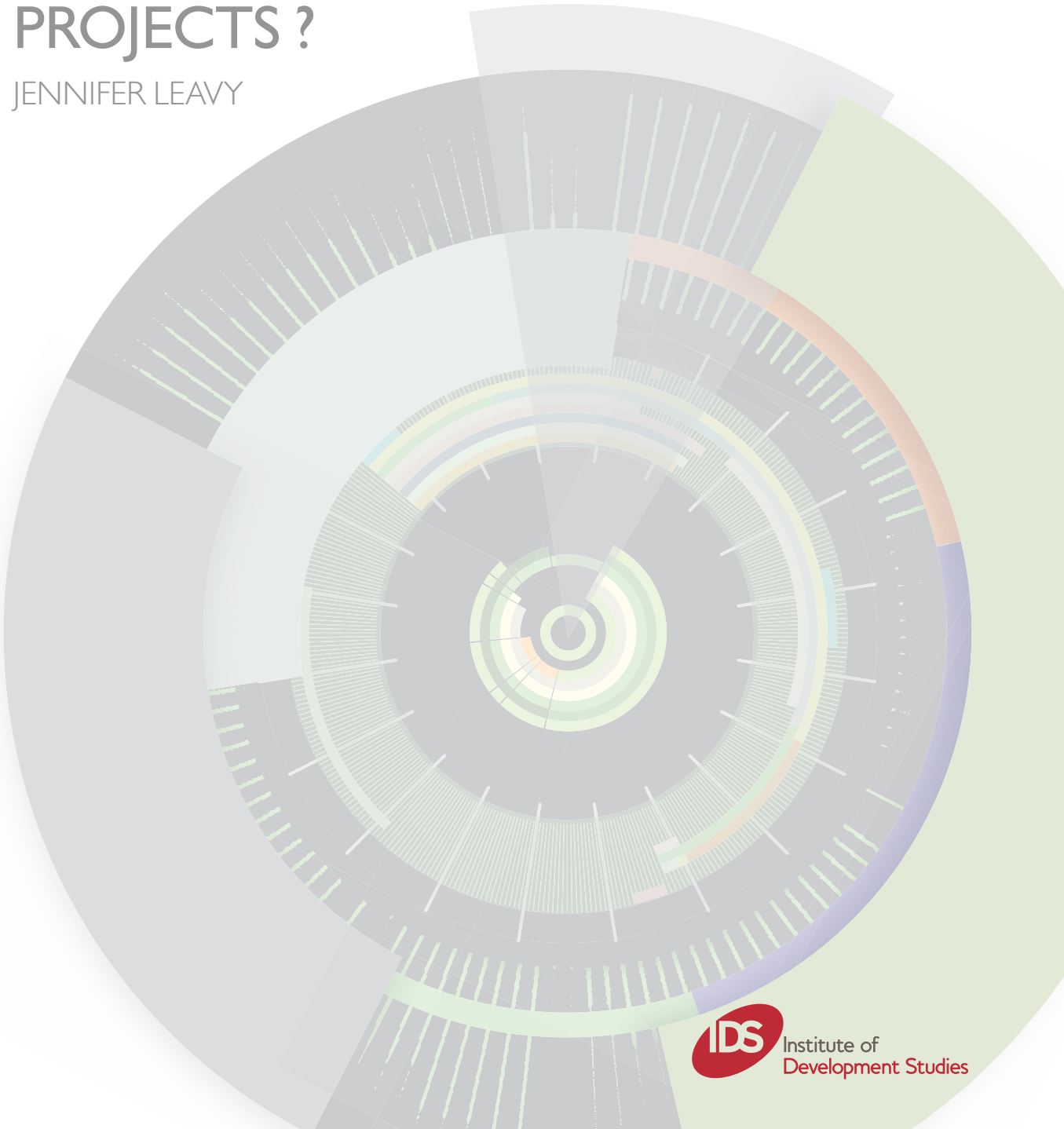
# MAKING ALL VOICES COUNT

## A GRAND CHALLENGE FOR DEVELOPMENT

# HOW USEFUL ARE RCTS IN EVALUATING TRANSPARENCY AND ACCOUNTABILITY PROJECTS ?

JENNIFER LEAVY



ISSUE 01 /// SEPTEMBER 2014
RESEARCH, EVIDENCE AND LEARNING
WORKING PAPER

IDS Institute of Development Studies

MAKING ALL VOICES COUNT

# Credits

**Author:** Dr Jennifer Leavy
*Dr Jennifer Leavy is a Senior Research Fellow in the School on International Development at the University of East Anglia. She is an economist whose research focuses on the intersection of social, cultural and economic life in the context of poverty, inequality and vulnerability. She is a mixed methods impact evaluation specialist.*
J.leavy@uea.ac.uk
School of International Development, University of East Anglia, Norwich, NR4 7TJ, United Kingdom.
https://www.uea.ac.uk/international-development/people/profile/j-leavy

# About Making All Voices Count

*Making All Voices Count* is a programme working towards a world in which open, effective and participatory governance is the norm and not the exception. We encourage locally driven and context specific change, as we believe a global vision can only be achieved if it is pursued from the bottom up, rather than the top down.

