 'majority of the Facility's M&E resources will be expended on… "**Testing the hypotheses on which the facility logic (and National Team logic) is founded** – see section on *Key Evaluation Questions*."'[69]

The Key Evaluation Questions are actually a set of underlying hypotheses to test. This also goes to the heart of the '*how* activities are working' dimension of impact. The design states that these 'should be made explicit so that ongoing evaluation can assess whether the programs supported through the Facility match AusAID's expectations'. From these major hypotheses (herein outlined) a set of key evaluation questions were to be drawn during implementation. From this list and broader consultations, hypotheses were to be selected and regularly assessed over the life of the facility:

- A smaller number of larger poverty programmes will be **more effective** at reaching the poor than a larger number of smaller programmes.
- Improved targeting measures will let programmes **reach** more of the poor.
- **Social capital** investments can improve the efficacy of poverty targeting.
- Special targeting measures and reforms will allow currently **excluded households** to gain access to safety net programmes.
- Targeted programmes will be more **cost-effective** than universal safety net programmes.
- Programmes that **smooth out shocks** can prevent families from falling into poverty.
- Smoothing shocks and reducing stresses will produce positive **second generation impacts** by helping to break the intergenerational transmission of poverty.
- **Combining asset-based and transfer programmes** will maximise poverty reduction impacts because different vulnerable populations benefit.
- Increasing **competition** among service providers for safety net programme improves their efficiency.
- **Direct** transfers are more effective than providing services in kind.

Importantly, when planning a successor for TNP2K (for after 2015), the inception design team highlighted that what was needed was analysis of how these activities were working and being able to make comparative, informed value judgements about the best pathway to the goal. All of these elements in the design and subsequent documentation suggest that it was important to understand *how* the activities were working and why, as well as whether they could work more effectively.

This ends the overview of what PRSF (and others) had planned to measure in terms of impact. What actually happened in reality once the programme was being implemented? It is important to remember that there were dramatic changes to the programme, not least a huge increase of funding and expansion of activities, in the first two years.

A.III How did PRSF actually measure impact?

What actually happened during implementation, as opposed to what was planned, has been derived from extensive interviews and documentation such as the inception report, mid-term independent progress review, and the independent completion report. Additional supportive evidence of implementation planning was drawn from 2013 investment proposal (for a AU$300 million scaled up version of PRSF). The analysis in the following sections is separated into formal and informal impact evaluation systems – and goes through the different approaches.

66 These terms are broadly based on Befani (2016) on the algorithmic approach, and are used because they are every day terms that do not require a strict background in evaluation to use/read.

67 DFAT Design Document for PRSF, p.13.

68 DFAT Design Document for PRSF, p.13.

69 DFAT Design Document for PRSF, p.21.

A.III.i Formal impact evaluation systems

The PRSF programme had substantial financial investment and significant reporting requirements. Reporting outputs were detailed and high quality. TNP2K, PRSF and DFAT each had several formal M&E mechanisms. However, while several detailed and thorough studies were conducted,[70] they were predominantly geared towards informing policy recommendations rather than assessing the programme's own impact, nor did they engage heavily in understanding how and why activities worked for policy influence.

A.III.i.i Summary of TNP2K formal impact evaluation systems

- Very comprehensive activity, output and quality monitoring. Significant amount of reporting produced.
- A good deal of evidence and analysis for *what* and *how much*, analysis of *why* for adjustments/policy recommendations, but no analysis of *why* or how for their own ways of working.
- The tacit good management of *how* (the programmes perceived overall success in influencing policy) was likely provided through good people with strong experience, networks and understanding of the context.

TNP2K had an M&E working group and cluster teams to conduct research and evaluations to test proposed changes to the large social protection programmes, which would inform TN2PK policy recommendations to the Vice-President. The amount of quality research and analysis produced was impressive. However, while the M&E working group used M&E approaches, its work was primarily to conduct research to inform policy recommendations, rather than assessing the impact of their own research uptake or policy influence as a think tank. In other words, the evaluation answers *what* adjustments they should recommend to Line Ministries based on *how much* impact certain activities were having, but did not measure the impact of those recommendations once implemented.

There were a few notable exceptions where TNP2K did conduct analysis on its impact. One example was the 2014 qualitative study investigating the uses of the UDB.[71] This study looked at the number and nature of requests for data from the UDB, user satisfaction, socialisation of the UDB, procedures to access UDB data, additional needs for support in using the UBD, and recommendations for future. This goes to the *how much* question of the impact of this activity, and some way towards the *how*. There was also the 2013 cash transfer for poor students programme (BSM) Policy Brief update which revisited the recommendations that had been made by TNP2K to see how they were being implemented. The reforms included adjustments like better targeting to increase coverage of students from poor families, and a change to the timing of BSM payments to align with the academic year. The monitoring found several implementation issues:

> *'In particular logistical delays and geographical barriers, as well as incomplete information of school aged children in Unified Database, that together caused a lower than expected take up rate of BSM cards for Junior Secondary school.'*

It discusses refining the targeting methodology to allocate quotas based on poverty incidence, age ranges, drop out and discontinuity rates, and education access variables in each district. It proposes learning from lessons during the first phase of reforms to improve the targeting. This is where the research to inform policies overlaps with the *how much* and *how* of their activities. Its purpose was to inform further recommendations, but it involved assessing their own impact.

In fairness, the programme design states that this type of impact assessment would be left for ministries themselves to measure: 'the Ministry itself will monitor whether the implemented work programmes are successful.'[72] So it was, according to the design, not TNP2K's role to assess their impact for the sake of M&E or reporting to DFAT.

Their analysis (which identified gaps in existing social protection programmes which they then tried to help fix – as in the example given) also goes some way towards the *how* of their own approach, in terms of which adjustments to recommend to the Vice-President's Office (to recommend introduction of identity cards rather than other adjustments). There was clearly a process by which recommendations were prioritised over others based on this research – determining which pathway would best lead to the goal of poverty reduction – but no explicit documented prioritisation process is available. However it is not technically a formal impact evaluation system of their own work; it was fundamentally used to inform what they would recommend.

The Indonesian government markers that were identified in the design as 'success indicators' (aforementioned), were to some degree met. An indicative sample of these include:

70 www.povertyactionlab.org.

