Neil Buddy Shah
Paul Wang
Andrew Fraker
Daniel Gastfriend

# Evaluations with impact

## Decision-focused impact evaluation as a practical policymaking tool

September 2015

Working Paper 25

3ie International Initiative for Impact Evaluation

## About 3ie

The International Initiative for Impact Evaluation (3ie) is an international grant-making NGO promoting evidence-informed development policies and programmes. We are the global leader in funding and in producing high-quality evidence of what works, how, why and at what cost. We believe that better and policy-relevant evidence will make development more effective and improve people's lives.

## 3ie working papers

These papers focus on current issues, debates and enduring challenges facing development policymakers and practitioners and the impact evaluation and systematic reviews communities. Policy-relevant papers draw on findings from impact evaluations and systematic reviews funded by 3ie, as well as findings from other credible and rigorous evaluations and reviews, to offer insights, new analyses, findings and recommendations. Papers focusing on methods also draw on similar sources to help advance understanding, design and use of rigorous and appropriate evaluations and reviews.

## About this working paper

This paper was commissioned by the William and Flora Hewlett Foundation to provide a perspective on the future of impact evaluation, based on IDinsight's work using rigorous impact evaluation as a practical decision-making tool for policymakers and non-governmental organisations in developing countries. The goal of this paper is to articulate how impact evaluations can achieve their full potential to improve social programmes in the developing world. All of the content is the sole responsibility of the authors and does not represent the opinions of 3ie, its donors or its Board of Commissioners. Any errors and omissions are also the sole responsibility of the authors. Any comments or queries should be directed to the corresponding author, Neil Buddy Shah, Neil.buddy.shah@idinsight.org.

Suggested citation: Shah, NB, Wang, P, Fraker, A and Gastfriend, D, 2015. *Evaluations with impact: decision-focused impact evaluation as a practical policymaking tool. 3ie Working Paper 25.* New Delhi: International Initiative for Impact Evaluation (3ie).

3ie executive editors: Emmanuel Jimenez and Beryl Leach
3ie managing editor: Annette N Brown
Production manager: Brigid Monaghan
Copy editor: Jaime L Jarvis
Proof reader: Yvette Charboneau
Cover design: John McGill
Printer: Via Interactive
Cover photo: Albert González Farran/UNAMID

# Evaluations with impact: decision-focused impact evaluation as a practical policymaking tool

Neil Buddy Shah
IDinsight

Paul Wang
IDinsight

Andrew Fraker
IDinsight

Daniel Gastfriend
IDinsight

**3ie Working Paper 25**
**September 2015**

## Acknowledgements

The authors would like to thank the William and Flora Hewlett Foundation for funding this paper, and in particular, Kristen Stelljes and Sarah Lucas from the Hewlett Foundation for their useful comments throughout the process of writing the paper. We would like to thank Annette Brown, Emmanuel Jimenez, Jo Puri, and anonymous reviewers from the International Initiative for Impact Evaluation (3ie) for their comments. We would like to acknowledge all of our colleagues at IDinsight, especially Ronald Abraham, Heather Lanthorn, Jeffery McManus and Esther Wang, whose perspectives and experiences were integral in writing this paper. Finally, we are grateful for the insightful thoughts from the numerous policymakers, donors, and NGO leaders who were generous in giving us their time to be interviewed for this paper.

## Executive summary

Impact evaluations have enhanced international development discourse and thinking over the past 15 years, but the full promise of impact evaluations to improve lives has yet to be realised.

Rigorous impact evaluations to date have largely focused on contributing to a global learning agenda. These 'knowledge-focused evaluations' (KFEs) – i.e. those primarily designed to build global knowledge about development interventions and theory – have catalysed a more sophisticated dialogue around results and refined important development theories. Evidence created by KFEs has also led to international scale-up of several interventions.

However, two challenges have limited the extent to which KFEs have informed policy and programmatic decisions. First, evaluator incentives are often misaligned with implementer needs. Second, many impact evaluation results do not generalise across contexts.

To more effectively inform development action, impact evaluations must be adapted to serve as context-specific tools for decision making that feed into local solution-finding systems. Towards this end, a new kind of impact evaluation has recently emerged, one that prioritises the implementer's decision-making needs over potential contributions to global knowledge. These 'decision-focused evaluations' (DFEs) are driven by implementer demand, tailored to implementer needs and constraints, and embedded within implementer structures. Importantly, DFEs mitigate generalizability limitations by testing interventions under conditions very similar to those in which the interventions could be scaled. By reframing the primary evaluation objective, they allow implementers to generate and use rigorous evidence more quickly, more affordably and more effectively than ever before.

Acknowledging that the distinction between KFEs and DFEs is not binary – any evaluation will exhibit varying KFE and DFE characteristics – we have developed these terms because they help elucidate the objectives of a given evaluation and offer a useful conceptual frame to represent two axes of rigorous impact evaluation. We argue that the future of impact evaluation should see continued use of KFEs, significantly expanded use of DFEs and a clear strategy on when to use each type of evaluation. Where the primary need is for rigorous evidence to directly inform a particular development programme or policy, DFEs will usually be the more appropriate tool. KFEs, in turn, should be employed when the primary objective is to advance development theory or in instances when we expect high external validity, *ex ante*. This recalibration will require expanding the use of DFEs and greater targeting in the use of KFEs.

To promote the use of rigorous evidence to inform at-scale action, we identify two strategies to increase demand and four strategies to increase supply of DFEs. To stimulate demand for DFEs:

- Funders should establish ‘impact first’ incentive structures that tie scale-up funding to demonstration of impact over a long time horizon; and
- Funders should allocate funds to support DFEs across their portfolios.

To build supply for DFEs:

- Universities and funders should build and strengthen professional tertiary education programmes to train non-academic impact evaluation specialists;
- Funders should subsidise start-up funds to seed decision-focused impact evaluation providers, and international evaluation organisations should support these organisations with rapid external quality reviews;
- Evaluation registries should publish the cost and length of impact evaluations and the actions they influence; and
- Funders, evaluators and implementers should collaborate to establish ‘build, operate, transfer’ evaluation cells to strengthen evaluation capacity in implementing organisations.

Overall, to maximise the social impact of impact evaluations, all involved stakeholders (implementer, funder and evaluator) should have clarity on each evaluation’s primary objective and select the appropriate evaluation type (DFE or KFE) accordingly. We subsequently envision DFEs supporting a robust innovation ‘churn’ whereby intervention variations are generated and rigorously assessed in rapid cycles. Both KFEs and DFEs – along with enhanced monitoring systems, big data and other emerging measurement developments – will play a critical role in tightening the link between evidence and action.

# Contents

# List of tables and figures

## Tables



## Figures

## Abbreviations and acronyms

| | |
|---|---|
| 3ie | International Initiative for Impact Evaluation |
| AusAID | Australian Agency for International Development |
| BSPHCL | Bihar State Power Holding Company Ltd. |
| CEC | Concurrent evaluation cell |
| DFE | Decision-focused evaluation |
| DfID | Department for International Development |
| IEG | Independent Evaluation Group |
| IPA | Innovations for Poverty Action |
| J-PAL | Abdul Latif Jameel Poverty Action Lab |
| KFE | Knowledge-focused evaluation |
| M&E | Monitoring and evaluation |
| NGO | Non-governmental organisation |
| RCT | Randomised controlled trial |
| USAID | United States Agency for International Development |

# 1. Introduction

This paper was commissioned by the William and Flora Hewlett Foundation to provide a perspective on the future of impact evaluation, based on IDinsight's work using rigorous impact evaluation as a practical decision-making tool for policymakers and non-governmental organisations (NGOs) in developing countries. The goal of this paper is to articulate our views on the strengths and shortcomings in the impact evaluation status quo[1] and to explore how impact evaluations can achieve their full potential to improve social programmes in the developing world.

We argue that the status quo of impact evaluation has been dominated by a global learning agenda focused on building a body of knowledge around intervention-outcome combinations and advancing development theory.[2] This focus has led to an emphasis on what we call 'knowledge-focused evaluations' (KFEs) – evaluations primarily designed to build global knowledge about development interventions and theory – that have catalysed a more sophisticated dialogue around results and have refined important development theories. Evidence created by KFEs has also led to international scale-up of several interventions.

However, two challenges have limited the extent to which KFEs have informed policy and programmatic decisions. First, evaluator incentives are often misaligned with implementer needs. Second, many impact evaluation results do not generalise across contexts.

To more effectively inform development action, we argue that rigorous impact evaluation should increasingly serve as a context-specific decision-making tool that feeds into local solution-finding systems. To this end, we argue that the future of impact evaluation should significantly expand the use of what we term 'decision-focused evaluations' (DFEs). The fundamental goal of a DFE is to inform a *specific* policy/programmatic decision of a *specific* implementer in a *specific* geography for a *specific* target population over a *specific* time horizon. DFEs are demanded by implementers and aim to generate the most rigorous evidence possible to inform a programmatic decision, within the time, budgetary and operational constraints of the implementer. Importantly, DFEs mitigate generaliszability limitations by testing interventions under conditions very similar to those in which the interventions could be scaled. By reframing the primary evaluation objective, we argue that DFEs have

[1] For this paper, we define impact evaluation as quantitative or mixed-methods studies that estimate the causal impact of an intervention on an outcome of interest by comparing outcomes of beneficiaries against outcomes of an estimated counterfactual. (The counterfactual refers to the hypothetical outcomes those beneficiaries would have exhibited had they not received the intervention.) We assume the reader is familiar with experimental and quasi-experimental impact evaluation methodologies, including randomised evaluations, statistical matching, difference-in-differences analysis and regression discontinuity.

[2] By 'development theory', we mean the various theoretical disciplines that inform the international community's understanding of how development occurs, under what circumstances and why. These include theories of economics, behaviour and political economy.

the potential to enable implementers to generate and use rigorous evidence more quickly, more affordably and more effectively than ever before.

Although we draw a distinction between KFEs and DFEs throughout this paper, by no means do we intend to create a false dichotomy between the two; any given impact evaluation will exhibit KFE and DFE characteristics. We have created this classification as a conceptual framework to help clarify an evaluation's primary objective. This clarity will help determine the evaluation's optimal structure (in terms of methodology, cost, timeline and other characteristics) and criteria to judge whether the evaluation achieved its goals.

The rest of this paper is organised as follows. Section 2 describes the scope of the paper and our methods. Section 3 is an analysis the current state of impact evaluation, focusing on the KFE theory of change and the successes and limitations of that type of evaluation. Section 4 defines DFEs, discusses how they address some of the shortcomings in the impact evaluation status quo and highlights some limitations of DFEs. Sections 5 and 6 conclude with a discussion of how KFEs and DFEs should be used in the future, pointing to wider changes in the development ecosystem that will be needed to optimally use KFEs and DFEs.

## 2. Scope and methods

This paper is a concept piece based primarily on the collective experiences of our colleagues at IDinsight (including their previous roles at other organisations). It is further informed by a review of the impact evaluation literature, assessing publications from leading impact evaluation organisations including the International Initiative for Impact Evaluation (3ie), the Abdul Latif Jameel Poverty Action Lab (J-PAL), Innovations for Poverty Action (IPA) and the World Bank's Development Impact Evaluation initiative. This review explores relevant research but does not attempt a comprehensive assessment of the existing literature.

These arguments are also informed by 27 interviews, conducted by phone and in-person with policymakers, practitioners, funders and researchers with varying levels of familiarity with impact evaluation. The respondents were purposely chosen from our networks to elicit a variety of viewpoints, so they should not be considered a representative subsample of impact evaluation stakeholders. We conducted the interviews to explore different perspectives on the current state of impact evaluation, solicit ideas on the future of impact evaluations and collect feedback on the hypotheses and arguments we advance in this paper. The interviews followed open-ended, semi-structured questionnaires tailored to the respondents' different roles. Given the limitations of the semi-structured interview, we use these interviews primarily to provide more colour to our conceptual arguments throughout the paper rather than as the primary source of evidence for our claims. We categorise interview responses thematically, as described in appendix A, and present the list of respondents in appendix B.

We have limited the scope of this paper in several ways. It is not a thorough review of the history, uses and impact of impact evaluation in international development. It touches only briefly on other measurement and learning tools besides rigorous impact evaluations (e.g. monitoring and process evaluations) that can help maximise social impact. It also does not comprehensively cover all possible impact evaluation purposes, such as to meet end of project reporting requirements for funders or to validate an organisation's model to help with fundraising.

# 3. The status quo: 'knowledge-focused' evaluation

## 3.1 Knowledge-focused evaluations: theory of change and successes

The rigorous impact evaluation status quo has been driven by a global learning agenda that seeks to build a global body of evidence of 'what works in development'.[3] This agenda assumes it is possible (at least to a certain extent) to predict outcomes across contexts based on impact evaluations of the same intervention-outcome combination. It also seeks to use impact evaluation to improve our understanding of human behaviour in developing country contexts. As a result of this focus, most rigorous impact evaluations to date have been KFEs, which are often conducted by university-based researchers or multilateral institutions, with results published in academic journals. Evaluation topics are motivated by gaps in the academic literature, and evaluators are typically external to the implementing organisation.

Figure 1 presents a stylised theory of change for how KFEs achieve social impact across four channels.

The first channel posits that, for a given KFE, the implementer managing the intervention under evaluation will scale, revise or discontinue the programme based on the evaluation results. Because the evaluation results clarify which approach is socially optimal, this decision will typically lead to improved social outcomes relative to the counterfactual action that would have occurred in the absence of the evaluation.

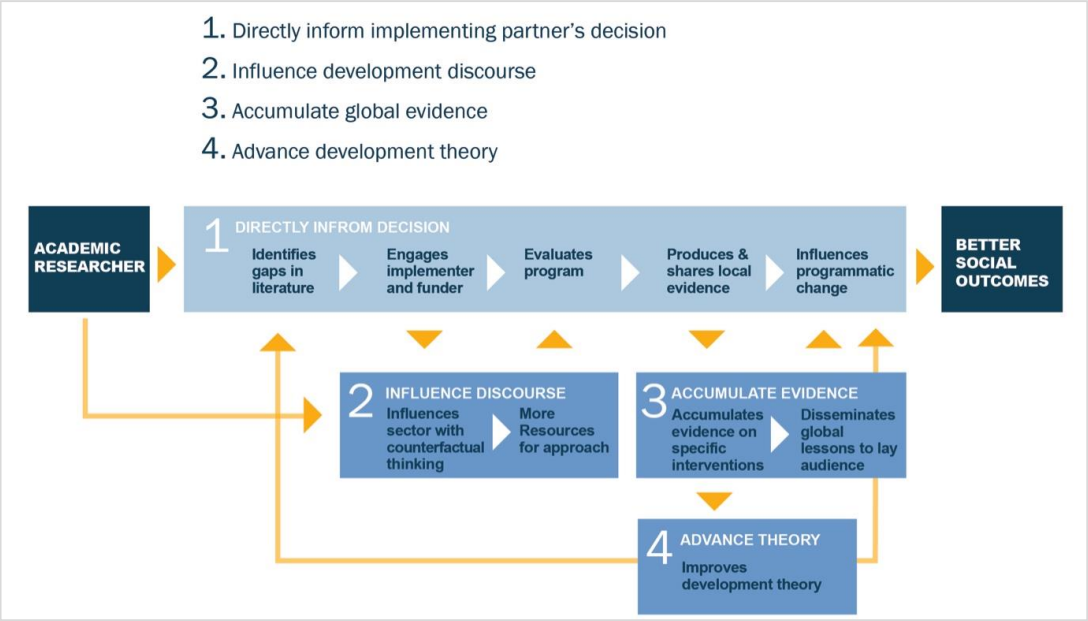
The second channel hypothesises that an emphasis on KFEs will move development practitioners to talk and think about impact measurement more critically. This shift will mobilise resources for further rigorous evaluations, which will influence action via the other channels described and generally lead to higher-quality systems for monitoring and evaluation (M&E).