The field of technology for Open Government is relatively young and the consortium partners, Hivos, Institute of Development Studies (IDS) and Ushahidi, are a part of this rapidly developing domain. These institutions have extensive and complementary skills and experience in the field of citizen engagement, government accountability, private sector entrepreneurs, (technical) innovation and research.

*Making All Voices Count* is supported by the U.K Department for International Development (DFID), U.S. Agency for International Development (USAID), Swedish International Development Cooperation Agency, Open Society Foundations (OSF) and Omidyar Network (ON), and is implemented by a consortium consisting of Hivos (lead organisation), the Institute of Development Studies (IDS) and Ushahidi.

# Research, Evidence and Learning component

The Research, Evidence and Learning component's purpose is to contribute to improving performance and practice and build an evidence base in the field of citizen voice, government responsiveness, transparency and accountability (T&A) and Technology-for-T&A. It is managed by the Institute of Development Studies (IDS), a global organisation with over thirty years of research into governance and citizen participation.

This publication is licensed under a Creative Commons Attribution 3.0 Unported License. This means that you are free to share and copy the content provided The Institute of Development Studies and originating authors are acknowledged.

© Institute of Development Studies 2014

Disclaimer: This document has been produced with the financial support of the Omidyar Network, the Open Society Foundation, the Swedish International Development Cooperation Agency (SIDA), the UK Department for International Development (DFID), and the United States Agency for International Development (USAID). The views expressed in this publication do not necessarily reflect the official policies of our funders.

RESEARCH, EVIDENCE AND LEARNING
WORKING PAPER

# Contents



## Tables

# List of abbreviations

| | |
|---|---|
| CDC | Community Development Committee |
| CSO | civil society organisation |
| E&A | empowerment and accountability |
| FGD | focus group discussion |
| FOI | Freedom of Information |
| GSDRC | Governance and Social Development Resource Centre |
| ICT | information and communication technology |
| ICT4D | Information and Communication Technology for Development |
| IE | impact evaluation |
| INGO | international non-governmental organisation |
| NGO | non-governmental organisation |
| PETS | Public Expenditure Tracking Surveys |
| R&E | research and evidence |
| RCTs | randomised control trials |
| SR | systematic review |
| T&A | transparency and accountability |
| T4T&A | Technology-for-Transparency-and-Accountability |
| VDC | Village Development Committee |
| WATSAN | water and sanitation |

# 1 Introduction

This paper, produced under the research and evidence (R&E) component of the Making All Voices Count programme,[1] reviews experience in the use of randomised control trials (RCTs) in evaluating transparency and accountability (T&A) initiatives, and where evidence exists, in evaluating Technology-for-Transparency-and-Accountability (henceforth T4T&A, also known as Tech-4-T&A) initiatives.

To date, there have been relatively few impact evaluations (IEs) of T&A programmes, despite the amount of donor funding and attention given to the field. Evaluations tend either to be concentrated in certain sectors and countries, making it difficult to make generalisations, or are in very early stages themselves and therefore too soon for lessons to be drawn. For technology-based initiatives the pool of evaluations is even smaller, reflecting the relative youth of these kinds of interventions.

The paper aims to contribute to understanding the potential usefulness of RCT approaches within Making All Voices Count and similar aid and research programmes and grant mechanisms. The review explores: (i) where and under what conditions RCTs might be the most appropriate approach; and (ii) where other approaches would be more effective and more robust, given the particular characteristics of T&A programmes.

Within the Making All Voices Count programme the paper is expected to have three uses. First, among other activities, the programme supports research that helps to build the evidence base about T&A and T4T&A, and some of this research is impact assessment, for which RCT-type designs are among the options that the researchers and research organisations might consider. Second, the programme also funds T&A/T4T&A projects intended to scale up or scale out innovative approaches to promoting citizen voice and securing government responsiveness; these grantees and partners, too, might consider RCTs or quasi-experimental designs for their scaling efforts. And third, it is expected that the external Evaluation Management Unit, which is charged with evaluating Making All Voices Count throughout its four-year duration, might consider using RCTs or quasi-experimental methods to assess impact in specific initiatives supported by the programme. Therefore, the paper is intended to meet some internal programme-related needs and also answer questions that are relevant to the T&A field and the impact assessment field more broadly.

While the primary concern of Making All Voices Count is *technology* for T&A, this review considers evaluations with an RCT or quasi-experimental component of a range of technology and non-tech-based T&A interventions. It focuses largely on interventions and initiatives in service delivery, explicitly oriented towards monitoring and demanding performance accountability in services widely accepted to be entitlements: Public Expenditure Tracking Surveys (PETS); complaint mechanisms; report cards; scorecards; information dissemination; community monitoring; public hearings and social audits. It also considers RCT evaluations of T&A initiatives related to corruption and electoral reform where there is a clear social accountability link.

1 Making All Voices Count is a global initiative that supports innovation, scaling, and research to deepen existing innovations and help harness new technologies to enable citizen engagement and government responsiveness.

## 1.1 Objectives of this review

The research and evidence (R&E) component of Making All Voices Count, which commissioned this paper, has two main purposes:

- to contribute to improving performance and practice in the field of T&A and T4T&A – 'programme learning'; and
- to build an evidence base and theory in the general field of voice and accountability, and specifically to inform the incipient fields of practice of T4T&A and Open Government – 'evidence- and theory-building'.