71 TNP2K, Bah et al. (2014). This looked at the number and nature of requests, user satisfaction, socialisation of the UDB, procedures to access UDB data, additional needs for support in using the UBD, and recommendations for future.

72 DFAT Design Document for PRSF, p.21. Clause 30 (M&E).

i. Indonesian government agencies use a unified standardised database for poverty targeting. **Arguably achieved** – recorded data from PRSF shows the UBD was being used by Line Ministries and many requests (over 600 by 2015) came from local government also.
ii. Indonesian government develops a high level system for M&E of poverty reduction progress. **Unclear** whether any progress was made on this – from preparatory work for this report.
iii. Indonesia introduces a social security card that enables eligible holders to access services. **Achieved** – the unified social protection card was issued (KPS).

These go to the *what*, rather than the *how much, how* or *why* but are important markers of success that were identified in the design from outset.

In the 2013 implementation planning report by the Inception Design team, M&E was described as important for understanding the highly complex political environment in which PRSF and TNP2K operate, and thus trying to understand (and later predict) what types of activities would get the most traction.[73] The report **specified a need in future for a mechanism that would trial different types of interventions in an attempt to learn more about what type of assistance is most likely to work in this context** (because *it was seen as lacking* from the programme in 2013). This means it will be important for future programming aimed at country systems building through portfolio-based programmes.

So TNP2K for all its excellent research could not be said to contribute comprehensively to assessing the impact of their own work – in either *how much* or *how* terms. At least in explicit documented terms that were reported to DFAT, there was not any formal system recorded that could be said to evaluate their own impact, though no doubt much tacit work was being done to direct the programme's energies by the members of senior management (including decisions about BEST PATHWAY).

A.III.i.ii Summary of PRSF formal impact evaluation systems

- An impressive effort to produce documentation and track progress across a vast and dynamic programme.
- Some *how much* analysis that is quite good – which relies heavily on work produced by TNP2K. Very little *why* or *how*, other than a few case studies (which tell you the *how* retrospectively of one activity, not the laboratory *how* across the group, nor in time to course correct).

PRSF reporting against this M&E system throughout the life of the programme was intensive. PRSF produced regular quarterly progress reports reporting on PRSF's progress, via TNP2K activities (not being wanting to be seen to evaluate an Indonesian government organisation). These quarterly reports evolved over time and became more detailed – with increasingly available information from TNP2K activities, and requests from DFAT about the style and content of reporting.[74] By late 2014 the quarterly reports included an executive summary which gave a poverty update and overview of progress, followed by a section detailing results achieved that quarter, and one detailing the work plan for the next quarter as well as annexes reporting against the outcomes monitoring matrix with and listing the research studies and evaluation outputs that TNP2K had produced.

The executive summary's poverty update showed whether the national poverty level had risen or fallen during the quarter, and projected estimates for coming months (based on World Bank analysis). This goes to the *how much* question of impact (in a 'before and after' type counterfactual approach), however cannot account for TNP2K contribution or the exclusion of any interfering factors (of which there would have been many). The poverty update also included assessments of the increased number of households being reached, and higher benefit levels delivered by social protection programmes over time (from year to year) implying a direct contribution by TNP2K policy recommendations. How this contribution occurred relied on TNP2K documentation of recommendations and then resulting changes in figures from Line Ministries, the Indonesian National Social Economic Survey (Susenas) and the Indonesian National Bureau of Statistics (BPS). It was a somewhat implicit line of causal inference that was rarely teased out.

The quarterly reports also detailed the results achieved over that quarter (against outcomes) which in part answers elements of the *how much* question also. The matrix of results is quite extensive in years 2013-2014, and documents a wealth of information about progress across cluster areas or cross cutting issues. It lists the activity funded, what progress could be identified against that activity (for example the number of data requests for UDB from local government to date), and then key findings, policy implications, any issues and lessons learned (all in succinct, dot point form).[75] These collate and curate the results of TNP2K work and are often not original research or evaluation on the impact of the programme. For example, one section reports the key findings of the UBD use by local government, based on the report that had been

73 PRSF Implementation Planning: Final Report Inception Design Team (2013), p.20.

74 Increasing from approximately 40 pages in late 2013 to 75 pages in late 2014.

75 Quarterly Progress Report, October – December 2013, p.6.

completed. It essentially summarises decisions made by TNP2K about policy recommendations, based on their research and flags any issues that might be challenging in future (like how observable success might be). These can be useful in showing what TNP2K was recommending, from which information can be gleaned for the *what* and *how much* impact TNP2K had, including their policy influence but such conclusions are not drawn. It is simply implied in the documentation that they are achieving policy traction and will continue to do so based on the merit of their work. This PRSF quarterly reporting does not pick up the fact that TNP2K recommendations to Line Ministries are rarely followed up on to determine actual impacts once reforms are implemented. Nor does it consider which of the activities is more successful in achieving the goal of accelerating poverty reduction, or any other analysis of the laboratory that TNP2K was designed to be. It should be stated that there are a vast number of activities reported on, which would have been a significant workload to track in itself – which leads to implications for resourcing.

In the annexes to the quarterly reports is a set of information that **does go towards measuring the policy influence** element of the programme, rather than the impact at the end of the line (i.e. the impact on people's lives as a result of the reforms to social protection programmes). These annexes on outcomes monitoring first appear in mid-2012 reporting and are phased out again in 2014. They detail not only poverty indicators from a baseline (2010) to the current quarter (depending on the report),[76] but also the number of key poverty policies and programs which have been changed as a result of research feedback and evidence brought up by TNP2K. For this they work from a baseline (of zero in 2011) and report the number of policies changed as a result (For example, five in Q2 2012). The annex also details the number of UDB data requests received and responded to as well as the number of initiatives organised/led by TNP2K aiming at fostering dialogue and debate on evidence-based policy-making. It is restricted to again using a type of before and after counterfactual, rather than any other tools. These are very clearly in the *how much*, quantitative side of measuring policy influence rather than any qualitative or *how* aspect of impact, but it is the only observable reporting on this element of the programme. That means it is the only reporting found on the sphere of control element of the TNP2K programme (highlighted in red in diagram below).

One important factor in the M&E reporting for the programme was that PRSF was operating within a three-tiered system: whereby PRSF was to conduct M&E on TNP2K activities (which they were not always directly involved in, but rather monitoring when invited into forums), for reporting to DFAT. This was problematic. They were sometimes seen as 'external' partners when there were sensitive political or policy issues being discussed. It meant that access to understanding how and why activities were working, or indeed, assessing any failures, was very challenging.