[3] See, for example, the World Bank's Knowledge in Development Note on Impact Evaluation: 'Impact evaluations are about understanding what works in development, under what circumstances, and why … Impact evaluations are essential for creating a "knowledge bank" … For example, are learning outcomes best improved through reducing poverty, building schools and other education-based policies, improving early childhood development through health interventions, or building rural roads?' (World Bank n.d.)

The third channel holds that researchers will aggregate results from KFEs into a global body of evidence on which models of intervention 'work' (i.e. successfully influence targeted social outcomes in the desired direction), which do not work and under what conditions they are most likely to succeed or fail. This information is made available to implementers via published studies, systematic reviews, policy briefs and expert opinion and is meant to inform their decisions on which types of programmes to pursue and how to structure them.

The fourth channel contends that researchers will use the body of evidence to answer theoretical questions relating to development, advancing the academic community's understanding of the dynamics underlying poverty. Researchers will communicate these insights directly to practitioners, influencing their thinking and action, and feed these insights back into the KFE generation process.

**Figure 1: KFE theory of change**




Note: Figure prepared by IDinsight.

KFEs have led to notable accomplishments along all four impact channels; we highlight successes in the following sub-sections. However, our analysis indicates that KFEs are best suited to influence development discourse (channel 2) and advance development theory (channel 4). In section 3.6 we discuss how flawed assumptions in channels 1 and 3 limit KFEs' impact through these mechanisms.

## 3.2 Channel 1: directly informing decisions

KFEs have directly informed several large-scale policy decisions and scale-ups of interventions, such as mass school-based deworming to improve school attendance (box 1), chlorine dispensers to reduce diarrhoeal disease and conditional cash transfers such as Progresa in Mexico (though some argue that Progresa, as box 2 explores, is not a straightforward triumph of evidence-informed decision making). In

some cases, impact evaluations may have also helped undercut support for less cost-effective interventions, although as Dean Karlan, it is difficult to know how many donor dollars were not invested as a result of null or negative results (IPA 2011).

**Box 1: Deworming the world: KFE plays a key role in catalysing policy change**

From 1998 to 2001, the NGO International Child Support delivered deworming drugs to some 30,000 students in 75 government primary schools in Kenya's Busia District to treat intestinal worm infections. The programme's phased-in design resulted in some schools randomly receiving the drugs years before other schools, enabling Miguel and Kremer (2004) to quantify the causal impact of deworming on health and education outcomes. Research on this programme found meaningful effects on self-reported health, school attendance and long-run labour market outcomes, demonstrating that deworming can be an extremely cost-effective health and education intervention (Baird *et al.* 2013). J-PAL estimates that deworming costs only US$7.19 per additional year of schooling gained and US$4.55 per disability-adjusted life year averted (J-PAL 2012).

After these findings were published, several international deworming initiatives were launched, most prominently Deworm the World, which formed after J-PAL researchers presented at the World Economic Forum in 2007 (J-PAL 2012). Deworm the World provides technical assistance to governments to implement mass school-based deworming programmes, which treated 37 million children in Kenya and India in the 2013–2014 school year, with plans to expand to more countries in the coming years (Evidence Action n.d.).

The Kremer and Miguel randomised evaluation likely played a key role in international take-up by policymakers (Ashraf *et al.* 2011). This example, in turn, demonstrates the contributions KFEs can make when they successfully influence international action.[4]

Several factors may have helped the study influence large-scale action:

- Low political controversy: deworming programmes do not challenge the status quo in government schools (unlike reforms such as teacher payment, increased classroom monitoring and curriculum changes), so there are few potential 'losers' who would resist implementation;
- Straightforward implementation: deworming programmes can be applied on top of traditional education services; and
- Sustained advocacy for scale-up: the researchers involved in the original evaluation did not conclude their involvement with a journal article but actively campaigned in multiple forums for development practitioners to expand deworming programmes.

[4] There is still controversy regarding the findings of this evaluation, its external validity and the broader impacts of deworming. We do not discuss the debate here, but note that it touches on some of the external validity arguments posed later in the paper.

**Box 2: The story behind Progresa: evidence as a political tool**

The Progresa conditional cash transfer programme in Mexico, later renamed Oportunidades and Prospera, is often hailed as one of the great successes of impact evaluations influencing policy. Launched by the Mexican government in 1997, Progresa disburses cash transfers to poor households that meet specified requirements, including having children attend school and receive medical check-ups. The programme was designed to include an impact evaluation from the outset (Gertler *et al.* 2011).

The evaluation found that the programme increased school enrolment and reduced disease morbidity among children in participating families. One analysis estimated that Progresa would increase schooling by 0.7 years for the average child participating between the ages of 6 and 14 (Behrman *et al.* 2005). Another analysis found that the programme reduced the risk of illness in the first six months of life by 25 per cent (Gertler 2004). Progresa was lauded internationally and expanded in subsequent years, even after the political party that had instituted the programme left power (Gertler *et al.* 2011).

Evaluation advocates have pointed to the programme as a leading example of evidence-based policy. The Progresa evaluation has been credited with driving scale-up within Mexico, promoting conditional cash transfers internationally and leading the Mexican government to use evaluations more frequently as a policy-making tool (Székely 2011; Gertler *et al.* 2011).

However, the evaluation also demonstrates how politics drive the use of evidence. In Mexico, new government administrations had typically dismantled and replaced existing social programmes soon after taking office (Székely 2011; Gertler *et al.* 2011). Given this environment, some argue that the government commissioned the evaluation as political cover to extend the lifespan of a programme it had already decided to implement, rather than as a scientific exercise to determine whether Progresa warranted implementation at scale. Others go further, arguing that the conditionality imposed on the cash transfers was included to protect the transfers from interference, and that the impact of those conditions on school enrolment was a political afterthought. Lant Pritchett (2012) writes,

> In other words, one common narrative – that the scaling up of CCTs [conditional cash transfers] is a good example of evidence based policymaking because the use of randomisation in the design of Progresa provided solid evidence that it was an effective program and hence other countries adopted a CCT because of this solid evidence – has it almost exactly backwards. The impact evaluation proved that Progresa was cost ineffective if it was considered as a mechanism to increase schooling. Everyone involved in the design knew this. They were not imposing the conditionality to get the behaviour conditioned upon, but to get the transfer itself …. Ironically, what was really learned from the experience of Progresa is that having a rigorous experiment attached to your program can be great politics … This increases your ability to resist partisan political meddling in design and implementation – even if you don't learn anything particularly special from the experiment.

The Progresa evaluation made waves in the international community and served as a model for future policy evaluation efforts, but the story also underscores how political interests can influence the use of impact evaluation results. Understanding the incentives for conducting an evaluation may help identify how and when evaluation results are likely to influence policy.

### 3.3 Channel 2: influencing development discourse

> 'Imbuing a culture of constantly questioning and testing whether something works … [and] bringing the social scientists' approach of curiosity and critical thinking into policymaking is a giant step, regardless of what the specific lessons from individual evaluations are'.
>
> *—Marc Shotland, director of research and training at J-PAL (interview with IDinsight)*

The growing prominence of impact evaluation has catalysed a more sophisticated dialogue around impact and the importance of rigorous measurement to quantify development results. KFEs contributed to a shift away from an input/output paradigm, which tracks resources used and deliverables produced to judge success, to a paradigm focused on outcomes and causal attribution. This shift continues, and remains one of the crucial contributions of KFEs to global development practice (Levine *et al.* 2015). Several development practitioners interviewed for this paper remarked that implementers are becoming increasingly conversant in evaluation concepts; other respondents highlighted increasingly serious approaches to impact measurement among pioneering funders such as the UK Department for International Development (DfID), the Development Impact Ventures initiative of the United States Agency for International Development (USAID) and the Global Innovation Fund.

### 3.4 Channels 3 and 4: accumulating global evidence and advancing development theory

KFEs have improved our understanding of human behaviour and development theory (channel 4) by building a global evidence base (channel 3). In behavioural economics, for example, a series of impact evaluations on commitment devices have driven insights into the role of cognitive capacity constraints and time-inconsistent preferences in driving microeconomic behaviour (World Bank 2015). Books such as Abhijit Banerjee and Esther Duflo's *Poor Economics* (2012) have synthesised evidence from many impact evaluations to enhance the development community's understanding of incentives in public service delivery, appropriate pricing of public health goods, and savings behaviour.

Clusters of papers around specific interventions have also provided valuable learning for the development community. For instance, seven randomised evaluations from around the world have provided strong evidence that microcredit does not reliably enable the poor to raise their incomes (J-PAL and IPA 2015).[5] In an interview for this paper, Bill Savedoff, senior fellow at the Center for Global Development, argued that contributing to the global knowledge base and development theory may be impact evaluation's most valuable function.

[5] However, the extent to which this knowledge accumulation has affected global flows of subsidised capital to microcredit is less clear.

## 3.5 Limitations of knowledge-focused evaluations

Impact evaluations have the potential to influence action to a greater extent than KFEs have yet done. More than 2,500 impact evaluations on development programmes have been published since 2000 (3ie n.d.), but we are aware of relatively few examples where the evidence generated by these evaluations has catalysed at-scale action by the implementing partner (channel 1).[6] For example, J-PAL has supported 626 evaluations since its inception, but its website lists only 7 intervention models that have been scaled up – and 15 total instances of scale-up – after demonstrating impact in a J-PAL affiliated evaluation (J-PAL n.d.). This is not an exhaustive list of interventions scaled up based on impact evaluations, nor does it capture decisions to discontinue cost-ineffective programmes based on evaluation evidence, but it is still indicative of the relatively low yield of evaluations directly leading to action. See box 3 for an example of a J-PAL KFE that failed to influence policy.

This shortcoming has been acknowledged by researchers such as Dean Karlan, founder of IPA, who recently stated:

> IPA and JPAL have been involved in five to six hundred randomised trials over the past 10 years now … and there's really only a handful – 5 to 10 things that have risen to the top – that have clear evidence, consistent evidence, and a clear theory of change behind them so that we have championed them into scale up (Ford Foundation 2014).

In an earlier interview, Karlan acknowledged, 'We are doing less scale-up [of proven programmes] than was originally intended … [research] doesn't always lead to a packaged "do this" answer' (IPA 2011).

Some of our respondents offered similar sentiments, reporting frustration at the lack of programmatic change that has come from the wave of impact evaluations over the last 15 years.

Below we discuss barriers to KFEs influencing action. We then discuss actions that funders and implementers of KFEs (including 3ie, IPA, J-PAL, the Center for Effective Global Action and the World Bank) are taking to address this shortcoming, before arguing in sections 4 and 5 for more fundamental changes.

[6] There are, of course, notable exceptions such as Pratham, an India-based NGO that has conducted a series of RCTs, in partnership with J-PAL, to inform its programming and strategy. CEO Rukmini Banerji noted in our interview that Pratham has revised its programming based on insights generated from the data collected for their RCTs. She said these studies have been particularly useful due to Pratham and J-PAL's focus on researching shared learning objectives, whereas it can be more difficult to work with researchers who bring an agenda that is not useful for the implementing partner.

**Box 3: Incentives for immunisation: failure of a KFE to influence policy**

From 2004 to 2007, J-PAL coordinated with the Indian NGO Seva Mandir to evaluate a programme aimed at increasing immunisation rates. The study evaluated two variants of the programme against a control group. In the first treatment arm, monthly reliable immunisation camps were set up in villages; in the second, the camps were combined with small awards offered to attending families. The study found that the camps alone increased the proportion of children receiving full immunisation from 6 per cent to 18 per cent, and camps with incentives increased this figure to 39 per cent. By driving volume, the incentives also dramatically increased the camps' efficiency, reducing the cost per fully immunised child from US$56 to US$28 (J-PAL 2011).

GiveWell labelled this evaluation as one of 'the most successful and policy-relevant studies J-PAL has implemented [in] the last ten years' (2014). However, despite the extremely positive results, to our knowledge there has been no significant scale-up of this programme or others based on it. Although Seva Mandir (n.d.) continues to implement the programme, it operates at a relatively small scale and immunised only 1,324 children in 2014.

J-PAL is planning several external replication studies of similar programmes, including one in Haryana, India, and another in Pakistan, and Evidence Action is exploring ways to support scale-up in Pakistan. Preliminary results from the Haryana study are not expected until the end of 2016, however, and as of December 2014, plans for research and scale-up in Pakistan were still in preliminary stages (GiveWell 2014).

The lack of significant action in the eight years following the study – despite its remarkable findings – offers an example of the challenge of influencing policy with KFEs. Without a direct consumer of the evidence willing and able to implement a study's recommendations at scale, even extremely compelling findings may fail to catalyse change in a reasonable time frame.

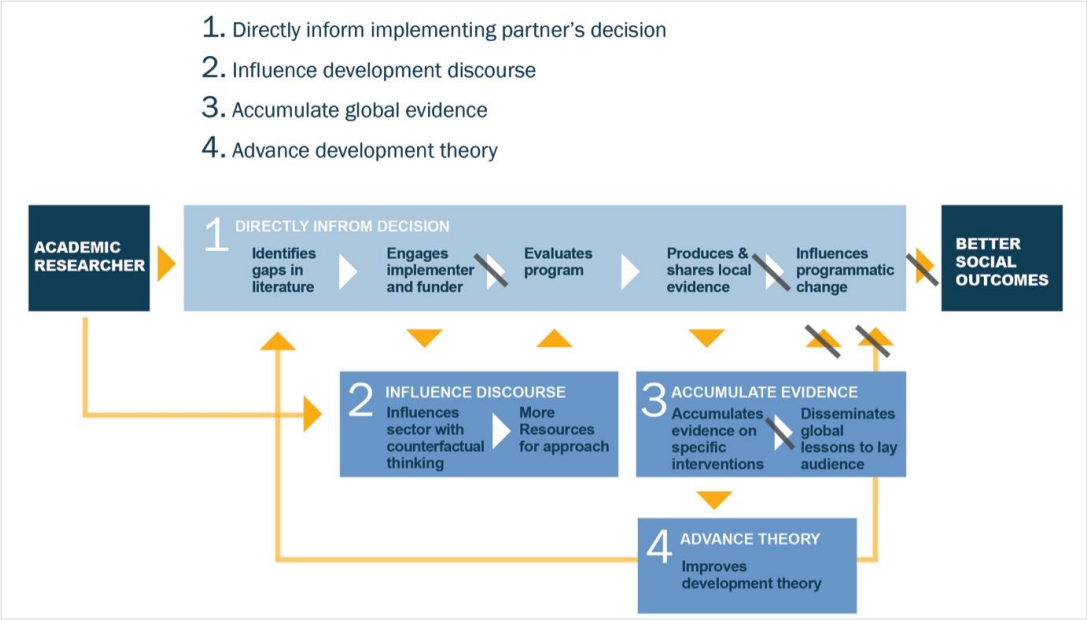
## 3.6 Weak links in the theory of change for knowledge-focused evaluations

Examining the KFE theory of change helps explain limitations in KFEs' ability to directly influence development action. Specifically, weaknesses are most prominent in assumptions underpinning channel 1 (direct impact) and channel 3 (global evidence accumulation). These weaknesses are outlined in table 1 and figure 2 and discussed in more detail below. When combined, these weaknesses have the potential to obstruct the key links between KFEs and programmatic change.