In contributing to both of these, the R&E component also informs the strategies of donors operating in this field. It supports both secondary and primary research. In terms of methodological approaches, the component's guiding principle is 'methodological appropriateness' in respect of the research question(s) addressed in any given case, informed by contemporary debates and scholarship as to appropriateness.

The motivation for this review stems from these twin mandates of 'programme learning' and 'evidence- and theory-building'. Via its grant-making function, Making All Voices Count could potentially be funding evaluative work that might use experimental and quasi-experimental approaches. Some evaluation questions might be amenable to this method, but an important question to address here is whether, overall, this is the right way to get answers, balancing methodological appropriateness with practical concerns such as cost. Given the relatively low value of the innovation and scaling grants (ranging from just under £40,000 to £100,000), is there justification in using an RCT to evaluate one of these projects given the comparatively high cost of conducting an RCT well (a common minimum would be £250,000: Howard White, pers. comm. August 2014). For evaluators of T4T&A, what might be the opportunity costs given that these methods tend to answer relatively narrow types of questions and therefore may have crucial limitations when it comes to taking an analytical or exploratory approach? In addition, what ethical issues – especially those peculiar to the T4T&A field – need to be considered that may also preclude this kind of approach (for example, the need to identify a control group and 'exclude' it or hold it back from an intervention)?

Bearing these questions in mind, it is important to review experience with RCTs in order to be able to bring to bear critical reflections on their applicability and usefulness in the context of T4T&A on the decision of whether or not to go ahead with, or support, an RCT. This review therefore seeks to: (i) help Making All Voices Count (and others) to understand those circumstances where RCTs might be appropriate and feasible to assess T4T&A (even if this is within a small percentage of such interventions); and (ii) highlight designs where RCTs have been innovative or combined with other methods to answer impact questions around T&A/T4T&A. In this latter respect it is expected to be relevant to other transparency and accountability programmes as well as to the governance field as a whole.

Who is this paper for? The Making All Voices Count R&E component has three major stakeholder groups:

- Practitioners of T4T&A/T&A – both development practitioners working on citizen engagement and social accountability, and tech practitioners working on applications of technology for T&A-related purposes.
- Scholars of T4T&A/T&A – academics, policy researchers, impact assessors and evaluators involved in exploring what T&A consist of and how they are achieved, and in

producing scholarly outputs and policy guidance and recommendations in this field. These are currently mainly northern-based but a programme such as Making All Voices Count is well-positioned to help nurture these actors and communities in southern countries.

- Funders of T4T&A/T&A programmes and implementing agencies – official donors, INGOs, philanthropic foundations and private donors contributing funds to programming in this field, interested in knowing about impact and understanding what works so as to shape their future strategies and investments.

More broadly, practitioners and knowledge brokers involved in the fields of T&A, Information and Communication Technology for Development (ICT4D), T4T&A, media, cultural and communications studies are potential interlocutors for the Making All Voices Count R&E component and likely audiences for its R&E outputs.

The paper attempts to address the overarching question:

> **How and in what ways might RCTs and other experimental methods be appropriate in the context of evaluating T4T&A programmes?**

In order to get at this, five sub-questions help to guide the discussion and subsequent analysis:

S1 What should impact evaluations be measuring in the context of T&A?
S2 What do we know about RCTs (and other experimental approaches) in evaluating T&A initiatives?
S3 How do RCTs measure [which elements of] the impact of T&A programmes?
S4 What can RCTs tell us about the effectiveness/success of T&A initiatives?
S5 What are the most appropriate ways to measure [which elements of] impact in T&A?

The discussion is structured as follows. Section 2 considers the evaluation 'problem'– what is impact evaluation? And specifically, what are RCTs and the conditions under which they can be carried out? Section 3 sets out definitions of transparency and accountability before describing key characteristics of T&A initiatives and the measurement of T&A. The current state of play in RCTs in evaluating T&A initiatives is described in Section 4, setting out what we have learned so far in the use of these methods in this field and implications for evaluation design. Section 5 analyses the effectiveness of RCTs in measuring the impact of T&A programmes via a review of 15 RCT and/or experimental evaluations of T&A initiatives including those available in T4T&A. Section 6 concludes.

# 2 Impact evaluation and RCTs

## 2.1 Impact evaluation definitions

What do we mean by 'impact evaluation' (IE)? Different definitions abound in the literature and are hotly contested. Defining impact evaluation is important as different definitions emphasise differing aspects of impact, including differing beliefs about what produces impact (causality) and how evaluations should be designed in order to measure this (for a summary of the debate see Stern *et al.* 2012).

The prominent definitions in international development are:

> 'Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended' (OECD-DAC Glossary[2]).
>
> 'The primary purpose of impact evaluation is to determine whether a program has an impact (on a few key outcomes), and more specifically, to quantify how large that impact is' (Abdul Latif Jameel Poverty Action Lab[3]).
>
> '… assessing changes in the well-being of individuals, households, communities or firms that can be attributed to a particular project, programme or policy' (World Bank, Poverty Reduction & Equity website[4]).
>
> 'Rigorous impact evaluation studies are analyses that measure the net change in outcomes for a particular group of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question that is being investigated and to the specific context' (3ie in its 'foundation document'[5]).