Beyond the quarterly reporting, one or two important documents were produced by PRSF in 2014-2015 assessing impact of the programme on the beneficiaries. One example is the Value for Money assessment of social welfare benefits delivered by TNP2K, written by PRSF. This entailed analysis collated by PRSF on the benefits (implicitly attributed to TNP2K) produced by reforms on four of the key social protection programmes. The report indicated that there had been a very high return for investment on interventions by TNP2K, depending on the programme, its coverage and the nature of benefits and the accuracy of targeting. As the ICR summarised, 'for every dollar DFAT spent, between $28 to a high of $487 of benefits were generated that would not otherwise have occurred …[and] add up to a net present value of between $345 million to $2.48 billion over a 5-10 year period.[77] Data for this report was sourced from both TNP2K and Susenas.

Table 3: Return of investment for PRSF support and TNP2K reforms

	DFAT investment	Every $1 invested yields:	Net present value
RASKIN	$12.2 million	$28	$345 million (2013-2017)
PKH	$23.3 million	$57	$1.33 billion (2012-2021)
BSM	$10.6 million	$157	$1.6 billion (2012-2021)
BLSM	$5.1 million	$487	$2.48 billion (2013-2017)

Source: PRSF 2015

76 Indicators include: (i) Poverty headcount ratio at national poverty line (% of population), disaggregated by gender and residence, (ii) Poverty gap at national poverty line (%), (iii) Inequality rate (GINI index) and (iv) unemployment rate (%). See Quarterly Report 2 (2012, p.43).

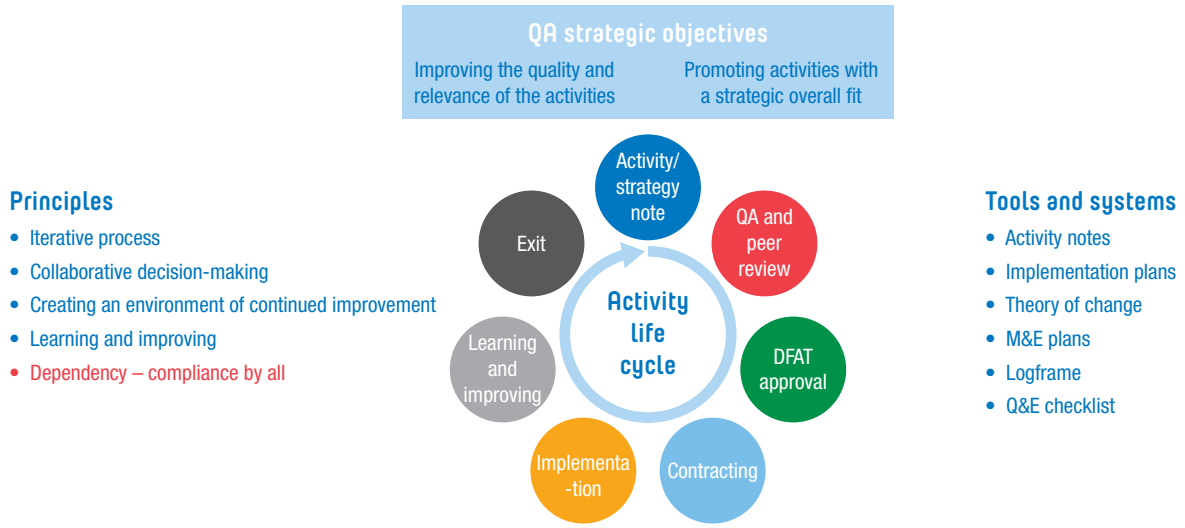
77 ICR, p.13, Ashcroft, V. (2015).

The ICR states that results show that benefits to the poor have been real and measurable – in the sense of *how much* positive impact was generated. This report also uses a comparison of *how much* was measured across the different reforms (although it does skip the why or how) and makes a pseudo ranking of which activity delivered better welfare returns benefits across the facility – estimating the best pathway or most effective investment across TNP2K activities.[78]

In 2012, when TNP2K was scaling up its work significantly, PRSF and DFAT identified that a much more robust quality assurance process was needed.[79] PRSF recognised that 'given the high profile and technically complicated work involved, TNP2K would benefit from an internal but independent process for quality assuring concepts, ideas and products.'[80] This QA process is captured in Figure 4.

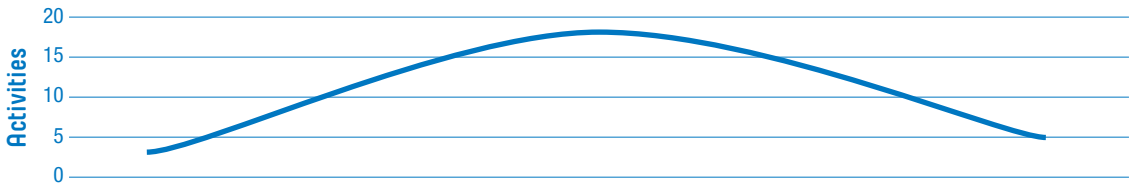
The QA process essentially became a kind of evaluation tool for monitoring BEST PATHWAY decision making. It was strengthened over time and today is accepted by almost all staff as an important part of TNP2K and PRSF's internal processes.

Figure 4: Quality assurance process strategic objectives



Source: PRSF 2015

Figure 5: Phases of the quality assurance process in PRSF



Phase 1 - Initiation (2011-2012)
Identifying the need for QA launch preliminary QA processes
Internal activity selection criteria
One-step process

Phase 2 - Expansion (2012-2014)
First revision of the QA system
Activity guidelines
Role of the mid-term review
External peer review mechanism
Investment designs for high risk
Second revision of the QA system
Two-step process

Phase 3 - Consolidation (2014-2015)
Expansion vs. consolidation
Apply additional strategic filters
Handover planning
Gender and disability lens
Focus on implementation planning, risk assesment
What is the value for money?

Source: PRSF 2015

78 As outlined in the beginning of this section, TNP2K and PRSF were trying to measure *what, how much* and *why* or *how* – as well as the laboratory effect, or best pathway, which is a cumulative product of these other questions.

79 The establishment of the QA System 'coincided with the arrival of the Social Protection Specialist in mid-2012 and was recommended to be strengthened by the independent progress report in early 2013.' ICR, p.16.