**Table 1: Weak links in the KFE theory of change**

| Assumption | Weakness |
|---|---|
| *Channel 1: direct impact* | |
| The questions researchers seek to answer align with the questions implementers seek to answer. | 1. Differing evaluator and implementer priorities may lead to research questions that are not directly relevant to implementers' decisions. |
| The typical structure of researcher-driven evaluations aligns with the decision-making needs and constraints of implementers. | 2. Academic-driven evaluations may take a long time to produce results and may miss an implementer's decision-making window.<br>3. Impact evaluations are often not conducted because they are too expensive, lengthy, or operationally demanding for the funder or implementer. |
| *Channel 3: global evidence accumulation* | |
| Replicating studies in many contexts can provide clear guidance on whether certain interventions 'work'. | 4. Results often do not generalise across contexts; meta-analyses have poor predictive power on whether an intervention will work in a specific context (Vivalt 2015). |
| Implementing organisations can readily access global evidence generated by KFEs. | 5. Results are frequently difficult for implementing organisations to access or interpret. |
| Implementers will act on accumulated evidence from KFEs in other contexts. | 6. Even when results generalise and are accessible, political barriers can prevent external evidence from influencing action. |

**Figure 2: Weak links in the KFE theory of change**




Note: Figure prepared by IDinsight.

### *3.6.1 Weak link 1: differing evaluator and implementer priorities*

> 'The academic question was, "Does [the programme] work?" We know now that it does. But follow-on questions like, "What's the optimal ratio of students to teaching assistants—should it be four to one; eight to one; twelve to one?" That's not a question you are going to get many academics excited about spending a year and a half and a couple hundred thousand dollars figuring out. But if you're a large organisation running remedial education, you really ought to know the answer … we're doing less of that type of research, things that are distinctly non-academic but that are necessary for policy'.
>
> *—Dean Karlan, on an evaluation of a remedial education program (IPA 2011)*

Evaluator interests often do not align with practitioner interests (Dhaliwal *et al.* n.d.). Most KFEs are conducted by academic researchers, whose incentive to conduct evaluations is advancement of their research careers (i.e. publication). Academic publications reward the novelty of the intervention being tested, relevance to development theory and use of innovative methodologies, incentivising researchers to push implementers towards evaluations with these characteristics. In contrast, policymakers and practitioners tend to want evidence on operational topics that may seem mundane to researchers (Dhaliwal *et al.* n.d.; Levine *et al.* 2015). In a 2011 interview, Dean Karlan acknowledged the divergent priorities of researchers and implementers, noting there are 'tests that are not academically interesting but [that] would be hugely useful to [development organisations] for program or product design' (IPA 2011).

These competing priorities can limit the extent to which researcher-driven studies influence action. In an AidData (2015) survey of almost 6,750 developing country policymakers and practitioners, respondents reported that alignment with the domestic leadership's priorities was the main factor determining whether external government performance assessments influenced subsequent reforms. Similarly, a 2009–2011 survey of 985 policy stakeholders across Africa, South Asia and Latin America demonstrated that many policymakers have highly specific questions that are not being answered by academic researchers (through impact evaluations or otherwise). These individuals – among them senior staff in government and non-governmental, multilateral, bilateral and private organisations – consistently demanded high-quality primary data they could use to answer their particular questions. The report consequently recommended that think tanks 'offer more specific and customised analytic services to deliver on the information needs of stakeholders' (Cottle n.d.).

> ‘When an impact evaluation person talks to Ministry of X in country Y, they usually think, well, what kind of paper can I get out of this? … The objective function of academic researchers is to publish, and that is not always going to coincide with asking the kind of questions policymakers want answered … It is not wrong that academics want to answer fundamental questions for theory. But let’s not pretend that the policy relevance is always high on those. Let’s call it what it is’.
>
> *—Markus Goldstein, lead economist at the World Bank (interview with IDinsight)*
>
> ‘It must be acknowledged that the set of research questions that are most relevant to development policy overlap only partially with the set of questions that are seen to be in vogue by the editors of the professional journals at any given time. The dominance of academia in the respected publishing outlets is understandable, but it can sometimes make it harder for researchers doing work more relevant to development practitioners, even when that work meets academic standards. Academic research draws its motivation from academic concerns that overlap imperfectly with the issues that matter to development practitioners’.
>
> *—Martin Ravallion (2009)*

### *3.6.2 Weak links 2 and 3: time frame and cost*

Researchers’ priorities can also lead KFEs to be longer and costlier than implementers need. A number of factors drive researchers to design longer, more expensive evaluations:

- Longer data collection periods enable measurement of downstream, rather than proximate, outcomes (Dhaliwal *et al.* n.d.);
- Larger sample sizes facilitate detection of academically relevant effect sizes, which can be smaller than the ‘policy relevant’ effect sizes;
- Longer and more expensive household survey questionnaires can tease out and test questions that are relevant to development theory but not directly related to an implementing organisation’s decision making; and
- Longer writing and review processes are used to adhere to publication conventions.[7]

These features can make KFEs prohibitively expensive for many implementers. In 2011 the Australian Agency for International Development (AusAID) estimated that the average 3ie-funded impact evaluation cost US$250,000, and in 2012 the Independent Evaluation Group (IEG) reported that the average World Bank impact evaluation costs US$500,000 (IEG 2012; AusAID 2011).

Publication schedules in particular illustrate the contrast between academic and decision-maker timelines. Cameron, Mishra and Brown (2015) assessed a random sample of 113 impact evaluations and found that the average study was published 4.17 years after end line data collection. Working papers were not much faster (3.63

[7] Working papers and presentation of pre-published materials to relevant stakeholders mitigate this problem.

years), and evaluations in social science journals were far slower (6.18 years). In contrast, evaluations commissioned by governments were published only a year after end line data collection, on average.

Such time frames are incongruous with many implementer contexts. Given the pace of decision-making cycles and operational learning, policymakers require more rapid feedback than most KFEs offer. Implementer decisions can rarely wait two or more years for evidence to arrive. During this time, the individual who commissioned the evaluation may have left his or her position, or the programmatic circumstances that prompted the study may have changed (Hallsworth *et al.* 2011). Moreover, as implementers conduct a programme, they regularly discover operational improvements that merit immediate piloting or deployment. Academic evaluations, however, typically require that implementation procedures remain consistent for the duration of the study to maintain a coherent and consistent model (Woolcock 2013).

### *3.6.3 Weak link 4: generalisability*

There is growing evidence that external validity (the extent to which findings from one context generalise to another) for impact evaluations of many development programmes is extremely low (see box 4). This lack of generalisability is perhaps the greatest weakness in the KFE theory of change.[8] KFEs are typically designed to maximise internal validity, with external validity treated as a second-order consideration (Ravallion 2009).

**Box 4: External validity concerns from practitioner interviews**

In our interviews with funders and practitioners, numerous respondents raised concerns regarding the generalisability of impact evaluations. For example, Andrew Youn, executive director and co-founder of One Acre Fund, remarked that although One Acre Fund is often able to use published evaluations as an initial screening for programme design, their utility is limited since they usually assess interventions that were conducted in a different geographic context, under ideal conditions and divorced from a scalable business model. Cormac Quinn, evaluation and results adviser at DfID, noted that there is serious discussion within DfID on the extent to which impact evaluation results are generalisable within the same country. Satyam Vyas, chief operating officer of Going to School, said he has questions about generalising findings even within his own organisation; he does not assume that programme success in one region means it will achieve the same results when scaled elsewhere.

Development economics borrowed randomised controlled trials (RCTs) from clinical science, and the differences between the disciplines highlight the challenge of generalising results from development evaluations. Clinical trials assess the efficacy of drugs, which have identical chemical compositions across implementation contexts and operate under similar physiological mechanisms in different sub-populations.[9]

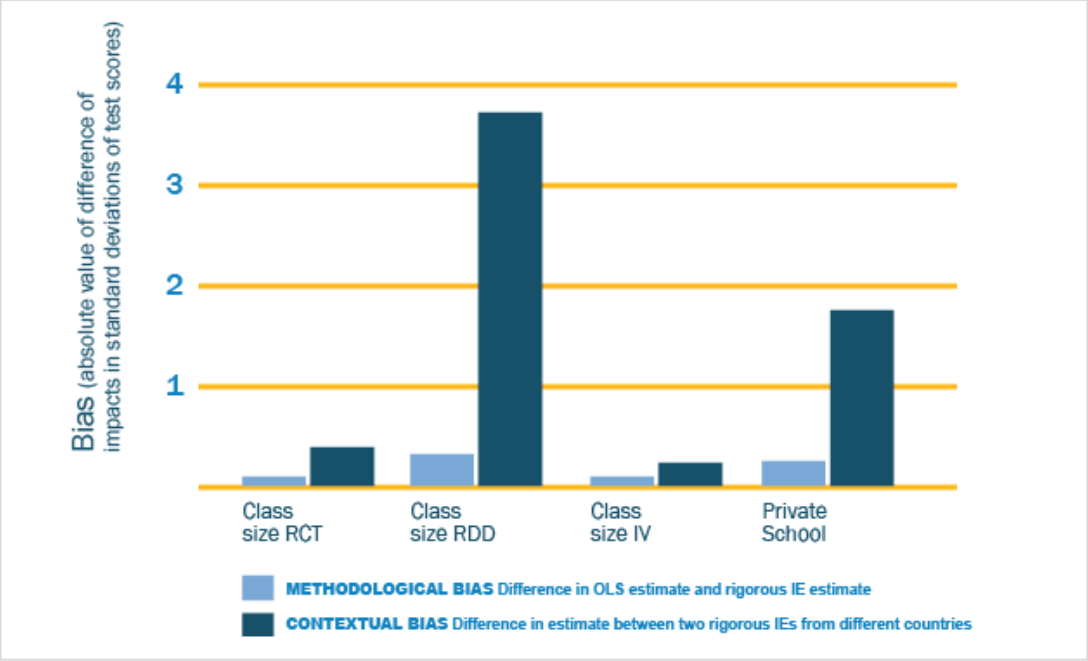
[8] This is not just a problem with RCTs; external validity need not be tied to method.

[9] Even so, external validity has proved challenging for clinical science as well. According to Dennis Whittle, 'medical companies have had exceptional difficulties replicating (in standard laboratory conditions) the results of preclinical RCT-based trials for interventions like cancer *[contd. on next page]*

Arguably, development interventions are more strongly affected by a broader array of contextual factors; a programmatic model implemented in multiple contexts can never be conducted in exactly the same way. Management quality and cultural, political, economic and demographic context are just a few of the many dynamics that influence whether a model can succeed in a given environment.

These concerns are increasingly gaining empirical backing. Pritchett and Sandefur (2013) demonstrate that studies with lower internal validity from the implementing context can offer more accurate impact estimates than studies with higher internal validity from different contexts. They first compare estimates of the impact of class sizes and private schools on test scores and the impact of schooling on earnings (Mincerian returns) between non-experimental ordinary least squares estimates and more rigorous experimental or quasi-experimental estimates from the same context. This comparison provides an estimate of the selection bias associated with non-experimental analysis in these studies. Pritchett and Sandefur then compare impact estimates between multiple experimental and quasi-experimental analyses from different contexts, to estimate the contextual bias introduced when extrapolating from one context to another. As figure 3 illustrates, they find that the bias incurred when extrapolating from one context to another is far greater than the bias incurred when extrapolating from less internally rigorous estimates in the same context.

**Figure 3: Pritchett and Sandefur's estimates of methodological and contextual bias**




Note: RDD stands for regression discontinuity design, and IV for instrumental variables analysis. Figure prepared by IDinsight using data from Pritchett and Sandefur (2013). We calculated the square root of the mean square errors presented in their figure 6 for ease of interpretation.

---

drugs. Even under laboratory conditions, scientists at the drug companies Amgen and Bayer, for example, were able to reproduce the results of only 11 per cent and 21 per cent, respectively, of the RCT-based trials they studied' (2013).

Vivalt (2015) examines a wide range of development impact evaluations, finding that impact evaluation results are highly heterogeneous and are poor predictors of evaluation results of the same intervention-outcome combination in different contexts. She finds that the average coefficient of variation (the standard deviation divided by the mean of a collection of results) for impact estimates from multiple studies on the same intervention-outcome combination was 1.9. This figure is substantially higher than most coefficients of variation from the medical literature, which according to Vivalt typically range from 0.1 to 0.5.

High heterogeneity across studies indicates that contextual factors are major determinants of programme effectiveness, limiting the extent to which one can extrapolate impact evaluation results from one context to another. Furthermore, Vivalt found that the predictive power (as measured by the $R^2$) of a meta-analysis of a particular intervention-outcome combination to an individual intervention is extremely low ($R^2$ ranging from 0.04 to 0.16 in a hierarchical Bayesian meta-analysis, depending on the exclusion criteria).[10]

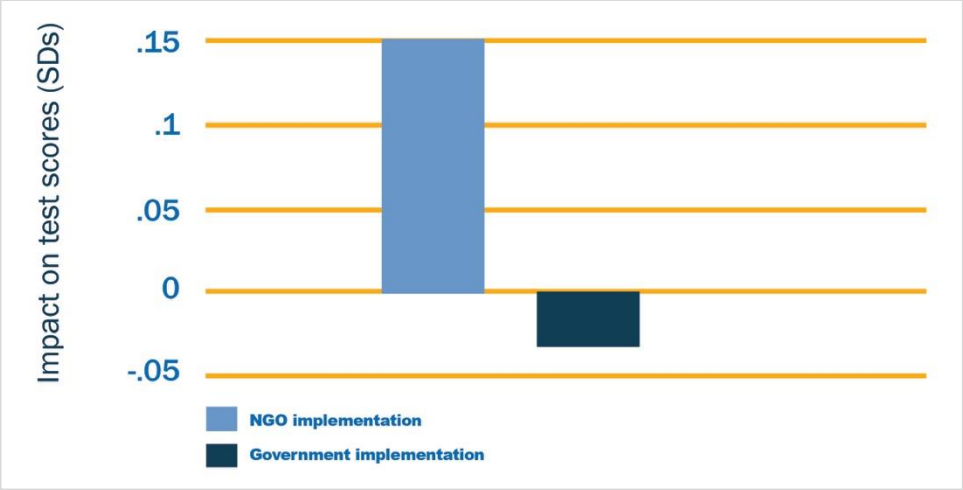
Similarly, Evans and Popova (n.d.) assess six systematic reviews of impact evaluations on the same topic – 'how to improve learning outcomes for children in low and middle income countries' – and found 'massive heterogeneity' within intervention categories. Finally, Bold *et al.* (2013) find that outcomes differ substantially when the same programmatic model is implemented by different organisations, even at the same time and in the same locations (see box 5).