The 3ie and Abdul Latif Jameel Poverty Action Lab definitions imply only a limited set of (experimental) methods – all of which imply a counterfactual framework of causal influence and a focus on attribution.

While IE definitions and methodologies are hotly contested, and the Making All Voices Count R&E component recognises methodological pluralism via 'appropriateness' in research and evaluation, this paper focuses on one method (RCTs) which tends to align with the latter two definitions of impact, and considers their applicability to T&A and T4T&A.

### 2.1.1 Causality and the counterfactual

Demonstrating causal links and explaining how these links work lie at the heart of IE. In basic terms, impact evaluation assesses how an intervention affects outcomes via causal links (no matter whether these effects are intended or unintended). In effect we are assessing the changes in wellbeing or welfare of individuals, households, communities, firms that can be

2 See http://www2.unescobkk.org/elib/UNESCO-MI-Course-Material/Session-16/Ref%2016.1.%20OECD_M&E_Glossory.pdf (accessed 27 August 2014).

3 See www.povertyactionlab.org/methodology/what-evaluation/impact-evaluation (accessed 27 August 2014).

4 See http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,menuPK:384339~pagePK:162100~piPK:159310~theSitePK:384329,00.html (accessed 27 August 2014).

5 See www.cgdev.org/doc/eval%20gap/3IEfound_doc.pdf (accessed 27 August 2014).

*attributed* to a particular programme or policy. The way we anticipate and measure these causal links and determine attribution in turn determines the evaluation design, which in turn determines the methods used to measure impact. In many cases an intervention is but one causal factor at play – it can be thought of as being a contributory cause insofar as it is a necessary and sufficient condition for the impact in question *as part of a package of other causal or 'supporting' factors* (termed an INUS cause: 'an Insufficient but Necessary part of a Condition that is itself Unnecessary but Sufficient for the occurrence of the effect' (Stern *et al.* 2012: 41; see discussion in Mayne 2012).

In many cases evaluation design hinges on the question: *What would have happened in the absence of the intervention?* Looking at what would have happened to those receiving the intervention (or 'treatment') if they had *not* received it, is a comparison of scenarios 'with' and 'without' the intervention. However, we cannot actually observe what would have happened to those receiving the treatment had they not received it after all – the 'without' scenario. Between two points in time people's outcomes might have changed anyway due to other factors that may not be attributable to the programme, so we need a way to isolate the change in outcomes that we can directly attribute to the programme. To achieve this we need to construct a plausible counterfactual – define a hypothetical situation that would occur in the absence of a programme and to measure the welfare levels of individuals or other identifiable units that correspond with this hypothetical situation (the 'control' group) – and do what we can to minimise any bias. That is, we would be aiming to attribute changes to a programme or policy while removing confounding factors. This allows us to estimate the magnitude of effects with clear causation. This causal analysis is essential for understanding the relative role of alternative interventions in achieving the desired policy outcomes.

In reality it is questionable whether the assumptions underlying counterfactual thinking hold. Indeed, 'finding an identical match for the factual world, e.g. the world where the cause and the effect have been observed, may be difficult or impossible' (Stern *et al.* 2012: 7). Counterfactual analysis tends only to be feasible where eligibility for an intervention varies along criteria that do not also determine outcomes, and therefore equivalent groups are excluded automatically (for example, so-called 'natural experiments').

## 2.2 Strengths and conditions of RCTs

RCTs are described by some to be the most robust way of measuring impact (see, for example, CGD 2006). Random allocation of an intervention to treatment groups removes selection bias. It controls for bias caused by confounding factors (that is, alternative causes are not confounded with the treatment) by making sure these have equal chance of occurring in both treatment and control groups. These groups therefore do not differ systematically prior to the intervention and so any differences that arise subsequently can be attributed to the intervention. In clinical trials originating in the medical sector and in the natural sciences, RCTs are seen to be the 'gold standard' in establishing a plausible, unbiased counterfactual against which to measure the impact of an intervention in that they 'consistently produce the most accurate results' (J-PAL 2014[6]). This 'gold standard' language has carried through to impact evaluations in other spheres, including development policy and programming (see, for example, Abdul Latif Jameel Poverty Action Lab J-PAL www.povertyactionlab.org/ and the discussion in Barahona 2010 and Camfield and Duvendack 2014).

6 Abdul Latif Jameel Poverty Action Lab (J-PAL), www.povertyactionlab.org/methodology/what-randomization (accessed 27 August 2014).

However, RCTs work best under somewhat stringent conditions that are not very easily met. These conditions include:

- there is one primary cause and one primary effect;
- a 'control' group can be identified without contamination, and comparison groups can be controlled;
- they examine the success of a specific intervention in a particular setting, rather than in wider generalisation;
- sample size is large enough to support statistical analysis;
- the focus is on causal rather than explanatory analysis (Stern *et al.* 2012: 38–9).