80 ICR, Ashcroft, V. (2015), p.16.

However, other than this analysis, the Inception Design team explained that the documentation was heavily weighted towards activity or output level reporting. They stated that a 'much more useful focus would have been on whether and how these activities led to changes in policy and practice.' Acknowledging that this was picked up in the TNP2K reports, but those remain isolated examples of success, and were not integrated with tools like outcome mapping. In its recommendations the 2013 report said that additional questions ought to be the focus of a future M&E plan, such as:

- Which type of studies/activities had the most influence and why?
- Were key targeted individuals engaged appropriately throughout the cycle of knowledge creation?
- When outcomes around improved policy and practice did happen, what happened, and what was the role of PRSF in this?
- To what extent were projects done in a manner to maximise the chance of uptake/influence?

So PRSF for all its strong research could be said to contribute significantly to assessing the impact of TNP2K's work in terms of *how much* (though considered heavily output focussed) but not in terms of *how* impact was generated – or the heart of the laboratory/facility role.

A.III.i.iii Summary of DFAT formal impact evaluation systems

- To a large extent monitored reporting that PRSF produced, provided strategic direction and had regular consultations with both PRSF and TNP2K including on formal mechanisms like the Steering Committee.
- DFAT also was interested in the *what* and *how much*, steered towards the how, through informal mechanisms and good people.

DFAT also conducted several studies that spoke to the evaluation of impact on this programme, including an early report by the Office of Development Effectiveness (ODE) and their own Independent Progress Review (IPR) in 2012-2013 (led by Steve Ashley (IDL Group) Francesca Bastagli (ODI) and Gatot Widyanto (management specialist).

The ODE report was a case study to underpin a broader report being conducted in AusAID at the time, *Thinking and Working Politically: An evaluation of policy dialogue in AusAID*. It was an interesting overview of what the programme planned to achieve, and several building blocks that can be used as a framework when assessing policy influence, but was written in the early stages of implementation (mid 2012 – the second year of PRSF implementation), so was unable to assess impact.[80] It explains what TNP2K aims to achieve, and what AusAID's role is in supporting and financing this programme, and has a one page bullet point list of achievements for the programme to date. It is not a detailed evaluation of impact, in that it does not assess *how much* nor the *how* question.

The IPR was tasked with three key evaluation questions which go more to the management of the programme than to assessing impact or *how* its activities are gaining success, or the merits of taking different pathways to achieving policy influence. The three evaluation questions for the IPR were: (i) Is the PRSF on-track to achieve its expected outcomes? (ii) How effective are the TNP2K, PRSF and AusAID management arrangements? (iii) What lessons can we learn to inform remaining programme time, and a possible scale-up of Australian support?

The IPR provides very helpful strategic insights and management accountability lessons which were taken on board and all largely implemented by DFAT. The M&E of policy influence however is lightly touched on, which is understandable given the scope. The section on 'learning' in PRSF states:

> *'There are few of the formal mechanisms for learning and sharing within the PRSF/TNP2K that would be required if it was to be considered a learning organisation. The culture at present is not one of sharing, questioning, thinking, learning, using information.*
>
> *Equally, there are few formal systems to ensure that learning is systematically used to enhance programme performance. Given the complexity of what is being attempted, and the challenging context in which it is taking place, the IPR suggest that all opportunities to reflect on what is and is not working, and how it might be improved, should be taken. But this requires solid systems to ensure this is well-planned and effective.'*[82]

There is limited available follow up guidance about how to implement these systems and the types of tools that could be used or approaches.

81 http://dfat.gov.au/aid/how-we-measure-performance/ode/Documents/case-study-tnp2k-fa.pdf.

82 PRSF IPR, p.ix.

A.III.ii Informal M&E mechanisms

There were a lot of M&E systems beyond this PRSF M&E Framework that DFAT had in place which were not formally acknowledged as such. For example, DFAT placed a social protection specialist within the TNP2K and PRSF offices two days per week from mid-2012. They also had regular meetings with TNP2K to discuss progress, as well as attending conferences and workshops on several cluster topics. These opportunities to reflect on successes, failures and what was working (perhaps how) were all part of the monitoring of the programme, which was largely verbal or not documented.

One excellent example of how this informal system helped to strengthen the formal M&E was when the social protection expert that had been semi-seconded to PRSF and TNP2K saw a gap in the QA systems. This ultimately led to the QA process introduced in 2012 that became one of the stronger tools for eliciting discussions and assessments of how different approaches were working.

A.III.iii Conclusions for what was implemented in practice

Largely PRSF's focus was on monitoring, less on evaluation. It was predominantly also activity based reporting and lots of paperwork. There was limited results-based reporting, analysis, or sense-making. What M&E tools TNP2K did use were in the form of research in order to make recommendations to VPO, not predominantly of their own work – which they had little interest in evaluating (as this was seen as PRSF's role). This meant they did not explicitly capture/document how or why activities achieved uptake in terms of policy influence.

Table 4: Listing the outputs against impact evaluation questions

STAGE	*How much* impact	*How* impact occurred
1. Relevant evidence based policy advice is produced by policy working groups	Strong volume of evidence. Hundreds of documents (all of high quality), contained in the TNP2K records management system.	Not relevant
2. Research informs policy	Some measurement, for example: 1. UDB usage (by TNP2K) - quantity and quality. 2. Number of policy changes made as a result of TNP2K recommendations (PRSF) – numerical only. 3. Opportunities to foster dialogue, eg 108 fora in 2013. (PRSF) – numerical only. 4. % of evidence based research leading to recommendations for implementation. Eg 0% in 2011, 61% in Q4 2013, target for 2014 90%.[83] (PRSF) – numerical only. 5. % of research findings leading to policy advice. Eg 0% in 2011, 100% Q4 2013, target was 80%.[84] 6. % of research feedback on gaps that have supported policy advice developed by TNP2K. Eg 0% in 2011, 74% in Q4 2013, target for 2014 80%.[85] 7. Medium term outcomes (eg policy advice is realistic) - design indicators only, not implemented.	Not measured 1. Reference to coordination mechanisms, but little info beyond that meetings occurred.
3. Existing programmes are improved and new programmes are created	Limited measurement, for example: 1. GOI indicators of success: e.g. social security card introduced; Ministry of Health restructures health insurance; microfinance programmes consolidated. (TNP2K) – ad hoc reporting – not systematically reported against. 2. Longer-term indicators, nominated in design, e.g. Social protection programmes are better targeted. (PRSF) – some ad hoc reporting on this.	Limited measurement, for example: 1. BSM update (2013) (TNP2K) – quantity and quality. 2. QA system for selecting activities.
4. Policy changes positively affect the lives of poor people[86]	Some measurement, for example: 1. BSM update (2013) (TNP2K) – quantity and quality. 2. Poverty update in quarterly progress reports (PRSF) – numerical only. 3. Longer-term outcomes (e.g. increased GOI appetite for integration of social protection programmes) – design indicators, not implemented. 4. VFM – measured well in two ways late in the programme (by PRSF at DFAT request).[87]	Limited measurement: 1. Pilot studies – e.g. RCTs (also categorised above in stage 1 as 'research'), assess how to modify social protection programmes.
5. Impact from shocks and stresses on the poor and vulnerable are cushioned and poverty reduction is accelerated	Yes. Good info available. For example reporting by the World Bank, National poverty indicators, produced by BPS and other national agencies.	Not measured