**Box 5: Contract teachers in Kenya: impediments to external validity**

An RCT study by Tessa Bold *et al.* (2013) in Kenya demonstrates that substantial impediments to generalisability may exist even when interventions are replicated in highly similar contexts. The study randomly allocated schools to receive a contract teacher programme managed by World Vision Kenya or by the Kenyan government. Previous evaluations in India and Kenya had found that contract teachers who are employed on short-term contracts demonstrated superior ability to raise test scores than traditionally employed civil service teachers. Similarly, Bold *et al.* found that introducing contract teachers positively impacted maths and English test scores in schools where World Vision implemented the programme. In schools with government-run programmes, there was virtually no improvement in test scores, despite the fact that the government conducted the intervention at the same time, in a similar geography and ostensibly using the same recruitment protocols.

[10] Vivalt (2015) calculates $R^2$ using various other methods; estimates fall in similar ranges.

**Figure 4: Impact of contract teachers on test scores in Bold *et al.* (2013)**




Note: Figure prepared by IDinsight.

Bold *et al.* argue the disparity in outcomes derives in part due to a campaign launched by the teachers' union to alter and dilute the programme in government schools. This response was triggered by a government decision to hire 18,000 contract teachers nationally during the study. World Vision's ability to insulate their hiring model from political pressures therefore proved necessary to achieve positive results. These results demonstrate that even a single contextual difference can threaten the external validity of a study from one environment to another, highlighting the value of obtaining evidence specific to a particular implementer.

If impact evaluations only weakly generalise to other contexts, much of the rationale motivating longer, more expensive KFEs falls into question. Even if a study identifies programmatic impacts with extreme precision, it shows only the impact that specific programme had at that time, in that location, with those implementers, for that population. Broader insights can be extracted from this information, especially if consistent behavioural factors influence the effectiveness of programmatic models across contexts. Implementers may also find it useful to know where a certain model of intervention has succeeded or failed in the past. However, as the literature on external validity is increasingly demonstrating, KFEs are usually poor predictors of whether an intervention will work in a new context and are therefore limited in the extent to which they can guide development action.

Table 2 catalogues the dimensions along which external validity fails. From a social welfare perspective, impact evaluations must have returns that justify the cost, and if the findings of large, expensive evaluations aiming to contribute to global knowledge do not generalise beyond an immediate context, then the cost and time involved may not be justified.

**Table 2: Dimensions along which external validity may fail to hold**

| Dimension | Description |
|---|---|
| *Intervention* | |

| Dimension | Description |
|---|---|
| Design | Interventions are rarely 100% identical across contexts, but small changes can greatly affect intervention impacts. |
| Scale | When moving from small pilots to national scale, many changes can reduce effectiveness or increase cost: process management may be more challenging, new layers of bureaucracy might be necessary, leadership might supervise the program less closely, etc. |
| *Implementer (supply side)* | |
| Capacity | Implementer capacity (e.g. NGO vs. government) can greatly affect intervention impacts. See, for example, Bold *et al*. (2013). |
| Incentives | Different organisations face different objectives (e.g. some look to maximise funding, whereas others try to earn votes or minimise risk). These incentives can affect interest in running certain programmes and ability to deliver. |
| *Beneficiaries (demand side)* | |
| Culture, attitudes and beliefs | Humans tend to respond to incentives everywhere, but even simple games can yield very different results in different cultures (Henrich 2000). The effectiveness of interventions such as public health outreach, which rely on beliefs and attitudes, can vary across cultures. |
| Household characteristics | Programme impact can depend on household characteristics such as education, wealth, occupation, location, socioeconomic status, gender and age. For example, better-educated and wealthier farmers might be better positioned to take advantage of complicated and risky technologies. |
| *Environment/context* | |
| Geography and climate | What works in one location does not always translate to another, e.g. deworming pills will not increase school attendance in areas with no worms, and interventions to increase school choice require a sufficient density of schools to affect education outcomes. |
| History | Historical factors may influence how populations respond to interventions. For example, free delivery of health products might be effective only where there is a history of intense social marketing relating to those products. |
| Politics | Political interests may influence interventions in a variety of ways. For example, implementers may have more political control over frontline workers during a pilot than at scale. Performance pay may 'work' when tested at a small scale, but once there are efforts to roll it out nationally, powerful frontline workers' interest groups (e.g. teachers and health care workers) may organise to block reforms that threaten their interests. |
| Regulatory environment | A small change can produce significant results in one context but fail to produce results in another due to obstructive regulations. For example, business training or loans might foster entrepreneurship in one environment, but entrepreneurial responses to these programmes might be constrained by hostile regulation elsewhere. |
| Complementary institutions | Interventions may require certain formal or informal practices to already be in place to succeed. For example, training farmers to use improved inputs requires that they have access to capital to purchase such inputs; such an intervention might have limited effect in areas with weak financial institutions (formal or informal). |
| Market/ economic context | The roles of the state and market vary dramatically from country to country, limiting the extent to which lessons from one context can be exported to another. For example, health care may be provided free in government clinics in one country and by the private sector in another, making it hard to share programmatic lessons. |
| Temporal changes in any of above dimensions | Changes over time along all dimensions can affect intervention impacts. |

### *3.6.4 Weak link 5: access to actionable evidence*

> 'Since the main focus of most research papers is on the design of the study and the results, many facts that most interest policymakers, such as context, implementation details and costs, are not covered in sufficient detail for policymakers to draw conclusions for their context'.
>
> *—Iqbal Dhaliwal, director of policy at J-PAL (Barder 2014)*

Accessing or understanding the evidence generated by impact evaluations is not a straightforward process for policymakers and practitioners. Findings are often presented in 'technical papers that are targeted for an academic, rather than a practitioner audience' (Dhaliwal and Tulloch n.d.). A survey conducted by InterAction (an international alliance of more than 180 NGOs) on members of its Evaluation and Program Effectiveness Working Group found that almost 60 per cent of respondents felt that impact evaluation reports crossing their desks were not user friendly (Bonbright 2012). A policy stakeholder survey found that improving research dissemination was the most common recommendation provided to improve think tank research in Africa and Latin America, and the third most common in South Asia (Cottle n.d.). The problem of poor research accessibility is compounded by the fact that sifting through evidence to glean relevant and appropriate takeaways is a labour- and skill-intensive task that is easily crowded out by more pressing demands.

> 'Purely evidence-based decision making competes with decision making based on managers synthesising experience, intuition and evidence, and often the evidence is not used because it is not accessible or trusted by the decision maker as much as experience and intuition. It is indeed a "minor miracle" for a purely evidence-based decision to be made, because of this accessibility and trust deficit and particularly if it must override experience and intuition'.
>
> *—Steven Chapman, director of evidence, measurement and evaluation at the Children's Investment Fund Foundation (interview with IDinsight)*

Organisations such as J-PAL, IPA, 3ie and the World Bank have addressed this challenge by writing clear and engaging policy briefs for policy audiences. However, the InterAction stakeholder survey found that policy briefs were among the least preferred mediums to communicate policy evidence. Instead, policymakers 'consistently [said] they prefer user-driven, self-directed information exchanges … to support their work in national policy' (Cottle n.d.). In our interviews with practitioners, a number cited expert consultation as the most important source of information on existing evidence, rather than the literature.

### *3.6.5 Weak link 6: political barriers to acting on external evidence*

Finally, even when evidence is generalisable and accessible to decision makers, there still may be political constraints to using evidence generated outside the immediate decision-making context. Based on research by Matt Andrews, Lant Pritchett, Michael Woolcock and himself, Owen Barder (2014) writes,

> For at least some problems, there is something useful about 'the struggle' – that is, the need for a community to identify its challenges and grapple iteratively with the solutions. If the process of adaptation and iteration is necessary, then solutions parachuted in from outside will not succeed. Furthermore, efforts to bypass the struggle might actually be unhelpful.

A 2015 AidData study finds that assessments of government performance are, on average, more likely to influence government reform if they analyse the government's own data and use country-specific tools rather than cross-country comparisons. The study authors argue that using data generated by the government 'increase[s] the local resonance' of the assessment (AidData 2015). Thus, for some programmes (particularly those that are complex or threaten existing power structures in an implementing organisation or community), even highly applicable evidence may not be taken up if the evidence is generated in a foreign context (Woolcock 2013).

## 3.7 Recent progress

Numerous efforts are underway to improve the link between KFEs and at-scale action. Researcher-practitioner convenings are becoming more commonplace to thoroughly explore the 'question space' that is relevant to both groups. External replication efforts have been undertaken to determine the external validity of different types of intervention, and many initiatives are synthesising existing evidence to make takeaways more accessible and interpretable. There has been increasing activity by organisations such as Evidence Action, J-PAL, IPA and GiveWell to promote specific evidence-based interventions around the world, with funding from a wide range of foundations and bilateral agencies. All of these activities are valuable and may enhance the use of rigorous evidence for at-scale action.

Despite these efforts, fundamental researcher-practitioner disconnects and the spectre of low external validity remain, and therefore will limit the extent to which 'matchmaking' events, systematic reviews and replication studies can address the underlying limitations of KFEs. Until evaluator and practitioner interests are aligned or new evidence indicating greater external validity of impact evaluations emerges, the potential for KFEs to directly influence development action will remain limited.

### *3.7.1 The appropriate role of knowledge-focused evaluations*

KFEs will continue to generate value for the international development sector, primarily through their contributions to development theory. They are better positioned to drive theory than to effect on the ground programmatic change because of the publication-focused incentive structures facing most researchers who design

KFEs.[11] KFEs can also build valuable evidence bases on interventions where we would expect relatively high generalisability, such as for the provision of health commodities (e.g. vaccines). Finally, KFEs can be of greater interest to policymakers when testing new approaches to development as proof of concept research (as with One Laptop per Child or large unconditional cash transfers such as GiveDirectly's programmes). Our interviews suggest that although the international development community is looking for impact evaluations that can influence policy more directly, KFEs retain their utility for refining development theories, building global evidence in specific cases and piloting innovative interventions.

> 'Adding to a body of knowledge is as important as impacting policy. Basically you have no idea what will be useful when. You may be doing something in Bihar that may be useful in Rwanda… [However, the] impact evaluation world needs a much higher dose of the policy vitamin. It's driven, inhabited and incentivised by the need to publish. This needs rebalancing'.
>
> *—A Santhosh Mathew, joint secretary (skills and information technology) at India's Ministry of Rural Development (interview with IDinsight)*

Efforts to externally replicate KFEs across different contexts are still young, and it is not yet clear how valuable they will be. While some replication efforts have demonstrated relatively consistent findings across studies, such as the Ultra-Poor Graduation programme pilots (Banerjee *et al.* 2015), the rising evidence of low external validity raises questions about the expected benefits of replication relative to its cost. We expect that replication initiatives will significantly improve our understanding of the external validity of impact evaluations, produce actionable information on select interventions whose impacts are highly generalisable across contexts, and offer insight into the contextual determinants of programmatic success where generalisability is lower. Replication may be most valuable if it focuses on the theoretical questions of why and how interventions work, rather than searching for models that are cost-effective across most contexts and merit international scale (which, the external validity literature suggests, may be rare).

Despite KFEs' successes, fundamental changes are needed if impact evaluations are to more directly influence action. Below we discuss what changes must be made for impact evaluations to serve as practical tools to inform implementer practices.

## 4. New paradigm: 'decision-focused' impact evaluations

To inform specific, context-dependent programmatic decisions with rigorous evidence, evaluations need to feed into local solution-finding systems. A new breed of impact evaluation – the DFE – has recently emerged. DFEs tailor impact evaluations to the decision-making needs and constraints of implementers. This

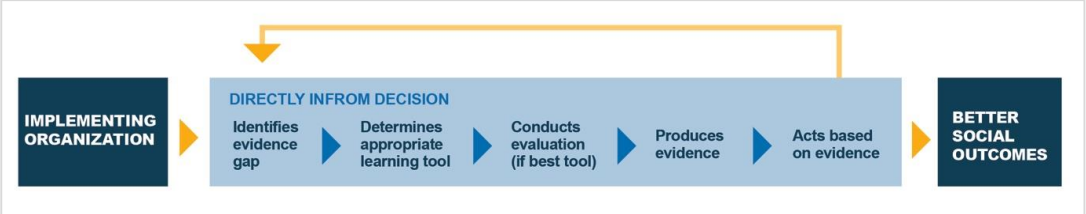
[11] A detailed assessment of the critiques of impact evaluations' ability to advance theory (such as the 'black box' argument) is beyond the scope of this paper. However, we believe that well designed evaluations have demonstrated the ability to answer important theoretical questions, and that these successes warrant continued use of KFEs towards this end.

section is based on IDinsight's experience using DFEs to inform the decision making of development organisations, as well as the views of others we have interviewed working in this space.

## 4.1 Decision-focused evaluations: theory of change

A stylised theory of change for DFEs flows as follows.

**Figure 5: DFE theory of change**




Note: Figure prepared by IDinsight.

DFEs subordinate all other potential objectives to the objective of informing a specific implementer's actions. These evaluations therefore focus only on the first channel described in the KFE theory of change – directly influencing the implementer's programming. Accordingly, the DFE theory of change begins with the implementing organisation rather than the researcher. The implementer identifies evidence gaps in its own knowledge that directly pertain to a pending decision, determines whether an evaluation is a cost-effective method of closing the gap, solicits external support to conduct an evaluation, if needed, directly acts on the results of the evaluation, and repeats this process iteratively.

It is possible for any DFE to contribute to global knowledge. However, such a contribution is secondary, and academic publication is pursued only if it does not interfere with the primary objective of informing the implementer's decision. Finally, although influencing development discourse (channel 2) is not a primary objective for DFEs, by showing that rigorous impact evaluation can be demand-driven and attentive to the priorities and constraints of implementers, DFEs can serve as powerful tools to advance sector-wide thinking around evidence-based decision making.