Further, critical voices have been arguing that there are threats to both the internal and external validity of RCTs. (These are discussed later. See, for example, Harrison 2013; Barrett and Carter 2010; Deaton 2010). There is therefore clear and growing recognition that RCTs and other experimental methods are not necessarily appropriate in all circumstances. This means other designs and methods, including those based on qualitative and participatory approaches, such as theory and case-based evaluations, are (re)gaining credibility in the impact evaluation toolkit.[7]

# 3 T&A initiatives

This section considers definitions of transparency and accountability, the key characteristics of T&A initiatives, and measurement of T&A in order to identify important considerations in designing and carrying out impact evaluations of T&A so that we can analyse whether and under what conditions RCTs work best in this field.

## 3.1 What are 'transparency' and 'accountability'?

'Transparency' specifically refers to governments and other organisations enabling 'the clear disclosure of information rules, plans, processes and actions' (Transparency International 2009: 44). This is achieved via well-defined processes and procedures to enable easy access to information for citizens.

Defining 'accountability' tends to be less straightforward. Definitions include the notion of actors (duty-bearers) being held responsible for their actions and behaving according to a set of standards (see, for example, Tisné 2010). The standards need to be enforceable.

Transparency initiatives tend to centre on placing formerly inaccessible information or processes in the public domain, in order for citizen groups, providers or policymakers to access and use this information. Accountability initiatives centre on the strengthening of the relationship between 'power holder (account provider) and delegator (account demander)' with mechanisms to foster standard-setting, information provision, judging and sanctioning 'unsatisfactory' behaviour and performance (Joshi 2011: 2). The term 'social accountability' is often used to distinguish accountability actions taken in the social domain from those taken in the political-electoral domain, and those initiated by citizens *claiming* accountability from those led by governments *giving* accountability.

7 See, for example, European Evaluation Society Statement (2007): 'The importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and in development interventions', www.europeanevaluation.org/news?newsId=1969406 (accessed 27 August 2014).

Conventional wisdom is that transparency creates accountability: in order for there to be accountability between state and citizens there must first be transparent government. However, in a review of empirical evidence of the apparent link between transparency and accountability, Fox (2007) concludes that transparency does not necessarily lead to accountability. Rather, we should be asking *under what conditions* does transparency lead to accountability? He refines this further into 'what types of transparency manage to generate what types of accountability?' (Fox 2007: 665). In framing the questions like this the differences in interpretation of the terms, and therefore different conceptual boundaries, by different stakeholders, can be accommodated: 'While critics call for accountability, powerful elites respond by offering some measure of transparency instead' (Fox 2007: 664). Related to this is the observation that many approaches to voice and accountability assume that merely providing information 'will inspire collective action with sufficient power to influence public sector performance'. However, in order to be effective, initiatives need to foster what Fox later termed both 'Voice' and 'Teeth':

> 'Voice' refers here to both the aggregation and representation of the views of under-represented citizens. Many need to exercise voice (aggregation) and they also need to have the capacity to dialogue & negotiate with authorities (representation).
> 'Teeth' refers to government capacity to respond to voice – which includes both positive incentives and negative sanctions to reform the public sector. That is: Can authorities deliver? (Fox 2014: 13).

The causal pathways between transparency (defined as the accessibility of information), accountability, and ultimately improvements in the quality of governance are still not well understood (Bellver and Kaufmann 2005), but it can be held that transparency is 'a necessary but insufficient condition for accountability, and does not automatically generate it' (McGee and Gaventa 2011a: 13–14). Transparency and accountability are not synonymous, and neither transparency nor 'openness' automatically leads to accountability. Indeed, Fox suggests that the challenge in achieving impact in voice and accountability work rests on 'How to trigger virtuous circles, in which enabling environments embolden citizens to exercise voice, which in turn can trigger and empower reforms, which can then encourage more voice?' (Fox 2014: 13).

Defining transparency and accountability and identifying the paths that lead from the former to the latter (including potential feedback loops) are crucial in enabling evaluation of the impacts of T&A initiatives. Gaps in our understanding of these causal links have implications for evaluation design. The way T&A initiatives are designed and the overall characteristics of programmes are based on various underlying assumptions about the causal chains, further highlighting the importance of elucidating clearly the underlying theory of change.

## 3.2 Characteristics of T&A initiatives

Having defined transparency and accountability we now turn to characteristics of T&A initiatives. These initiatives have largely emerged in response to the perception that traditional political and bureaucratic[8] forms of accountability have been inadequate. These 'demand side', citizen-led initiatives present a number of mechanisms aside from elections and other bureaucratic procedures by which states can be held to account by citizens and other social actors (Peruzzotti and Smulovitz 2006; Joshi 2008). Initiatives span a number of different subfields, which overlap in approach: service delivery (a strong focus of this review);

8 Also referred to as state-side, supply-side or institutional.

freedom of information; budget accountability and transparency (also covered here in technology initiatives for T&A related to corruption/election monitoring); and public finance management more broadly. Service delivery mechanisms considered in this review encompass Public Expenditure Tracking surveys (PETS), complaint mechanisms, report cards, scorecards, information dissemination, community monitoring, public hearings and social audits. It is in T&A programmes within the area of service delivery that most IE has been conducted, where failures in service delivery are essentially failures in accountability relationships (see Joshi 2011). Different initiatives tend to play on and/or tap into providers' main motivators, and therefore deliver different things. For example, report cards or scorecards may be used where rankings are important to providers. Community monitoring focuses on creating watchdog roles for the citizens. PETS work on the underlying assumption that fear of exposure improves provider behaviour.