83 Quarterly progress report, Q4, 2013, p.33.

84 Quarterly progress report, Q4, 2013, p.33.

85 As above – unclear how this differs from the previous monitoring.

86 According to the design, this was Line Ministry role to monitor.

87 Value for money assessment; and the Quality Assurance mechanisms installed by social protection consultant from DFAT.

Table 4 indicates the five stages of the PRSF programme logic (drawn from the five step model above), and presents information on what documentation there is to support impact evaluation across these, in the categories of *how much*, and *how* impact occurred.

A.IV Why there was a difference between what was planned and implemented

A.IV.i The how is really difficult to measure

Here is an example of *how* the how is challenging to measure…

One activity conducted by the rice subsidy programme RASKIN was to introduce identity cards for participants. TNP2K wanted to know if the introduction of identity cards would have an effect on uptake of the rice subsidy. The Abdul Latif Jameel Poverty Action Lab (J-PAL) led a counterfactual analysis with TNP2K.[88] 572 villages took part in the study, with 194 being randomly assigned to the control group, which was the original programme design with no identity cards, and the remaining 378 villages were assigned the modified programme with identity cards. The study found that the introduction of ID cards led to a 26% greater uptake of benefits. This led TNP2K to recommend that the Line Ministry adjust the programme to include identity cards. It showed the *how much* as well as *how* at this micro level. And so counterfactual analysis seems to have worked well (see Figure 6).

However, if we zoom out and take into account that TNP2K was not only introducing identity cards into its rice subsidiary programme, it was undertaking another intervention at the same time – improving targeting of the Unified Database (UDB) – undertaking counterfactual analysis becomes more complicated. TNP2K were able to compare the benefits of the RASKIN programme by looking at the benefits received (in the same locations) before and after the interventions. So the counterfactual was using the same location, at a different time. The multiple interventions makes it more difficult to determine how impact was achieved, as there are multiple factors to take into account. The challenge is to be able to identify possible alternative explanations of the measured changes and rule them out (for example, the many political or other contextual changes over this time frame). This could be done in theory, but would require a lot of work on the part of the programme M&E staff. So, again a counterfactual at this micro level is theoretically possible within the programme, but largely for answering *how much* impact was achieved, rather than how impact came about.

If we step back even further to take the full programme into account, it becomes even more complicated. As well as multiple interventions within the RASKIN programme, TNP2K was working on multiple social protection programmes.

Within the policy reforms recommended by TNP2K (for example, adjustments to programmes like RASKIN, or the creation of the unified database), it is important to remember that they also impact upon each other (the creation of the unified database helped with targeting on reforms to the RASKIN programme for example). There is interplay between all the activities, and failures and successes across the group – failure in one activity may lead to increased success in another activity for example.

Figure 6: Counterfactual analysis of impact of programme improvement on beneficiaries in a single intervention is possible

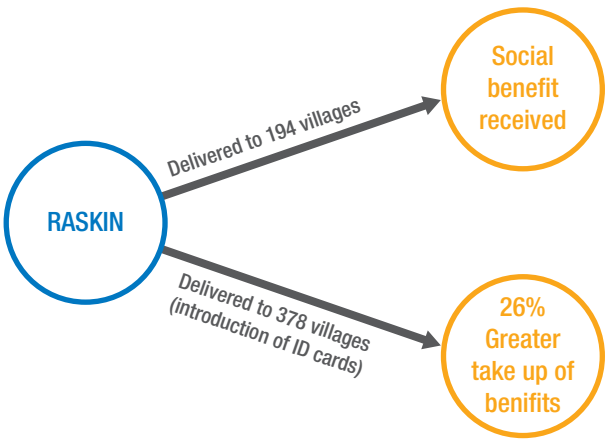
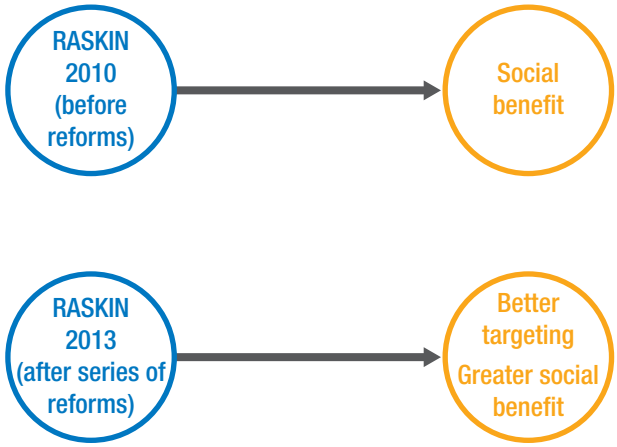
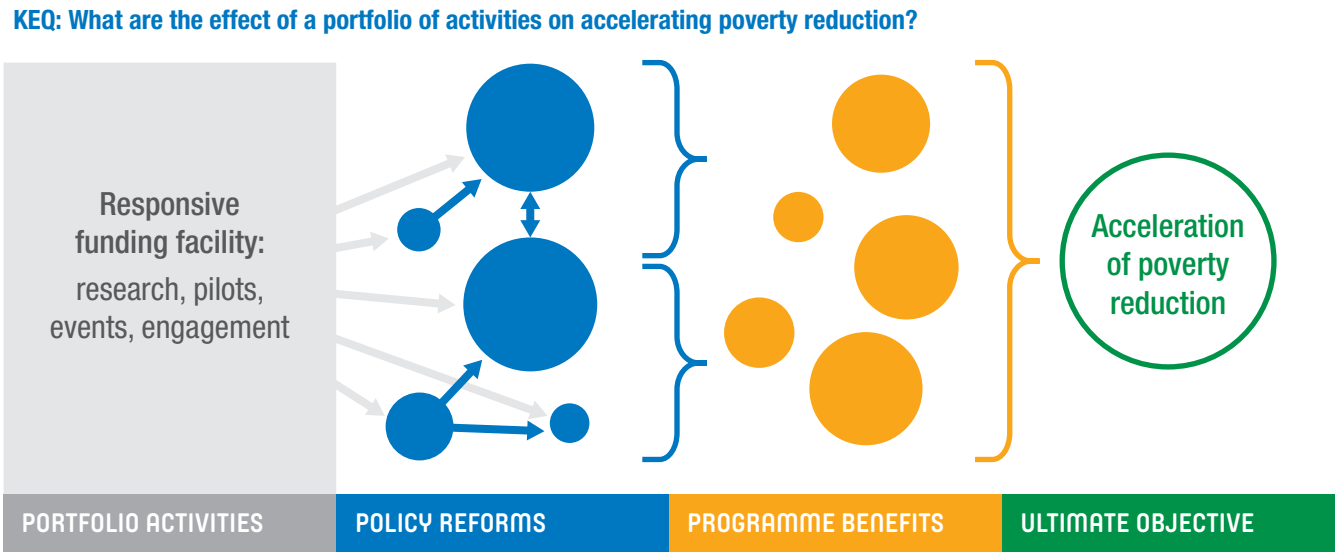