## 4.2 Characteristics of decision-focused impact evaluations

DFEs have four key qualities:

- Demand-driven – conducted only when an implementer desires evidence to inform future action[12];

[12] DFEs add most value where there is genuine equipoise (uncertainty regarding the optimal course of action) and implementers are willing to act on results in any direction – scaling up interventions that exhibit positive results and restructuring or abandoning interventions that exhibit mixed or negative results. Impact evaluations conducted to lend legitimacy to a *[contd. on next page]*

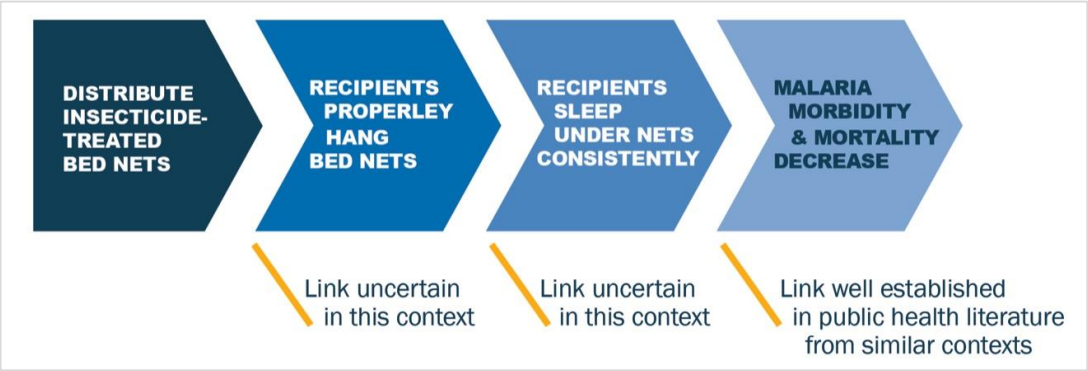
- Tailored – generating decision-relevant evidence within the temporal, budgetary, operational and political constraints of the implementer;
- Embedded – within the implementer's operational and decision-making structures; and
- Cost-effective – aiming for a positive social return on investment (with the evaluation considered as the 'investment').

### *4.2.1 Characteristic 1: demand-driven*

DFEs must be driven by demand, meaning the implementer dictates which evaluation question and approach would be most valuable. Prioritising the implementer's needs from the outset increases implementer engagement with and ownership of evaluation activities and gears evaluations to produce operationally and politically feasible recommendations. It therefore maximises the likelihood that the results will be acted upon.

These knowledge gaps should flow from an explicit mapping of the implementer's theory of change for the given intervention. Mapping helps the implementer identify evidence gaps in their context. It also helps design DFEs that not only assess whether a programme works but also sheds light on the causal mechanisms through which it seeks to effect change. Figure 6 presents a simplified example for a hypothetical bed net distribution programme, illustrating how such a theory of change can identify knowledge gaps a DFE could address.

**Figure 6: Example intervention theory of change**




Note: Figure prepared by IDinsight.

A broad swath of organisations can demand and benefit from DFEs. Any direct implementer of policies or programmes (e.g. a government, an NGO or a social business) can commission a DFE to determine programmatic strategy. See box 6 for an example. We find a growing number of examples of this across sectors. A number of governments, including the United States, have used impact evaluations to directly inform policy (Council of Economic Advisers 2014). J-PAL has partnered with the

programme the implementer has already decided to pursue – although they are demand-driven – do not qualify as DFEs by our definition, as they do not truly inform a decision.

government of Tamil Nadu to use RCTs for policy decisions (Luke 2015), and NGOs such as One Acre Fund, Educate! and BRAC have started conducting their own impact evaluations to refine operations.

**Box 6: 'Mama kits': demand-driven evaluation to inform national policy**

In Zambia, improving rates of facility-based delivery is a strategic government priority to reduce maternal and infant mortality. In 2013 Zambian health officials identified the possibility of using small non-monetary incentives – 'mama kits' – to encourage facility deliveries, but it was unclear whether mama kits would be cost-effective.

Acknowledging this evidence gap, Zambia's health ministries commissioned IDinsight to conduct a cluster RCT in two rural districts to measure the impact of mama kits on facility delivery rates. Several aspects of the evaluation were tailored for policy relevance. To maximise financial viability, sustainability and cost-effectiveness, the government chose to evaluate kits whose contents could be purchased locally for about US$4 per kit. The evaluation was powered to detect an effect size at which the kits would be as cost-effective as other commonly scaled public health interventions.

The three-month study estimated that the kits increased institutional deliveries by 47 per cent, with a cost-effectiveness of US$3,414 per death averted based on existing evidence from Zambia on the link between facility delivery and maternal mortality. Given these results, the government made mama kits available to all health facilities by adding them to the Essential Medicines List. Zambia's health ministries also issued a letter to all cooperating partners recommending they use the kits in their maternal and health programmes. Nine months elapsed between the evaluation being commissioned and the change in Zambia's national health guidelines.

Enabling organisations, such as foundations and multilateral organisations, can also use DFEs to inform their actions. For example, the USAID Development Innovation Ventures initiative and the Global Innovation Fund explicitly tie their scale-up funding decisions to impact evaluation outcomes. In all cases, the fundamental characteristic is that the expected users of the evaluation evidence asks for the evidence for their decision-making purposes.

### *4.2.2 Characteristic 2: tailored*

DFEs are tailored to optimise the evidence available to decision makers within their temporal, budgetary, operational and political constraints. Subjecting evaluations to decision-making constraints often necessitates faster, cheaper and simpler studies (see box 7). If time is short, a DFE may rely on proximate indicators rather than longer-term indicators.[13] To reduce costs, DFEs may use smaller samples designed

[13] Proximate indicators are not always strongly correlated with downstream outcomes. It is therefore important to assess the link between proximate and downstream indicators before relying exclusively on proximate variables. KFEs can improve DFEs by investigating these links, e.g. the link between usage of bed nets and reductions in mortality.

to detect policy-relevant effects and may rely more heavily on routine and pre-existing data. If operations preclude randomisation, quasi-experimental approaches are thoroughly considered.

**Box 7: Approaches to make evaluations faster, cheaper and less operationally disruptive**

- Smaller, focused questionnaires
- Proximate outcome variables (when there is a strong link to the ultimate outcome)
- Skip baseline survey if not needed
- Rolling sample sizes, early stopping rules
- Larger, policy-relevant effect sizes
- Non-randomised designs (e.g. matching, regression discontinuity)
- Electronic surveys/mobile data collection
- Use of existing data collection systems (when there is minimal incentive to report false data; with or without independent audits on a subsample)

The focus on informing discrete decisions clarifies which 'bells and whistles' to incorporate in the evaluation. For example, long household surveys may provide plenty of interesting data that are unlikely to be actionable. A larger sample may facilitate sub-analyses that will not influence any impending decision. In these cases, a DFE will opt for lighter data collection and a smaller sample even if options exist to augment the robustness of the evaluation.

While many DFEs are as internally valid as KFEs, methodological flexibility is sometimes required to accommodate the implementer's constraints and priorities. For example, randomisation of treatment may alter the operational model to such an extent that it no longer mimics the at-scale implementation model. In this scenario, a less rigorous quasi-experimental methodology may be superior to inform decisions (see box 8).

Overall, any methodology that would not meet the implementer's decision-making requirements should not be considered for a DFE. The guiding principle is to improve the evidence available for a given decision. If the existing evidence base is weak and constraints preclude 'more rigorous' approaches, then a methodology with technical flaws may still provide tremendous value.

**Box 8: Evaluating d.light: methodological flexibility to improve evidence for decision making**

The social enterprise d.light delivers affordable solar power and lighting products for low-income households. A requirement of their funding from USAID Development Innovation Ventures was to quantify their model's social impact via a rigorous, independent impact evaluation.

d.light was advised that a randomised evaluation would be the preferred methodology from Development Innovation Ventures' perspective, but all options to randomise had material disadvantages. d.light and its distribution partner recognised that it would be operationally difficult to randomise product rollout. Forcing independent retailers to selectively sell to some customers but not others proved infeasible since the pilot intended to introduce a new product and build a consumer market. Refusing the product to some customers would have created confusion during the product launch. Moreover, randomisation at the geographic level was financially infeasible, as it would have effectively doubled d.light's geographic footprint.

d.light also considered options to offer the products free of charge or at heavily subsidised prices to randomly selected households. This approach might have provided interesting findings on whether solar systems could improve household welfare, but the evaluation would have investigated a sales model d.light would not pursue in any other scenario, with unsustainable price points, in a geography where d.light did not intend to work and in households much poorer than their typical clients. For these reasons, the findings would not have been relevant to d.light's actual market priorities.

Ultimately, d.light and Development Innovation Ventures worked with IDinsight to design a prospective matching study that measured the impact of d.light's actual operations. IDinsight interviewed 500 households that had just purchased d.light home solar systems at baseline and used statistical matching to identify 500 similar households in nearby villages as a comparison group. This design enabled a difference-in-differences analysis on outcomes such as lighting usage, energy expenditure and socioeconomic and health metrics.

In conclusion, an RCT might have had higher internal validity but would have produced a result with limited utility to d.light or their funder, given Development Innovation Ventures' interest in market-based solutions. As a result, they used a methodology with larger technical weaknesses but with much higher relevance to their operations and decision-making needs.

### *4.2.3 Characteristic 3: embedded*

The most effective DFEs occur when the evaluation function is embedded in an organisation's decision-making apparatus. External and independent viewpoints can enhance evaluation quality and credibility, but integration with routine organisational processes ensures alignment with organisational needs and constraints.

In this scenario, evaluators are akin to trusted strategic advisors. Just as a chief financial officer uses her financial understanding to guide organisational actions, an evaluator uses her expertise to guide critical intervention design and scale-up decisions. The relationship between the evaluator and the executive should be dynamic, continuous and highly consultative, allowing the evaluator to maximise her influence on action and enabling the executive to integrate social impact

considerations in all her major decisions. To function properly, this setup requires at least one ‘champion’ within the implementing organisation who recognises the importance of basing decisions on evidence and can help integrate the embedded evaluation unit into decision-making processes.

To ensure alignment, decision-focused evaluators should use an iterative consultative process to design any evaluation and should regularly consult implementers throughout all evaluation phases. An design ‘menu’ that communicates critical trade-offs between competing designs in non-technical language can be useful to ensure implementer interests inform the evaluation design (table 3 is an illustrative example). Beyond design, the evaluator should maintain communication with the primary decision makers to solicit feedback and make course corrections when necessary. This regular engagement increases implementer ownership and maximises the fit between the evaluation design and the decision-making context.

**Table 3: Illustrative evaluation menu**

| Evaluation options | Rigor | Time | Cost | Operational demands |
|---|---|---|---|---|
| *High rigor:* RCT with long-term follow-up | • High internal validity<br>• Measure final outcome | 4 years | US$1 million | Randomise on village level |
| *Faster option:* RCT with short-term follow-up | • High internal validity<br>• Requires proximate indicator to link with final outcome | 1 year | US$300,000 | Randomise on village level |
| *Faster, simpler option:* differences-in-differences with short-term follow-up | • Requires more assumptions<br>• Requires proximate indicator | 1 year | US$300,000 | Preserve treatment areas |
| *Fastest, cheapest option:* regression using historical program data | • Requires assumptions about counterfactual<br>• Need to verify data quality | 3 months | US$30,000 | None |

Thus, the decision-focused evaluator must be both technically and interpersonally proficient. To tailor evaluations to decision-making needs, she ideally accesses the full methodological toolkit while maintaining deep understanding of and influence over the decision-making context. Unfortunately, our experience indicates that many implementing organisations lack staff sufficiently skilled in impact evaluation, and few evaluators sufficiently emphasise interpersonal proficiency and a client-service orientation as critical skills. We discuss this situation in greater depth in section 5.

In the short and medium term, most implementers will continue to rely on external providers for DFEs. In such scenarios, it is important for the providers to maximise their integration with decision makers in the implementing organisation. Ideally, evaluation contracts would span multiple studies to foster deep relationships, encourage testing of new ideas as they arise and dilute fixed contracting costs. In

this way, evaluation engagements can facilitate tighter evidence-generating and decision-making cycles. Box 9 illustrates the benefits of such a relationship.

**Box 9: Concurrent evaluation cell: embedding evaluators in a public company**

In 2013 the Bihar State Power Holding Company Ltd. (BSPHCL), a public electric utility company based in Patna, India, sought analytic support to address a number of challenges. Foremost were low levels of revenue collection: for every dollar billed, the company received only US$0.48.

BSPHCL worked with IDinsight to build a concurrent evaluation cell (CEC) housed at BSPHCL headquarters and consisting of several full-time IDinsight staff. The CEC originally had four objectives: evaluate pilot programmes aimed at increasing revenue; conduct process evaluations of ongoing electricity distribution projects; analyse BSPHCL data as requested; and advise on general data collection and management practices.

The CEC helped BSPHCL evaluate a programme that assigned community members to manage meter reading, bill distribution, and revenue collection for households in their neighbourhoods and provided compensation based on the amount of money collected. The evaluation found that the pilot programme expanded the base of paying electricity consumers and increased collection efficiency (the amount of revenue collected per the amount billed) for small, non-arrears paying customers. The CEC has performed a number of other analyses for BSPHCL, including ranking the performance of infrastructure repair agencies, depicting consumer payment method preferences and conducting process evaluations on consumer complaint offices.

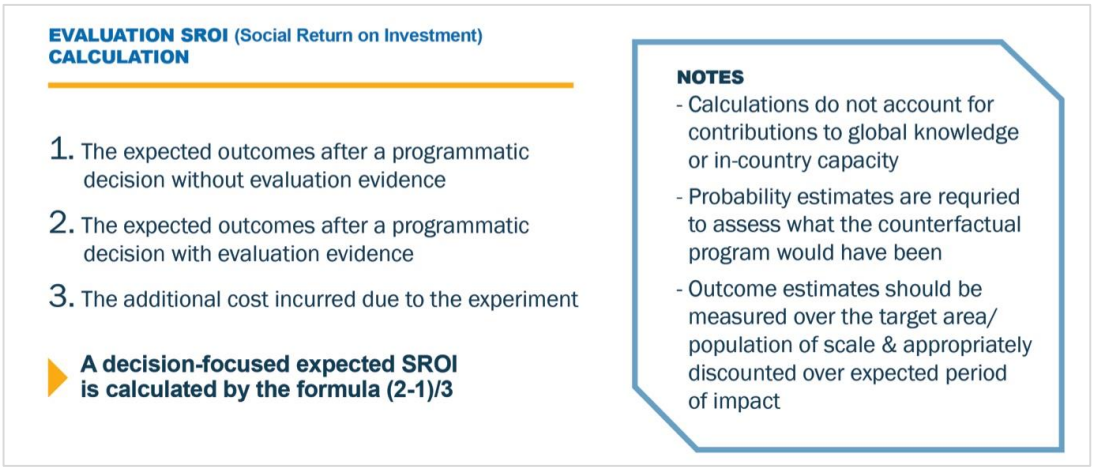
IDinsight is working on expanding the scope of the CEC to design a data collection and management framework for the state that can be used to increase employee accountability and automatically flag issues for management to investigate. This system will help energy sector leaders use data to make decisions without the support of external consultants. Although this goal will only be achieved over the long term, the energy sector leadership has expressed enthusiasm for this data system.

Other evaluation organisations are working to embed in-house evaluation capacity in government bodies. J-PAL, for example, is helping the Haryana State Education Department in India establish an internal monitoring and evaluation unit (Dhaliwal and Tulloch n.d.).