The range of initiatives that fall under the umbrella of T&A share a number of common elements, both technical elements and issues of power and politics. These have been explored extensively in various reviews, including McGee and Gaventa (2011a: 27) and Joshi (2011). To summarise here, T&A initiatives tend to:

- Include measures for improving transparency and information access designed to hold to account the state and its agents (for example, private-sector service providers);
- Engage citizens with more powerful actors, including state and private sector entities contracted by the state;
- Engage across social 'interfaces' as opposed to taking political, institutional or bureaucratic routes, although they may activate political and institutional mechanisms, for example internal government audits (see Claasen and Alpín-Lardiés 2010; Houtzager, Joshi and Gurza Lavalle 2008; McNeil and Malena 2010);
- Unfold within and seek to change context-specific social and political processes that tend to be complex and non-linear;
- Be based on theories of change that are often 'implicit or invisible' posing challenges in assessing impact and effecting learning (McGee and Gaventa 2011a: 27).

### 3.2.1 Technology for T&A

Recently, across all the T&A subfields there has been what is described as a wave of initiatives that use a range of information and communication technologies (ICTs) as tools for achieving T&A aims (ICT4T), including internet, global positioning systems, social media and mobile phones (McGee and Gaventa 2011a: 8; McGee and Carlitz 2013; Avila *et al.* 2010).

While T4T&A initiatives share some characteristics of T&A inititaives that rely on mechanisms that are not technology-based or technology-driven, a review by McGee and Carlitz (2013) suggests that the theories of changes underlying these initiatives 'rely even more heavily than those of non-tech-enabled T&A initiatives [TAIs] on assumptions about users and uptake' (McGee and Carlitz 2013: 8). This has implications for both programme impact and in evaluating impact, in that these implicit assumptions tend not to be well articulated in programme impact pathways and therefore do not tend to be systematically investigated, even though they are crucial to the programme achieving expected outcomes and impacts (McGee and Carlitz 2013: 15). This particularly relates to the tendency to neglect consideration of whether and how ordinary people currently use technology, and what the barriers to uptake and action might be.

## 3.3 Measuring (the impact of) T&A

What should impact evaluations be measuring in the context of T&A? And how do we know if an evaluation is any good? This section first explores important overarching issues in measuring impact in T&A programmes, exploring what T&A programmes are intending to achieve and from here what needs to be taken into account in identifying indicators and selecting data collection and measurement tools. This is important in establishing an analytical framework for considering experience to date in using RCTs to evaluate T4T&A initiatives, couched in an overall consideration of the whole suite of approaches available to conduct 'rigorous' impact evaluation that explicitly encompasses the specific characteristics of the T&A or governance context.

We are not concerned here with setting out prescribed indicators for measuring T&A – these will be specific to the initiative in question.[9] This important and complex debate has been well covered elsewhere.[10] Rather, we are concerned with elucidating the overarching steps and underlying mechanisms needed in order to be able to map out outcome and impact indicators and thus identify the types of data needed to capture these indicators, the methods needed to gather these data and thus answer evaluation questions of interest. For 'the realities of unaccountable governance, unproven accountability programming and uncertain evidence of impact' (McGee and Gaventa 2011a: 3) pose serious challenges in evaluating the impact of these programmes.

In their review, McGee and Gaventa (2011a) set out key explanatory variables on the state (or supply) and citizen (or demand) sides that determine the success, or otherwise, of T&A initiatives, not forgetting the importance of synergies between the two (McGee and Gaventa 2011a: 21, 2011b; Joshi 2011; Carlitz 2011; Calland 2011; Mejía Acosta 2011). It is helpful to set these out here:

State- or supply-side factors:

- Level of democratisation;
- Level of political will: political will and a political environment that favours a balanced supply- and demand-side approach to accountability;
- Enabling legal frameworks, political incentives and sanctions.

Citizen- or demand-side factors:

- Capabilities of citizens and civil society organisations (CSOs) to take up opportunities offered by T&A;
- The degree to which TAIs form part of multi-stranded and collective strategies;
- Engagement of citizens in the 'upstream' as well as the 'downstream' stages of governance and policy processes.

These factors emphasise further the crucial importance of considering the way citizen-led T&A initiatives play out in different political contexts and enabling environments, against the

9 This also relates to the importance of context in T&A initiatives that underpins arguments against developing an evaluation model to be applied across all T&A initiatives, see O'Neil, Foresti and Hudson (2007) and McGee and Gaventa (2011b).

10 Key papers on governance measurement include: Hawken and Munck (2009) on measuring corruption; Holland and Thirkell (2009) on measuring change in V&A; Ibrahim and Alkire (2007) on agency and empowerment indicators; Amin, Das and Goldstein (2008) for service delivery measurement tools; Alsop and Heinsohn (2005) on measuring empowerment; and Jupp *et al.* (2010) for an innovative methodology using qualitative self-assessment to measure empowerment as a key outcome of rights-based approaches to development.

backdrop of existing power relations, and the nature of the social contract between state and citizens, in determining positive impacts on state accountability (McGee and Gaventa 2011a: 20). The importance of power and politics as key components of the context of T&A initiatives has implications for what impact means in relation to these initiatives.