Figure 7: Counterfactual analysis of impact of a series of programme improvements on participants may not be possible



88 www.povertyactionlab.org/scale-ups/raskin-improving-targeting-and-distribution-subsidized-rice.

Figure 8: Counterfactual analysis of impact of a portfolio of activities on beneficiaries is not possible



This makes the key challenge explicit: PRSF is not just interested in improving a single programme, but improving the **entire social assistance sector** to accelerate poverty reduction. It does this through a flexible and responsive portfolio of programmes and activities. This is not like, for example, a standard immunisation programme using a single vaccine, with a clear cut theory of change and manageable control group. Activities are unknown at the start and there is risk involved – some things will work as planned, but it is expected from the outset that some things will not.

A.IV.ii Additional factors

This report acknowledges the need for pragmatism and the realities of programme delivery. There were several realities that geared the M&E of PRSF and TNP2K towards these decisions. There were time constraints in the original period of implementation, and a need to begin work urgently. The sheer scope of TNP2K's activities is important to note – and sheer number of policy recommendations or reports produced in the life of the programme. Furthermore, the nature of the programme kept changing (for example they began with 30 people but had closer to 250 people by mid-2014). The speed and pace of the think tank left little time for a reporting culture.

PRSF found it hard to access the decisions being made about the work, and to gauge the factors surrounding the uptake of recommendations/research produced by TNP2K. TNP2K had little or no incentives to keep PRSF informed under the programme design. Effectively PRSF found itself report to one government and support another, with very little authority or influence on either one. Connected to these sensitivities, were the political constraints – requests and decisions came from the Vice-President and DFAT did not want to be seen to be evaluating a foreign government's activities, only their programme.

A.IV.iii Conclusion for this PRSF case study

There are several conclusions in this case study which are important to note about PRSF and TNP2K's approach to evaluating impact. The first, relies on an acknowledgement is that more reporting was done on this programme than almost any other DFAT programme, as perceived by key informants at interview. The significant funding given to TNP2K to produce research relevant to policy was a huge investment by the Australian Government that we are unlikely to see again so it is unhelpful for a report to simply recommend 'more investment'. Instead this paper suggests an alternative balance in future programming (with a focus that includes evaluating *how* things worked) and appropriate allocation of the funding to reflect this. The research that was produced by TNP2K was considered high quality (with world leaders like the Abdul Latif Jameel Poverty Action Lab (JPAL) and Oxford Policy Management (OPM)). However the research had limited focus on the programme's own performance, and rather was targeting how to improve Indonesian Government programming. In future, there should be very clear delineation between what was 'funding for research' (such as appraisals and randomised control trials of social protection programmes in Indonesia), and what funding was to be spent on evaluating the programme's own impact (actually very little in PRSF and TNP2K's case).

The second conclusion is that PRSF at some point detracted from the design's intention to focus on uptake and influence. The reporting and focus of sense-making on this important aspect of the programming was lost. This was pointed out in one of the reviews during

implementation, but was never really addressed and brought back on track.

Thirdly, TNP2K and PRSF had very strong focus on *how much* impact had been generated by the programme, particularly towards the end of its mandate in 2015 (for example this was when PRSF produced the informative value for money exercise). However, they could follow through on their own impact more, and measure it – for example, they could follow their recommendations (to programmes such as RASKIN and BSM) into the line ministry reporting and seek information on what their recommendations achieved in practice once applied. Furthermore, their focus on *how much* should not distract from the need to set up a system to capture *how* activities worked, and the focus on where to look for their impact (within the complex objective of country systems building).

Fourthly, the informal M&E systems played a very important role that should not be forgotten. Having the Social Protection Advisor from DFAT Indonesia based in-house in TNP2K for two days per week led to real changes that vastly improved the programme's capacity to evaluate impact (not least the quality assurance system that was established in 2012). These remained largely undocumented and it was only through interview that such mechanisms were illuminated.

Fifthly, PRSF's focus was largely on monitoring rather than an evaluative role. This was in large part due to structural incentives for reporting on the programme and relationships. PRSF found it difficult to gauge the uptake of recommendations made by TNP2K to Indonesian government, and TNP2K had few incentives to report to PRSF. The reporting culture needs to be geared towards evaluative roles, as well as monitoring, and the support mechanism empowered to do so.

Annex B: List of people interviewed for the study

With thanks to Vincent Ashcroft who, when conducting the Independent Completion Review for PRSF, allowed the authors to join many of his interviews, which provided excellent access.