#### *4.2.4 Characteristic 4: cost-effective*

DFEs should be viewed as investments in evidence generation to improve social outcomes. Assessing evaluation cost-effectiveness can be done using a very simple social return on investment framework that compares expected outcomes following a ‘less informed’ decision to those following a decision informed by impact evaluation evidence. The framework shows how the optimal evaluation may not use all the money available for evaluation and how it is possible for a cheaper evaluation that yields less robust findings to be superior to a more rigorous but also more expensive evaluation. Investment framing also encompasses issues of equipoise and potential for scale up. The social return on investment framework can be augmented in many ways, but even the simplest version (see figure 7) provides very useful guidance to understand whether a proposed DFE is likely to be justified.

**Figure 7: Simple framework to assess a DFE's social return on investment**




Note: Figure prepared by IDinsight.

Box 10 offers an example of one outcome from a cost-effective DFE.

**Box 10: Latrines in Cambodia: DFE improving cost-effectiveness**

iDE is an international NGO that uses market-based solutions to address agriculture and sanitation challenges. In Cambodia, iDE trained local manufacturers to make and market latrines for rural households and was seeking new ways to increase latrine purchases.

In 2013, iDE and IDinsight launched a randomised evaluation to assess the potential impact of retail financing on latrine purchases. Conducted in 30 villages, the study measured consumer willingness to pay and latrine uptake when they were offered one of two options: cash on delivery or microfinance. The results were produced three and a half months after iDE identified latrine financing as a priority question; the evaluation cost roughly US$60,000. The study showed that microfinance substantially increased willingness to pay for latrines and uptake of latrines. The average household without a latrine would be willing to pay US$30 for a latrine under cash on delivery, whereas the average household offered a microfinance loan would be willing to pay US$50 for the same latrine. At the US$50 market price, 50 per cent of targeted households purchased latrines when offered financing, but only 12 per cent of households purchased when they were required to pay up front.

These results indicate that, at scale in rural Cambodia, financing can open up 280,000 additional households to latrine purchases. Moreover, due to the higher conversion rate, iDE could decrease its administrative and marketing costs per latrine sold by as much as 70 per cent. Achieving the same volume of latrine sales at scale without financing would have required a substantial subsidy on the latrines, costing an estimated US$5.7 million. Given that the evaluation cost only US$60,000, this information offers a substantial return on investment even if it increased the probability that iDE would scale financing by only a small amount, relative to the counterfactual of not conducting the evaluation.

iDE has since been investigating how best to scale latrine financing across their Cambodia operations. This effort has encountered new operational challenges: iDE would like to sell at least 36,000 latrines on microcredit each year, but their microfinance partners have the capacity to offer only 10,000 loans. iDE is therefore exploring alternative scaling mechanisms, including in-house financing.

According to Yi Wei, WASH innovation manager at iDE, ‘What’s notable about this experience … is that iDE has been confident about the impact of financing despite all the implementation challenges because of the evidence from the evaluation’. Translating DFE findings to at-scale implementation is often not easy, but knowledge from the evaluation can help determine whether the effort is worthwhile and what resources should be devoted to ensure success.

## 4.3 Advantages and limitations of decision-focused evaluations

The advantages and limitations of DFEs can be assessed across the four primary channels outlined in the KFE theory of change. DFEs enable immediate and efficient use of evaluation evidence, which has been an area of weakness for KFEs. DFEs will also shape sector dialogue around impact measurement in a manner similar to KFEs. However, DFEs are somewhat less well suited to share intervention-specific knowledge globally, and their contributions to development theory will be limited.

### *4.3.1 Channel 1: directly informing decisions*

Improving social impact within the implementing organisation is the primary objective of a DFE. Being demand-driven, tailored, embedded and cost-effective enables DFEs to fully align with decision-making needs. DFEs are therefore better suited to influence immediate programmatic action than KFEs.

External validity poses less of a threat to DFEs than KFEs. In most cases, DFEs assess interventions in the context in which they would ultimately be scaled. Decisions to scale, modify or discontinue the programme are made immediately after results are established, mitigating concerns about conditions changing over time. Although external validity weaknesses can remain, as implementation conditions may still change significantly during scale-up or programme modification, they are typically much less extreme than when attempting to transport findings across time, space and implementing institutions.

For DFEs to be demand-driven, however, the implementing organisation must have an awareness of its own evidence gaps (see box 11). Implementers may not recognise when they lack evidence to support their decisions, and they may not understand when impact evaluations are appropriate for their needs. DFEs are therefore most likely to be used by organisations possessing some familiarity with impact evaluations and their uses, or when there is a close relationship and reciprocal communication between the implementer and evaluator.

**Box 11: Zambia's distribution of insecticide-treated nets: rapidly generating rigorous, actionable evidence**

In 2014 Zambia's National Malaria Control Centre (NMCC) planned to distribute 6–7 million insecticide-treated nets nationwide. Existing guidelines dictated that the nets be distributed door-to-door, with community health workers hanging up each one, but the NMCC wondered if there were a more cost-effective approach. One alternative was to distribute the nets at 'fixed points', e.g. schools and churches. However, there were concerns that fixed-point distribution would mean low household attendance, low retention of the nets and low use.

Four months before the 2014 distribution guidelines would be set, the NMCC commissioned IDinsight to conduct a randomised evaluation examining two distribution options: fixed-point distribution with no community health worker visit to hang the nets and fixed-point distribution plus a 'hang-up' visit 1-3 days, 5-7 days, 10-12 days or 15-17 days after distribution. The study was implemented in three communities in Zambia's Eastern Province; households were randomly assigned to groups that differed on the timing of the hang-up visits. The study found high household attendance (96 per cent of pre-registered households) and high insecticide-treated net ownership and use across all study arms.

Preliminary findings were delivered to the NMCC four months after the study was commissioned, in time to influence the 2014 distribution guidelines. In response, the NMCC decided to loosen the guidelines to allow communities to distribute the nets through fixed points. The evaluation and policy engagement cost US$150,000, but it enabled the Zambian government to enjoy savings of up to US$2 million and achieve material efficiency gains nationwide.

#### *4.3.2 Channel 2: influencing development discourse*

The conversation catalysed by DFEs can be seen as an extension of the discourse started by KFEs. The issues DFEs raise dovetail cleanly with many ongoing debates around results-based financing, sector incentives and social sector measurement activities. Most DFEs involve governments, funding agencies or other 'market setting' organisations whose thinking and practice are likely to evolve due to their involvement with such evaluation activities. Furthermore, by demonstrating that rigorous impact evaluation can be demand-driven, affordable and responsive, DFEs can help stem some of the backlash against rigorous evaluation within the development community. The avenues through which DFEs can influence the sector are described in greater detail in section 5.

#### *4.3.3 Channel 3: accumulating global evidence*

While results from DFEs will often be informative beyond the immediate implementation context, they are much less likely to influence the decisions of outside implementers than those of the commissioning organisation, and they may contribute less to the global evidence base than KFEs, because their demand-driven nature of DFEs may lead them to answer questions that are specific to the circumstances of the implementer, limiting their international relevance.

In addition, internal and external validity outside the implementer context can be lower in DFEs than in KFEs. DFEs are often as rigorous as KFEs in terms of internal validity, but as discussed above, it sometimes makes sense to reduce rigor on one or more dimensions if doing so enables better information to be used at the decision-making moment. There can also be greater uncertainty about external validity if a DFE fails to collect robust contextual information.

Finally, given publication costs, results from many DFEs may not be disseminated widely. However, efforts can be made to disseminate their results to relevant audiences in a way that does not distract from the immediate goal of helping a particular implementing organisation make a better decision.

#### *4.3.4 Channel 4: advancing development theory*

DFEs are unlikely to address questions that advance overarching development theories, because wider knowledge generation objectives are subordinated to decision-related objectives. And because results from DFEs may not be published or disseminated widely if such activities require significant marginal investments, they are less likely to inform academic discourse. This does not rule out the possibility for certain DFEs to advance theory, but the vast majority will not do so.

### 4.4 The future of impact evaluation: clear objectives, appropriate actions

It is important to define any evaluation's priorities and objectives up front, as trade-offs quickly appear between pursuing a 'decision' versus a 'knowledge' agenda. Once evaluation objectives are established, the decision of whether the evaluation

should have more KFE elements or more DFE elements becomes clear. Table 4 demonstrates the relative strengths of DFEs and KFEs for different impact channels.

**Table 4: Appropriate uses of DFEs and KFEs**

| Impact channel | Primary objective | Evaluation type |
|---|---|---|
| Channel 1 | Inform decision to fund or scale a specific intervention | DFE |
| Channel 1 | Inform implementer's decision of which programmatic variant to scale | DFE |
| Channel 3 | Build evidence on interventions with high expected external validity | KFE |
| Channel 3 | Explore the impact and mechanisms of new interventions | KFE |
| Channel 4 | Contribute to development theory | KFE |

Unfortunately, the objectives of impact evaluations are often not clarified *a priori*. Several practitioners told us how an initial demand for an evaluation started with a decision focus, but it became increasingly knowledge-focused and less aligned with their original goals over the course of engaging with evaluators. Stakeholders may commonly expect decision-related outcomes from KFEs and global knowledge-related outcomes from DFEs. Such confusion about evaluation objectives may reduce the overall effectiveness of any impact evaluation.

Table 5 compares stylised characteristics of KFEs and DFEs.

**Table 5: Characteristics of KFEs and DFEs**

| Characteristic | KFE | DFE |
|---|---|---|
| Question source | Evaluator (with input from implementer) | Implementer (with input from evaluator) |
| Evaluator | Outside technical expert | Embedded policy advisor |
| Time to release findings | 1–5 years | 1–24 months |
| Cost | US$100,000 – US$5 million | US$10,000 – US$500,000 |
| Methodology | Lower diversity with emphasis on more robust methodologies and downstream outcomes | Higher diversity with greater emphasis on proximate outcomes and practical considerations |
| External validity | Significant concerns due to intention to apply findings across contexts | Reduced concerns, since action is intended to occur in implementers' context |
| Definition of success | Contribution to development theory, contribution to high-level policy debates, scale-up of generalisable interventions | Informed decision and at-scale action, or programme discontinuation, in implementer's context |

Finally, it is useful to acknowledge complementarities between KFEs and DFEs. In many cases, for example, KFEs provide the theoretical foundation to construct different intervention-outcome combinations from which DFEs can be used to refine intervention design and implementation. KFEs can also shed light on which contextual factors are necessary for programmatic success, informing when and how organisations conduct DFEs.

# 5. Realising the new paradigm

To achieve an ecosystem that fosters the appropriate use of DFEs, development organisations – especially funders – will need to build a market for their use. To do so, development organisations should aim to both **stimulate demand for** and **build the supply of** DFEs**.**

## 5.1 Stimulating demand for demand-focused evaluations

Although we have observed a growing number of funding and implementing organisations looking to conduct DFEs, demand remains low due to several inefficiencies across the development sector. The InterAction survey of its member organisations on the use of impact evaluations, for example, found that more than 70 per cent of respondents did not feel their organisation emphasised the use of evaluation results, more than 60 per cent felt that impact evaluations did not raise their organisation's credibility, and more than 40 per cent reported low comfort with evaluation fundamentals (Bonbright 2012). Broad efforts to better align the incentives of development organisations to achieve impact, combined with targeted efforts to encourage the use of DFEs, are needed to promote optimal usage levels.

**Table 6: Strategies to spur demand for DFEs**

| Market gap | Solution | Key actors |
|---|---|---|
| 1. Implementers are not incentivised to achieve impact | Create 'impact first' incentive systems that tie scale-up funding to demonstration of impact over a long time horizon | Funding organisations |
| 2. Implementers lack funding to learn through DFEs | Allocate dedicated portion of M&E/project funds for DFEs | Funding organisations |

### *5.1.1 Recommendation 1: improve implementer incentives to demonstrate and achieve impact*

> 'If a program can already generate sufficient support to be adequately funded then knowledge is a danger'.
>
> *—Lant Pritchett (2002)*

Under current funding and accountability systems, most implementers have little reason to conduct impact evaluations. Implementers typically reap few rewards from demonstrating impact, and there are material risks to discovering that their impact is small. Our interviews indicated

that impact measurement activities are frequently tangential to the dynamics of running and sustaining an implementing organisation. While a growing number of pioneering organisations are commissioning DFEs, the number of implementers using impact evaluations to inform decisions will remain small until this situation changes.

Two themes repeatedly emerged in our practitioner interviews:

- Most funders emphasise financial accountability and 'countable' metrics (e.g. inputs, activities and outputs) and do not directly incentivise implementers based on rigorously demonstrated impact; and
- Implementers may welcome a greater emphasis on impact as long as there remains room to admit failure and experiment to improve.

The emphasis on countable metrics, rather than outcomes and impact indicators, stems from accountability structures within the funding bodies themselves, where staff are not held accountable for demonstrating impact. It is therefore no surprise that impact-oriented incentives are not passed down to implementers, and that most donor-supported implementers view impact evaluations as secondary to their immediate objectives.

To reverse this trend, governments, foundations and other 'market setting' funding bodies must restructure accountability systems to collectively incentivise evidence generation and use. These organisations should aim to create an 'impact first' funding ecosystem that measures progress and allocates resources based on rigorous evidence of social impact. Establishing these incentive structures throughout the funding chain would promote effective downstream use of DFEs.

> 'Because program officers at foundations build relationships with their implementers and oftentimes must advocate for grants to the board of the foundation, negative results can be hard to swallow and can be swept under the rug'.
>
> *—Senior officer at a leading foundation (interview with IDinsight)*

To function properly, 'impact first' accountability systems must balance the priorities of allocating funding to its most impactful uses, while allowing implementers to demonstrate and learn from failure. Our interviews suggest that staff at implementing and funding organisations face disincentives to conduct impact evaluations due to the risk of exposing failure. Efforts to promote impact evaluation as a decision-making tool are therefore likely to face substantial resistance unless implementers feel safe to fail without fear of immediately losing funding, and funding staff have the freedom to evaluate and expose failed programmes in their portfolios. 'Embracing' failure is especially important to allow organisations to try innovative approaches and to encourage stakeholders to learn as much as possible from these attempts.

The freedom to fail need not be at odds with tying funding decisions to demonstrated impact. Implementers in our interviews agreed that funding should be tied to an organisation's evidence of impact, but that there should be longer time horizons (e.g.

5–7 years) over which they have free range to experiment, fail, learn and improve before funding decisions are directly tied to evidence. This is particularly important for newer programmes and organisations that are still refining their programme models.

An ‘impact first’ funding ecosystem would therefore allow far greater flexibility in activities, inputs and outputs, giving implementers freedom to collect whichever monitoring data are most useful to inform daily operations. It would demand a more rigorous assessment of impact, after an appropriate and context-dependent period of experimentation and iteration in the programme model (the learning phase).