This brings us back once again to the importance and necessity of a well-articulated theory of change, in the sense of having a clear map of causal pathways ('programme logic') underlying an initiative. A clearly articulated theory of change can be key for both enhancing the effectiveness of an intervention by establishing a clear direction and focus, and also enabling impact assessment (monitoring and evaluation, progress tracking). Many T&A initiatives, however, lack a clear elucidation of the outcomes and impacts being sought, or of the assumptions that underlie the causal links between actions and inputs to outcomes and impact. This has been particularly highlighted as a problem in service delivery (see Joshi 2011; McGee and Gaventa 2011a: 15). A lack of theory of change 'can make it difficult to analyse retrospectively the existence or nature of connections between the *ex post* situation and the inputs made by the intervention, and thus reduce the possibility of learning… this means that even the effective implementation of the initiative may be hard to demonstrate, and that it will be harder still to demonstrate links between it and any apparent impact' (McGee and Gaventa 2011a: 27). See also the discussions in Vogel (2012), Guijt (2007) and Taylor *et al.* (2006) on the importance of theories of change.

# 4 RCT evaluation of T&A initiatives

This section summarises recent reviews and critiques of RCT evaluations of T&A initiatives, before turning to what the implications are for impact evaluation design in T&A.

## 4.1 What do we already know?

Randomised control trials have been used increasingly in the evaluation of specific interventions in governance programmes where outcomes are well-defined (Duflo, Hanna and Ryan 2008; Björkman and Svensson 2009; Banerjee *et al.* 2010; Olken 2007; Pandey, Goyal and Sundararaman 2009), highlighting once again that overcoming measurement challenges of governance and empowerment are key in determining the suitability of this approach. As acknowledged early in this paper, the methodology has certain advantages. However, a recent (2011) Governance and Social Development Resource Centre (GSDRC) review of the extent to which RCTs have been used to measure specifically impact of empowerment and accountability (E&A) processes and programmes suggests a tension within the academic literature regarding the suitability of RCTs for measuring impact of these interventions. Some see the potential for RCTs in enabling us to draw universal and generalisable lessons (for example, Moehler 2010: 42); others question their appropriateness: 'the interventions that are assessed tend to be quite narrow and the results have to be supplemented by qualitative work that can unearth processes through which the impacts are actually achieved' (Joshi 2011: 13). Key criticisms in relation to empowerment and accountability contexts identified in the GSDRC review include:

- RCTs do not tell us why or how an E&A initiative works and so 'in many cases, makes the experimental approach [of which the RCT approach is one kind] less useful for policy design or broader application' [than other approaches might be] (Bardhan 2011 cited in GSDRC 2011);

- The need to identify a plausible counterfactual, with interventions modelled as delivering a discrete and homogenous output, means that while an RCT can effectively measure short-term results with short causal chains, such approaches are less amenable to assessing complex interventions where many factors bring about change. This is highly likely to be the case with transparency and accountability programmes (Jones 2009);
- In the context of using RCTs in community-driven development programmes, electoral reform and corruption, results from RCTs are likely to be 'inconsistent and erratic' – partly related to variations in interventions and populations with every trial (Blattman 2011[11]).

Relevant too are the ethical questions surrounding the use of RCTs in development more generally, in 'real-world, field applications' (see the discussion in Barrett and Carter 2010 on RCTs in development economics). These include: (i) the risk of 'doing harm' by withholding an intervention from a group of people (although in the case of stepped-wedge designs, whereby the control group is brought into the programme at the end, means this becomes less of an issue) or trialling something that is less effective than current best practice; (ii) wasting resources on 'treating' individuals known not to be in need of the intervention based on local knowledge, in the interests of preserving randomisation; (iii) the impact on the wellbeing and, importantly, behaviour of people in 'unblinded' trials – for example, distress caused by awareness of being in a control group and excluded from an intervention (Barrett and Carter 2010).

The studies consulted for the GSDRC review suggest mixed results from RCTs in measuring impact of governance programmes. In addition to the ethical concerns outlined in the previous paragraph, the main reasons why RCTs might not be the most appropriate methods in evaluating governance (encompassing transparency and accountability) programmes are summarised below (drawing in part on GSDRC 2011: 3):

In relation to governance programmes specifically:

- Scale: governance programmes tend to be 'small-n' rather than the 'large-n' needed for an RCT ('large-n' referring to cases where the sample size is large enough to make statistical inference);
- Context: importance of context and problems with external validity in scaling up T&A interventions;
- Complexity of governance initiatives and the tendency for them to be designed as a bundle of activities. This is often driven by important policy and political imperatives. A development agency often cannot just implement an intervention that is simple to measure in order to make it more amenable to randomised evaluation; it has to satisfy a bureaucracy and/or a hierarchy of other often complex institutional pressures;
- Time frame: governance processes are generally long term;
- Interdependency: there tends to be coordination between governance interventions making it difficult to isolate the effects of a single programme;
- Power issues are important. Often governance and T&A interventions are fundamentally about power shifts, following a non-linear, often longer term, somewhat unpredictable and harder to measure process.