Vanya Abuthan: DFAT A/g Unit Manager - PNPM Unit
Wali Akbar: PRSF Grant Manager
Vivi Alatas: World Bank PREM Team Leader
Martyn Ambury: PRSF Deputy Team Leader – Technical
Vincent Ashcroft: Independent Consultant
Surya Aslim: PRSF Outreach Specialist
Vivi Yulaswati Bappenas: Director for Social Protection and Welfare
Mehnaz Bhaur: PRSF Quality Assurance Manager
Barbara Befani: Secretary General of European Evaluation Society
Pak Boediono: Former Indonesia Vice President (2009-2014)
Julien Colomer: IUCN Monitoring and Learning Officer
Jess Dart: Clear Horizons Managing Director
Bethany Davies: Clear Horizons Senior Consultant
Rick Davies: Independent Monitoring and Evaluation Consultant
Irene Guijt: Research Associate, ODI
Hannah Derwent: DFAT Unit Manager - Women in Leadership Unit
Oetami Dewi: MOSA - Head of Cooperation Unit
Jo Dowling: DFAT Unit Manager - Education Unit
James Gilling: Minister (Development Cooperation) DFAT Jakarta
Ruddy Gobel: TNP2K Head of Communication Unit
Scott Guggenheim: World Bank Lead Social Specialist
Simon Henderson: IOD PARC Director
Caroline Hoy: DFID Monitoring and Evaluation Specialist
Patrick Hermanus: DFAT Senior Program Manager - PNPM Unit
Muhammad Ikhsan: Former Deputy for Vice President's Office
Mohammad Ilyas: TNP2K Head of Unified Database
Michael Joyce: TNP2K Mobile Money Specialist
Thamrin Kasman: Secretary, Basic Education Directorate, MOEC
Megha Kapoor + team: PRSF Knowledge Management Manager
Martin Kurnia: PRSF Procurement Manager
John Leigh: DFAT Counsellor Health Section
Theo van der Loop: TNP2K SME Project Design Specialist
Fiona MacIver: AusAID First Secretary, Social Protection Unit
Oliver Mathieson: GRM International Country Director Indonesia
Muhammad Maulana: TNP2K CDD Specialist
Agus Munawar: TNP2K Cluster 3 Working Groups
Suahasil Nazara: Head of Indonesia Fiscal Policy Agency (former TNP2K Head of Working Groups)
Stewart Norup: Sustineo Monitoring and Evaluation Consultant
Ari Perdana: Former TNP2K Head of Cluster 3
Sri Kusumastuti Rahayu: TNP2K Head of Cluster 1 Task Force
Peter Riddell-Carre: PRSF Deputy Team Leader – Management
Jean-Charles Rouge: PRSF M&E Specialist
Elan Satriawan: TNP2K Head of Working Groups
Cipta Setiawan + team: PRSF Human Resource Director
Carli Shillito: Director Indonesia Health and Education Section DFAT
David Smith: PRSF Operation Manager
Prastuti Soewondo (Becky): TNP2K Health Working Group
Dewi Sudharta: Former DFAT Program Manager - Education Unit
Dewi Susanti: KIAT Guru Manager
Sudarno Sumarto: TNP2K Senior Policy Advisor
Patrick Sweeting: PRSF Team Leader
Abdurahman Syebubakar: RSF Senior Policy and Planning Adviser
Andi Yoga Tama: TNP2K Cluster 2
Matt Wai-Poi: World Bank PREM Team
Bambang Widianto: TNP2K Executive Secretary
Emmy Widayanti, MPd: MOSA – Senior Adviser on Inter-institution Relations
Imma Yuliajati: PRSF Gender Adviser

References

Ashcroft, V. (2015) *Independent Completion Report*. DFAT. http://dfat.gov.au/about-us/publications/Documents/indonesia-poverty-reduction-support-facility-icr-2015-man-resp.pdf

Bah et al. (2014) 'An Evaluation of the Use of the Unified Database for Social Protection Programs by Local Governments in Indonesia'. *TNP2K Working Paper 06* (www.tnp2k.go.id/images/uploads/downloads/WP%206%202014%20Evaluation%20of%20the%20local%20uses%20of%20the%20UDB.pdf).

Bennett, A. (2010). 'Process Tracing and Casual Inference' in Brady and Collier (eds) *Reasoning with Causes in the Social Science*. Pittsburgh (http://philsciarchive.pitt.edu/8872/1/Bennett_Chapter_in_Brady_and_Collier_Second_Edition.pdf).

Befani, B. (2012) *Models of Causality and Causal Inference*. Department for International Development (http://betterevaluation.org/resources/guide/causality_and_causal_inference).

Befani, B. (2016) *Appropriate Methods for Impact Evaluation: Criteria and Standards for Choice*, London: Bond (forthcoming).

Belcher, B., Young, J., and Suryadarma, D. (2016 forthcoming) 'Assessing the Contribution of Research to Improved Policy and Practice: An Evaluation of CIFOR's Climate Change Research', in Palenberg, M. and Paulson, A. (eds) *Evaluation and the pursuit of impact*.

Britt, H. (2013) Complexity-Aware Monitoring. USAID Discussion Note, Monitoring & Evaluation Series. USAID.

Brown, T. et al. (2012) *Providing the environment for evidence-based policy making in Indonesia*. AusAID (http://dfat.gov.au/aid/how-we-measure-performance/ode/Documents/case-study-tnp2k-fa.pdf).

Buffardi, A. L. and Hearn, S. (2015) *Multi-Project Programmes: Functions, Forms and Implications for Evaluation and Learning*. Methods Lab. London: Overseas Development Institute.

Collier, D. (2011) 'Understanding Process Tracing', *Political Science and Politics* 44(4): 823-830 (http://polisci.berkeley.edu/sites/default/files/people/u3827/Understanding%20Process%20Tracing.pdf).

Dart, J. and Roberts, M. (2014) Collaborative Outcomes Reporting (http://betterevaluation.org/plan/approach/cort).

Dart, J. (2013) 52 weeks of BetterEvaluation: Week 37: Collaborative Outcomes Reporting (http://betterevaluation.org/blog/collaborative_outcomes_reporting).

Davidson, E. J. (2014). Evaluative Reasoning, Methodological Briefs: Impact Evaluation 4, UNICEF Office of Research, Florence.