Such an ecosystem would call for DFEs both during and at the end of the learning phase to guide future funding decisions. During the learning phase, implementers would use DFEs (along with other tools) to help experiment with and test different operational models and make course corrections to continually improve their programme. At the end of this initial phase, a different type of DFE – an independent evaluation of aimed at determining whether the intervention warrants further funding – would be conducted. While many programmes cannot be easily evaluated in an impact evaluation (e.g. infrastructure investments), the subset that can is large enough that such an accountability system could substantially improve development efforts.

Impact-oriented funding structures, such as development impact bonds and results-based financing, could play a major role in shifting the funding ecosystem towards this ‘impact first’ model. Standardisation of reporting formats across funding bodies to request implementers to make a case for impact could also help synchronize this effort across the sector.

### *5.1.2 Recommendation 2: allocate dedicated funding for decision-focused evaluations*

Under a funding ecosystem that perfectly incentivised the maximisation of social impact, optimal usage rates of DFEs would naturally emerge. This is an impossible ideal, given the challenges associated with measuring impact and the flaws inherent to any bureaucratic accountability structure. Especially while DFEs are a relatively new concept, targeted efforts to promote their use can speed the process of uptake.

Though they are cheaper than KFEs, DFEs are still expensive for most implementing organisations. On large projects, funders (whether national governments or international donors) can promote adoption by setting aside a portion of the programmatic implementation budget for one or more DFEs (contingent on implementers seeing value in a DFE), just as they would for basic monitoring activities. Funders can also create crosscutting funding windows to support high-impact DFEs across their portfolios, to which implementers or funding staff can apply on a case by case basis. Finally, funders can launch grants supporting multiple DFEs across a given implementer’s project portfolio over a set period, allowing the grantee to determine which areas of their programming and which questions would benefit most from rigorous evaluation.

Funders and implementers should keep in mind that DFEs are not the appropriate tool for every programmatic question, or even most programmatic questions. The long-term objective, therefore, should be to develop robust solution-finding systems in which DFEs represent one of many tools. However, given that experience with DFEs is still rare among implementing organisations, dedicated funding for DFEs in the coming years could move the sector closer towards the optimal DFE usage rate. Such funding will help implementing organisations internalise the merits of DFEs, develop the capacity to conduct more DFEs and experiment with different approaches to impact evaluation, which may further inform how the development community approaches and structures the tool.

## 5.2 Building the supply of decision-focused evaluations

We have argued that to maximise the social impact of impact evaluations, there must be a dramatic expansion of DFEs. Since academic incentives to pursue KFEs are unlikely to change soon, this begs the practical question: who will have the skills and incentives to conduct DFEs? This is a medium-term challenge for the international development community that requires directly building the supply of individuals and organisations capable of carrying out DFEs.[14]

**Table 7: Strategies to build supply of DFEs**

| Market gap | Solution | Key actors |
|---|---|---|
| 1. Need for non-academic impact evaluation specialists | Develop professional tertiary education programmes to train evaluators | Universities and funding organisations |
| 2. Dearth of high-quality impact evaluation organisations operating on a demand-driven basis | A) Subsidise start-up funds for DFE organisations to create a competitive market of quality DFE providers | Funding organisations |
| | B) Provide rapid external quality reviews of evaluation designs and analysis plans | International evaluation organisations (e.g. 3ie) |
| 3. Evaluators are not incentivised to prioritise decision-relevance | Publish the cost, length and actions influenced by impact evaluations in evaluation registries | Evaluation registries |
| 4. Low capacity to generate and use evidence among implementers | Fund 'build, operate, transfer' evaluation cells to embed evaluation capacity into implementing organisations | Funding, evaluation and implementing organisations |

[14] Our recommendations are most relevant for Africa and Asia, where we have more experience. They may be less relevant for Latin America, where there has been greater institutionalisation of impact evaluation. For more information on Latin America, please see the 3ie working paper by GRADE, which is part of this series.

### *5.2.1 Recommendation 3: develop professional tertiary education programs to train evaluators*

We have observed in our own hiring processes that relatively few development professionals possess quantitative impact evaluation expertise, management experience and client-facing skills, each of which is crucial to running a successful DFE. This gap may be partly due to the fact that many programmes teaching impact evaluation methodology are geared towards academic career tracks. There are strong professionally oriented programmes that teach these skills (e.g. the public administration/international development master's programme at Harvard University's John F Kennedy School of Government), and individuals from PhD programmes often do pursue professional rather than academic lines of work, but there are relatively few graduates from these programmes.

An effort to expand tertiary education programmes (especially master's degrees or certificates) that train development professionals to conduct high-quality impact evaluations would reduce the costs of DFEs and make it easier to manage them at a high level of quality. This objective is especially important because, as we have observed in our own work, management staff typically comprise a major portion of DFE budgets.

Organisations such as J-PAL, IPA, the Centers for Learning on Evaluation and Results and the World Bank are making important steps in the right direction by publishing books, launching online courses and leading training programmes that teach the fundamentals of evaluation. Universities can go several steps further by expanding their course selections on evaluation methodology in public policy and development programmes, and by creating one- or two-year master's programmes focused primarily on impact evaluation.[15] It would be especially useful for tertiary education programmes in developing countries to expand such offerings. Funding organisations can further promote these efforts by funding programmes.

### *5.2.2 Recommendation 4: develop organisations equipped to conduct high-quality DFEs*

Implementers commonly bemoan the lack of providers who can design and implement impact evaluations aligned with their decision-making needs. We heard this complaint in multiple interviews with funders and practitioners, some of whom worried that without an increase in the number of organisations equipped to conduct high-quality impact evaluations, a drive to increase demand for DFEs could risk a proliferation of low-quality studies. For DFEs to effectively inform action across the development sector, a larger, more diverse pool of impact evaluation providers with decision-oriented incentives is needed.

[15] See, for example, Oxford University's master of science in social policy, which places a strong emphasis on impact evaluation.

The international development community can expand the supply of organisations capable of executing high-quality DFEs by providing subsidised start-up funding to seed new demand-driven impact evaluation service providers. The ultimate goal is to create a robust, competitive market of impact evaluation providers that respond to the needs of discerning 'consumers' (funders and implementers) of impact evaluation services.

To support these organisations and offer sector-wide quality control, international organisations that promote and specialise in impact evaluation – especially 3ie – can review impact evaluation methodologies and analysis plans on a demand-driven basis. These reviews should offer rapid turnarounds to fit into the decision-oriented timelines of DFEs. They can serve as a general quality check, providing the funder and implementer with confidence in the rigor of the evaluation, and answer complex methodological questions that require the input of expert statisticians.[16] These reviews lessen the need to employ such expertise full-time, reducing the barriers to entry and the costs of conducting DFEs.

#### *5.2.3 Recommendation 5: reform evaluator incentives to encourage DFE*

Most impact evaluation forums and structures do not facilitate a holistic view when judging impact evaluations. Academic publications judge only on technical rigor, methodological innovation and contributions to development theory, and evaluation registries typically describe only data, evaluation methods and findings.

To encourage more impact evaluation providers to adopt decision-focused approaches, it is important to refine the evaluation community's definition of a successful evaluation. In most cases, technical rigor and contribution to global knowledge are the pre-eminent criteria for success. However, time, cost and use of the evidence produced should also be considered when judging the success of any evaluation, and evaluators should be professionally rewarded based on these factors as well.

To optimise the value that evaluations contribute to development practice, structures can be changed to incentivise relevance to practice. For example, in addition to research questions, methods, results and analysis plans, evaluation registries could publicly record the cost of the evaluation, time from evaluation start to final findings and tangible use of findings for at-scale action. Such transparency can influence observers to take a holistic view of the value (both in terms of knowledge generation and decision influence) produced by any given evaluation.

#### *5.2.4 Recommendation 6: build implementer capacity to conduct impact evaluations with 'build, operate, transfer' evaluation cells*

In addition to soliciting support from external evaluation organisations, implementers may choose to establish in-house capacity to conduct their own DFEs. To promote

[16] 3ie has provided such services to IDinsight in the past.

these efforts, funders and evaluation organisations should adopt ‘build, operate, transfer’ models that enable deep and sustained capacity building.

In our experience, impact evaluation capacity building for implementing organisations has been largely limited to ‘training workshops’. These are valuable as a first step to ensure that funders, implementers and government officials understand key measurement and evaluation concepts. Unless policymakers and programme managers understand evaluation fundamentals (e.g. the need to account for the counterfactual) and the particular advantages of evidence produced by impact evaluations, the likelihood that DFEs will be used is low. Many organisations – including 3ie, the World Bank, IPA and J-PAL – have been active in practitioner training. Such activities are essential to expose development organisations to impact evaluation’s uses, but they do not effectively transfer capacity to conduct impact evaluations.

To actually transfer the capacity to conduct impact evaluations and use their results to inform programmatic decision to implementing organisations, a more sustained, deliberate model of capacity building is needed. Adapting the ‘build, operate, transfer’ model of public-private partnerships in infrastructure projects could be a valuable approach for transferring significant capacity to understand, conduct and use impact evaluations within implementing organisations:

1. **Build:** The evaluation organisation builds and staffs an evaluation unit *within* an implementing organisation;
2. **Operate:** The evaluation organisation operates the evaluation unit within the implementing organisation for several years;
3. **Transfer:** The implementing organisation ‘seconds’ an appropriate number of its staff to the unit run by the evaluation organisation, to be trained on the job over several years.

The goal of this model is to transfer the full capabilities needed to execute all of the elements of a DFE: prioritising evaluation questions, designing and conducting impact evaluations, and interpreting and using findings to inform programme operations. However, given current capacity in most implementing organisations, developing in-house evaluation capabilities will be a long-term endeavour.

## 6. Conclusion

Rigorous impact evaluations have risen to prominence as part of a global learning agenda. These efforts have deepened our understanding of human behaviour, driven the development community to think more critically about impact measurement and promoted international scale-up of several highly cost-effective development interventions.

This paradigm has not, however, transformed how development is conducted on the ground. This is, in large part, because the academic incentives that drive most KFEs are frequently misaligned with the priorities and constraints facing decision makers.

Moreover, external validity challenges limit the usefulness of KFE results to practitioners outside the setting of the original study, who face highly context-specific operational questions. KFEs will continue to advance development theory, generate globally useful evidence where external validity is high and explore innovative interventions. However, impact evaluations can and should serve a more influential role in directly informing development action.

We have argued that a paradigm shift is needed – towards one that increasingly employs impact evaluations as tools embedded in localised solution-finding systems. These DFEs prioritise the needs of decision makers, tailor research methodologies to implementers' operational constraints, embed learning into decision-making processes and seek to maximise their own cost-effectiveness. They tend to be shorter, cheaper and more relevant to the decision-making needs of the implementing organisation than KFEs, although in the process they may forego some internal validity or relevance to theoretical questions.

We envision an impact evaluation landscape that strategically employs KFEs and DFEs where they add the most value. KFEs should remain a mainstay of academic efforts to deepen our understanding of development theory; DFEs should serve as a tool (among many others) for implementers to tighten the link between evidence and action. While KFEs and DFEs overlap substantially, greater clarity on their respective strengths and weaknesses will help development organisations identify the most appropriate evaluation to achieve their objectives.

Despite the powerful role DFEs can play in improving development efforts, imperfect incentives and limited capacity to conduct impact evaluations outside academic settings constrain their use. We have therefore advanced several strategies by which funders, implementers and research organisations can promote the use of DFEs. To stimulate demand, funders should hold implementers accountable to achieving impact over longer time frames and allocate dedicated funding pools to support the use of DFEs. To increase supply, universities should offer more professional courses in evaluation science, foundations should advance seed funding for DFE providers; evaluation support organisations such as 3ie should provide on-demand methodological review services for DFEs, evaluators should be incentivised to prioritise the decision-relevance of their studies, and implementers should experiment with 'build, operate, transfer' models to develop in-house evaluation capacity.

The DFE approach to impact evaluation comes with limitations and risks. Rigorous impact evaluations may prove prohibitively complex for most organisations to take on themselves, and it is yet to be seen whether a 'build, operate, transfer' model would bridge this capacity gap. A poorly conducted evaluation may lead decision makers astray, especially if it legitimises incorrect prior beliefs. Raising demand for DFEs without increasing the supply of high-quality DFE service providers could exacerbate this risk. Even when DFEs are conducted at high quality, decision makers may fail to act on findings that are counterintuitive or politically threatening. Finally,

implementation conditions necessarily change when scaling up a program validated in an impact evaluation, and external validity therefore remains a concern for DFEs.

Despite these limitations, the imperative that drove the rise of KFEs – the need to learn what works in development and to let that information direct development strategy – is as critical today as ever. Most development organisations are still flying blind on social impact measurement, and the impact evaluation wave of the past two decades has not equipped many of them to fill this gap. DFEs represent a promising new approach in international development, offering decision makers a powerful, practical and incentive-aligned method of learning how to more effectively improve social outcomes.

## Appendix A: Qualitative interviews

The arguments in this paper are informed by a series of semi-structured qualitative interviews with impact evaluation users from across the development sector. The interviews were conducted with a targeted, non-representative sample of policymakers, practitioners, funders and researchers of varying levels of familiarity with impact evaluation. Respondents were asked to comment on current uses and limitations of impact evaluations and changes that could improve the development community's use of evidence. On average, each interview took about 1 hour.

To assess interview responses, we compiled the main themes and ideas arising from each interview in a spreadsheet. We broke each interview into distinct comments or ideas, tagged each comment under one or more thematic categories and wrote a brief summary (5–20 words) of each comment to facilitate review. Tentative thematic categories were prepared in advance of the interviews and then revised to better fit the responses we received. Responses were also categorised by contact role (funder, policymaker, practitioner or researcher). These categorisations were completed by a single reviewer.

We identified 172 distinct comments/ideas across 27 interviews, across the following thematic categories (comments were tagged under multiple categories).

| Comments relating to | Number of comments |
|---|---|
| Use of evaluations by funders | 50 |
| Knowledge-focused evaluations | 48 |
| Decision-focused evaluations | 38 |
| Incentives driving the design of impact evaluations | 36 |
| Incentives driving implementation and funding decisions | 31 |
| Usefulness of M&E to implementers and funders | 28 |
| How organisations consult existing evidence | 27 |
| Flexibility of implementers and funders to change direction | 18 |
| External validity of impact evaluations | 13 |
| Availability of evaluation service providers | 7 |

## Appendix B: Individuals interviewed

| | Name | Organisation | Role |
|---|---|---|---|
| 1. | Tom Adams | Acumen Fund | Director of impact |
| 2. | Bill Savedoff | Center for Global Development | Senior fellow |
| 3. | Andrea Guariso | Centre for Institutions and Economic Performance, KU Leuven | PhD candidate in economics |
| 4. | Steven Chapman | Children's Investment Fund Foundation | Director of evidence, measurement and evaluation |
| 5. | Simon Berry | ColaLife | Founder, CEO |
| 6. | Cormac Quinn | DfID | Evaluation and results adviser |
| 7. | Desiree Winges | Educate! | M&E manager |
| 8. | Pranav Kothari | Educational Initiatives | Vice president, large scale assessments and Mindspark platform |
| 9. | Adam Ross | Bill & Melinda Gates Foundation | Senior program officer |
| 10. | Molly Kinder | Global Innovation Fund | Co-founder |
| 11. | Satyam Vyas | Going to School | COO |
| 12. | Yi Wei | IDE Cambodia | WASH innovation manager |
| 13. | Duncan Rhind | IDE Zambia | Country director |
| 14. | Musa Kpaka | International Institute of Tropical Agriculture<br><br>Bill & Melinda Gates Foundation (formerly) | Project coordinator and M&E specialist<br><br>Associate program officer |
| 15. | Jodi Nelson | IRC<br><br>Bill & Melinda Gates Foundation (formerly) | Senior Vice President of Policy and Practice<br><br>Director of strategy, measurement and evaluation |

| | Name | Organisation | Role |
|---|---|---|---|
| 16. | Marc Shotland | J-PAL | Director of training and senior research manager |
| 17. | Pascalina Chanda-Kapata | Ministry of Health, Government of Zambia | Principal surveillance and research officer |
| 18. | Rajit Punhani | Ministry of Home Affairs, Government of India | Joint secretary |
| 19. | A Santhosh Mathew | Ministry of Rural Development, Government of India | Joint secretary |
| 20. | Eva Vivalt | New York University | Post-doctoral researcher |
| 21. | Andrew Youn | One Acre Fund | Co-founder, executive director |
| 22. | Matt Forti | One Acre Fund USA | Co-founder, managing director |
| 23. | Paul Sparks | Peripheral Vision International<br><br>BRAC Uganda (formerly) | Director of research<br><br>Program manager for the Research and Evaluation Unit |
| 24. | Rukmini Banerji | Pratham<br><br>ASER Centre | CEO<br><br>Director |
| 25. | Gulzar Natarajan | Prime Minister's Office, Government of India | Director |
| 26. | Sharath Jeevan | STIR | Founder, CEO |
| 27. | Markus Goldstein | World Bank | Lead economist, Africa region and Research Group |

## References

3ie. *Impact Evaluation Repository*, International Initiative for Impact Evaluation. Available from: <http://www.3ieimpact.org/en/evidence/impact-evaluations/impact-evaluation-repository/> [Accessed 6 March 2015].

AidData, 2015. *The marketplace of ideas for policy change: who do developing world leaders listen to and why?* Available from: <http://aiddata.org/sites/default/files/marketplaceofideas_fullreport.pdf>

Ashraf, N, Gordon, R and Shah, N, 2011. *Deworming Kenya: translating research into action*. Harvard Business School.

AusAID, 2011. 3ie and the funding of impact evaluations: a discussion paper for 3ie's members. Available from: <http://mande.co.uk/blog/wp-content/uploads/2011/12/Discussion-Paper-3ie-and-the-funding-of-impact-eval-FINAL.pdf> [Accessed 22 September 2015].

Baird, S, Hicks, JH, Kremer, M and Miguel, E, 2013. *Worms at work: public finance implications of a child health investment*, J-PAL. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.435.2686&rep=rep1&type=pdf> [Accessed 22 September 2015].

Banerjee, A, Duflo, E, Goldberg, N, Karlan, D, Osei, R, Parienté, W, Shapiro, J, Thuysbaert, B and Udry, C, 2015. A multifaceted program causes lasting progress for the very poor: evidence from six countries. *Science*, 348(6236).

Banerjee, A and Duflo, E, 2012. *Poor economics: a radical rethinking of the way to fight global poverty*. New York: PublicAffairs.

Barder, O, 2014. Is 'the Struggle' the baby or the bathwater? *Center for Global Development blog*. Available from: <http://www.cgdev.org/blog/struggle-baby-or-bathwater> [Accessed 25 March 2015].

Behrman, JR, Sengupta, P and Todd, P, 2005. Progressing through PROGRESA: an impact assessment of a school subsidy experiment in rural Mexico. *Economic Development and Cultural Change*, 54(1), pp.237–75.

Bold, T, Kimenyi, M, Mwabu, G, Ng'ang'a, A and Sandefur, J, 2013. Scaling up what works: experimental evidence on external validity in Kenyan education. *Working Paper 321*. Center for Global Development. Available from: <http://www.cgdev.org/sites/default/files/Sandefur-et-al-Scaling-Up-What-Works.pdf> [Accessed 22 September 2015].

Bonbright, D, 2012. *Use of impact evaluation results*, InterAction and the Rockefeller Foundation. Available from:

<http://www.interaction.org/sites/default/files/Use%20of%20Impact%20Evaluation%20Results%20-%20ENGLISH.pdf> [Accessed 22 September 2015].

Cameron, D, Mishra, A and Brown, A, 2015. The growth of impact evaluation for international development: how much have we learned? *Journal of Development Effectiveness*, pp.1–21.

Cottle, G. *Report on the policy community survey*, GlobeScan. IDRC Think Tank Initiative. Available from: <http://www.idrc.ca/EN/Documents/IDRC-Global-Report.pdf> [Accessed 22 September 2015].

Dhaliwal, I and Tulloch, C. *From research to policy, using evidence from impact evaluations to inform development policy*, J-PAL. Available from: <http://www.povertyactionlab.org/publication/research-policy> [Accessed 22 September 2015].

Evans, D and Popova, A. What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. World Bank Group. Available from: <http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2015/02/26/090224b082b5cbf1/1_0/Rendered/PDF/What0really0wo0n0systematic0reviews.pdf>.

Evidence Action. *What We Do*. Available from: <http://www.evidenceaction.org/#about> [Accessed 12 March 2015].

Ford Foundation, 2014. *Researchers highlight graduation research, say time is right to scale up*, CGAP Ford Foundation Graduation Program. Available from: <http://graduation.cgap.org/2014/02/21/researchers-highlight-graduation-research-say-time-is-right-to-scale-up/> [Accessed 12 March 2015].

Gertler, P, 2004. Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *American Economic Review*, 94(2), pp.336–41.

Gertler, P, Martinez, S, Premand, P, Rawlings, L and Vermeersch, C, 2011. *Impact Evaluation in Practice*. Washington, DC: World Bank.

GiveWell, 2014. A conversation with the Abdul Latif Jameel Poverty Action Lab and Evidence Action. Available from: <http://files.givewell.org/files/conversations/J-PAL%20EvAct%2012-22-2014%20(public).pdf> [Accessed 24 March 2015].

Hallsworth, M, Parker, S and Rutter, J, 2011. *Policy making in the real world: evidence and analysis*, Institute for Government. Available from: <http://www.instituteforgovernment.org.uk/sites/default/files/publications/Policy%20making%20in%20the%20real%20world.pdf> [Accessed 22 September 2015].

Henrich, J, 2000. Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review*, 90(4), pp.973–979.

IEG 2012. *World Bank Group impact evaluations: relevance and effectiveness*, Independent Evaluation Group. Available from: <http://ieg.worldbankgroup.org/Data/reports/impact_eval_report.pdf> [Accessed 22 September 2015].

IPA 2011, *Impact and scale-up: an interview with Annie Duflo and Dean Karlan, part 2*, Innovations for Poverty Action. Available from: <http://www.poverty-action.org/node/4926> [Accessed 12 March 2015].

IPA 2015, *Frequently asked question*s, Innovations for Poverty Action. Available from: <http://www.poverty-action.org/about/faqs> [Accessed 22 June 2015].

J-PAL 2011, *Incentives for immunization*, The Abdul Latif Jameel Poverty Action Lab. Available from: <http://www.povertyactionlab.org/publication/incentives-immunization> [Accessed 22 September 2015].

J-PAL 2012, *Deworming: a best buy for development*, The Abdul Latif Jameel Poverty Action Lab. Available from: <http://www.povertyactionlab.org/publication/deworming-best-buy-development> [Accessed 22 September 2015].

J-PAL 2015, *Scale-Ups*, The Abdul Latif Jameel Poverty Action Lab. Available from: <http://www.povertyactionlab.org/scale-ups> [Accessed 12 March 2015].

J-PAL and IPA 2015, *Where credit is due*, The Abdul Latif Jameel Poverty Action Lab. Available from: <http://www.povertyactionlab.org/publication/where-credit-is-due> [Accessed 22 September 2015].

Levine, R and Savedoff, W, 2015. *The Future of Aid: Building Knowledge Collectively*, Center for Global Development. Available from: <http://www.cgdev.org/publication/future-aid-building-knowledge-collectively> [Accessed 22 September 2015].

Luke, C, 2015. *Putting evidence into policymaking: RCTs as a tool for decision-making*, Devex. Available from: <https://www.devex.com/news/putting-evidence-into-policymaking-rcts-as-a-tool-for-decision-making-84952> [Accessed 24 March 2015].

Miguel, E and Kremer, M, 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), pp.159–217.

Pritchett, L, 2002. It pays to be ignorant: a simple political economy of rigorous program evaluation. *Journal of Policy Reform*, 5(4), pp.251–269.

Pritchett, L, 2012. Impact evaluation and political economy: what does the 'conditional' in 'conditional cash transfers' accomplish? *Center for Global Development blog*. Available from: <http://www.cgdev.org/blog/impact-evaluation-and-political-economy-what-does-'conditional'-'conditional-cash-transfers'> [Accessed 24 March 2015].

Pritchett, L and Sandefur, J, 2013. *Context matters for size: why external validity claims and development practice don't mix*, Center for Global Development. Available from: <http://www.cgdev.org/sites/default/files/context-matters-for-size_0.pdf> [Accessed 22 September 2015].

Ravallion, M, 2009. *Evaluation in the Practice of Development*. International Bank for Reconstruction and Development.

Seva Mandir. *Health*. Available from: <http://www.sevamandir.org/health> [Accessed 22 June 2015].

Székely, M, 2011. *Toward Results-Based Social Policy Design and Implementation*, Center for Global Development. Available from: <http://www.cgdev.org/content/publications/detail/1425010> [Accessed 22 September 2015].

Council of Economic Advisers, 2014. *Chapter 7: Evaluation as a tool for improving federal programs*. Economic Report of the President.

Vivalt, E, 2015. *How much can we generalize from impact evaluations?* New York University draft paper.

Whittle, D, 2013. *How feedback loops can improve aid (and maybe governance)*, Center for Global Development. Available from: <http://www.cgdev.org/sites/default/files/WhittleFeedbackessay_1.pdf> [Accessed 22 September 2015].

Woolcock, M, 2013. Using case studies to explore the external validity of 'complex' interventions. *Evaluation* 19, pp.229–248.

World Bank, 2015. *World development report 2015: mind, society, and behavior*. Washington, DC: World Bank.

World Bank. Knowledge in development note: impact evaluation. Available from: <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:22451297> [Accessed 12 March 2015].

## Publications in the 3ie Working Paper Series

The following papers are available from
http://www.3ieimpact.org/3ie_working_papers

*Evaluations with impact: decision-focused impact evaluation as a practical policymaking tool, 3ie Working Paper 24.* Shah, NB, Wang, P, Fraker, A and Gastfriend, D, (2015)

*What methods may be used in impact evaluations of humanitarian assistance?*, *3ie Working Paper 22*. Puri, J, Aladysheva, A, Iversen, V, Ghorpade, Y and Brück, T (2014)

*Impact evaluation of development programmes: Experiences from Viet Nam, 3ie Working Paper 21*. Nguyen Viet Cuong (2014)

*Quality education for all children? What works in education in developing countries, 3ie Working Paper 20.* Krishnaratne, S, White, H and Carpenter, E (2013)

*Promoting commitment to evaluate, 3ie Working Paper 19*. Székely, M (2013)

*Building on what works: commitment to evaluation (c2e) indicator, 3ie Working Paper 18*. Levine, CJ and Chapoy, C (2013)

*From impact evaluations to paradigm shift: A case study of the Buenos Aires Ciudadanía Porteña conditional cash transfer programme*, *3ie Working Paper 17*. Agosto, G, Nuñez, E, Citarroni, H, Briasco, I and Garcette, N (2013)

*Validating one of the world's largest conditional cash transfer programmes: A case study on how an impact evaluation of Brazil's Bolsa Família Programme helped silence its critics and improve policy*, *3ie Working Paper 16*. Langou, GD and Forteza, P (2012)

*Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework, 3ie Working Paper 15*. White, H and Phillips, D (2012)

*Behind the scenes: managing and conducting large scale impact evaluations in Colombia*, *3ie Working Paper 14*. Briceño, B, Cuesta, L and Attanasio, O (2011)

*Can we obtain the required rigour without randomisation? 3ie Working Paper 13*. Hughes, K and Hutchings, C (2011)

*Sound expectations: from impact evaluations to policy change, 3ie Working Paper 12*. Weyrauch, V and Langou, GD (2011)

*A can of worms? Implications of rigorous impact evaluations for development agencies, 3ie Working Paper 11.* Roetman, E (2011)

*Conducting influential impact evaluations in China: the experience of the Rural Education Action Project*, *3ie Working Paper 10*. Boswell, M, Rozelle, S, Zhang, L, Liu, C, Luo, R and Shi, Y (2011)

*An introduction to the use of randomized control trials to evaluate development interventions*, *3ie Working Paper 9*. White, H (2011)

*Institutionalisation of government evaluation: balancing trade-offs*, *3ie Working Paper 8.* Gaarder, M and Briceño, B (2010)

*Impact Evaluation and interventions to address climate change: a scoping study*, *3ie Working Paper 7*. Snilstveit, B and Prowse, M (2010)

*A checklist for the reporting of randomized control trials of social and economic policy interventions in developing countries*, *3ie Working Paper 6.* Bose, R (2010)
*Impact evaluation in the post-disaster setting, 3ie Working Paper 5*. Buttenheim, A (2009)

*Designing impact evaluations: different perspectives, contributions, 3ie Working Paper 4.* Chambers, R, Karlan, D, Ravallion, M and Rogers, P (2009) [Also available in Spanish, French and Chinese]

*Theory-based impact evaluation*, *3ie Working Paper 3*. White, H (2009) [Also available in French and Chinese.]

*Better evidence for a better world*, *3ie Working Paper 2.* Lipsey, MW (ed.) and Noonan, E (2009)

*Some reflections on current debates in impact evaluation*, *3ie Working Paper 1*. White, H (2009)

In this paper, impact evaluation firm IDinsight argues that, in order to more effectively inform development action, impact evaluations must be adapted to serve as context-specific tools for decision-making that feed into local solution-finding systems. Towards this end, a new kind of impact evaluation has recently emerged, one that prioritises the implementer's specific decision-making needs over potential contributions to a global body of knowledge. These 'decision-focused evaluations' are driven by implementer demand, tailored to implementer needs and constraints and embedded within implementer structures. By reframing the primary evaluation objective, they allow implementers to generate and use rigorous evidence more quickly, more affordably and more effectively than ever before.

The authors suggest strategies for involving all stakeholders, increasing demand for and supply of decision-focused evaluations, effectually using knowledge- and decision-focused evaluation methods and incorporating other systems and considerations to maximise the social impact of impact evaluations.

**Working Paper Series**

International Initiative for Impact Evaluation
202–203, Rectangle One
D-4, Saket District Centre
New Delhi – 110017
India

3ie@3ieimpact.org
Tel: +91 11 4989 4444

www.3ieimpact.org