Davies, R. (2012) *Where there is no single Theory of Change: The uses of Decision Tree models* (http://mande.co.uk/blog/wp-content/uploads/2012/11/Decision-Trees-and-ToCs-Vs-20121227-NPW1-1.pdf).

Dawson, S. (2009) Draft discussion paper: *Design, Monitoring and Evaluation of Facilities*. Available from the author upon request.

DFAT Strategic Framework 2015-2019. http://dfat.gov.au/about-us/department/Pages/strategic-framework-2015-2019.aspx

Gerring, J. (2007). *Case Study Research: Principles and Practice*. Cambridge University Press.

Gillies, J. and Alvarado, F. (2012) Country Systems Strengthening: Beyond Human And Organizational Capacity Development. Background Paper for the USAID Experience Summit on Strengthening Country Systems. USAID.

Global Environment Facility (2015) *Impact Evaluation of GEF support to protected areas and protected area systems*, 49th GEF Council Meeting, Washington D.C., GEF/ME/C.49/Inf.02 (www.thegef.org/gef/sites/thegef.org/files/documents/EN_GEF.ME_C.49.inf_02_Biodiversity_Impact_Eval_Report_2015.pdf).

Hearn, S. and Buffardi, A. (2016) *What is Impact?* Methods Lab. London: Overseas Development Institute.

Homes, R., Febriany, V., Yumna, A. and Syukri, M. (2011) 'The role of social protection in tackling food insecurity and under-nutrition in Indonesia'. ODI report. London: Overseas Development Institute http://www.odi.org/publications/6184-social-protection-food-insecurity-undernutrition-indonesia

Hummelbrunner, R. and Jones, H. (2013) *A guide to managing in the face of complexity*. ODI Working Paper. London: Overseas Development Institute.

Jones, H. and Hearn, S. (2009) *Outcome Mapping: a realistic alternative for planning, monitoring and evaluation*. ODI Background Note. London: Overseas Development Institute.

Mayne, J. (2008) Contribution Analysis: An approach to exploring cause and effect, *ILAC methodological brief*.

OECD (2010) 'Country Systems, and Why We Need to Use Them', *Development Co-operation Report*, OECD Publishing (www.oecd.org/dac/stats/developmentco-operationreport2010.htm).

OECD (2002) *Glossary of Key Terms in Evaluation and Results Based Management*. Paris: OECD (www.oecd.org/dac/2754804.pdf).

DFAT (2010) Poverty Reduction Support Facility Design Document (http://dfat.gov.au/about-us/publications/Pages/poverty-reduction-support-facility-design-document.aspx).

Poverty Reduction Support Facility (PRSF): Implementation Planning Product 1: PRSF to end 2014 Final Report (2013). Inception Design Team. http://dfat.gov.au/about-us/grants-tenders-funding/tenders/business-notifications/Documents/prsf-to-end-2014.pdf

PRSF Quarterly progress report to DFAT Q2 (2012). Available upon request from authors.

PRSF Quarterly progress report to DFAT Q4 (2013). Available upon request from authors.

PRSF Monitoring and Evaluation Plan (2012). Available upon request from authors.

Quinn Patton, M., McKegg, K., Wehipeihana, N. (2015) *Developmental Evaluation Exemplars: Principles in Practice*. New York: Guilford Press.

Regulation of the President of the Republic of Indonesia Number 5 of 2010 Regarding the National Medium Term Development Plan 2010-2014 (2010). Ministry of National Development Planning/ National Development Planning Agency. http://theredddesk.org/sites/default/files/rpjmn-2010-2014-english-version__20100521111052__2608__0.pdf

Rogers, P. (2008) Using program theories to evaluate complicated and complex aspects of interventions. *Evaluation* 14(1): 29-48.

Rogers, P. (2014) Overview of Impact Evaluation, Methodological Briefs: Impact Evaluation 1,

Florence: UNICEF Office of Research (http://devinfolive.info/impact_evaluation/ie/img/downloads/Overview_ENG.pdf).

Scriven, M. (2008) A summative evaluation of RCT methodology and an alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, 5 (9): 11-24 (http://admn502a2010a01.pbworks.com/f/2008_scriven_on_causation_and_RCTs.pdf).

Snowden, D. J. and Boone, M. E. (2007) A Leader's Framework for Decision Making. *Harvard Business Review*, November 2007. Harvard Business School.

Stame, N. (2010) What Doesn't Work? Three Failures, Many Answers. *Evaluation* 16(4): 371-387.

Stern, E. et al. (2012) Broadening the range of designs and methods for impact evaluations. *Report of a study commissioned by the Department for International Development*. DFID Working Paper 38 (www.gov.uk/government/uploads/system/uploads/attachment_data/file/67427/design-method-impact-eval.pdf).

Williams, B., Hummelbrunner, R. (2010) Systems Concepts In Action: A Practitioner's Toolkit. Stanford University Press. ISBN: 9780804770637.

Wilson-Grau, R. and Britt, H. (2012) *Outcome Harvesting*. Ford Foundation (www.outcomemapping.ca/download/wilsongrau_en_Outome%20Harvesting%20Brief_revised%20Nov%202013.pdf).

White, H. and Phillips, D. (2012). Addressing the attribution of cause and effect in small n impact evaluations: towards an integrated framework. *International Initiative for Impact Evaluation*. Working Paper 15 (www.3ieimpact.org/media/filer_public/2012/06/29/working_paper_15.pdf).

Young, J. and Court, J. (2004) *Bridging Research and Policy in International Development*. ODI Briefing Paper. London: Overseas Development Institute (www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/198.pdf).

Young, J. and Bird, N. (2015) Informing REDD+ Policy: An assessment of CIFOR's Global Comparative Study. London: Overseas Development Institute.

Websites accessed:

BetterEvaluation (Accessed Nov 2015) http://betterevaluation.org/

Climate & Development Knowledge Network (Accessed Nov 2015) http://cdkn.org/

Eval3C Website (Accessed Nov 2015) http://evalc3.net/

MAMPU Website (Accessed Nov 2015) http://www.mampu.or.id/en

Poverty Action Lab (Accessed Nov 2015) http://www.povertyactionlab.org/

ODI is the UK's leading independent think tank on international development and humanitarian issues.

ISSN: 2052-7209
ISBN: 978-0-9941522-1-3

All ODI Reports are available from www.odi.org

Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
Tel +44 (0)20 7922 0300
Fax +44 (0)20 7922 0399

www.odi.org