11 http://chrisblattman.com/2011/03/24/behavioral-economics-and-randomized-trials-trumpeted-attacked-and-parried/ (accessed 27 August 2014).

Other criticisms of RCTs are not at all limited to governance or T&A sectors:

- Selection issues: the programmes put forward for evaluation tend to be the 'successful' programmes, i.e. those that (are expected to) have had good results;
- Theoretical gap: reasons for the success or failure of an intervention are often not gleaned by the evaluation, limiting its usefulness;
- Bias caused by the way treatment is (perceived to have been) assigned – for example, random compared to targeted;
- Proportionality: RCTs are relatively cost-intensive to do well, so the RCT needs to be proportional to the intervention, the questions, and the quality/certainty of evidence needed for a decision (for example, is it a pilot that has broad policy potential?);
- In addition, one practical consideration is reluctance of many NGOs and other implementing agencies to roll out RCTs as they generally involve considerable 'interference' in their activities, especially in implementation.

It could be argued that for some of these factors where RCTs apparently fall down, it is not the RCT approach *per se* that is the problem but, returning to our earlier discussion, the lack of well-elaborated theories of change, for example where so-called 'complexity' is perceived to be an issue. However, in cases where the theory of change is weak, this is a fact that should not be ignored or glossed over in taking a decision about which methods to use for IE.

## 4.2 Implications for evaluation design

Key features identified in the previous section relating specifically to governance and T&A attributes cover a broad range of factors that would suggest RCT conditions are more difficult to establish in the T&A context. They apply also to T4T&A. These are: *scale, context, complexity, time frame, interdependency and power*, and have implications for evaluation design. Take, for example, the fact that programmes differ in the numbers and kinds of interventions being implemented, and these may be independent of each other or interdependent. There may also be single or multiple intended outcomes or impacts. For this reason, approaches may not only be used singly but also combined in hybrid impact evaluation designs, whereby a combination of different approaches, 'wrapped' around a range of methods, may be used to address multi-layered evaluation questions.

These considerations notwithstanding, it is possible to indicate broadly in which circumstances particular designs would be most appropriate.

First, what questions can we answer via counterfactual analysis, and do these correspond to the kinds of impact questions we are interested in? The extent to which a change in outcome can be attributed to an initiative might call for an experimental or quasi-experimental design in order to isolate impacts due to the programme from confounding factors. However, counterfactual analysis has key limitations: while this approach allows us to answer causal questions related to whether or not a particular intervention has made a difference, or worked, in a particular circumstance ('internal validity'), generalisation to other contexts and times ('external validity') tends to be weak (see Ravallion 2009; Cartwright and Munro 2010). This means it is very difficult to answer questions about how an intervention could be improved (Jonas *et al.* 2009). More detailed knowledge about the underlying causal mechanisms is necessary in order to uncover why and how particular cause(s) lead to effects (Stern *et al.* 2012: 8). However, RCTs and other experimental approaches, unless combined with theory-based approaches that help to unpack the causal chain, are limited to the

narrowly defined underlying questions or hypotheses around which cause has led to which effect and for whom.

So we need other designs in order to uncover the answers to important impact questions such as how an impact came about, why it happened in this way, what else happened. These can be thought of as 'the complex connections between variables, the social and political dynamics and transmission belts by which impact is being attained and how this impact – political in nature – could be enhanced' (McGee and Gaventa 2011a: 27).

Theory-based evaluation may overcome some of the shortcomings of experimental and statistical approaches in relation to contextualisation, in that it enables us to map out causal pathways between implementation and outcomes. Theory-based approaches allow the nature of linkages between a complex set of causes and effects to be analysed in detail, affording an examination of whether and how objectives are being met and outcomes are achieved. It also enables identification and analysis of unexpected outcomes, both failures and successes (Rogers 2009; White 2009; Coryn *et al.* 2011; Pawson 2006, 2007; White and Phillips 2012).

For relatively new areas such as T4T&A, theory-based approaches may also help to build up new knowledge where there is little existing prior knowledge, especially where programmes are complicated insofar as the interventions tend to span disciplines with multiple agencies and cross-jurisdictions (Stern *et al.* 2012). These approaches can contribute towards unpacking the multiple causal strands often necessary in producing the impacts (Rogers 2009: 219, Table 1). These may happen through:

- Multiple sequential interventions;
- Multiple simultaneous interventions; or
- Multiple levels of intervention; or
- Different causal mechanisms operating in different contexts.

A key implication here is that it is important therefore when evaluating (T4)T&A initiatives to ask *what works for whom in what situations* (Rogers 2009: 219).

Non-linearity between cause and effect, including feedback loops, as well as emergent outcomes means initiatives may also be considered to be 'complex'.[12] In acknowledging this lack of linearity in accountability relationships, impact evaluations need to go beyond the question of how much of a desired impact was achieved and consider also what happened as the result of an initiative, why, and what it means. An important question given the fundamental importance of politics and power relations is to what extent and in what ways did power relations need to change in order for the programme to have impact. This poses a challenge not only for impact evaluation, especially in identifying relevant indicators, but for programme design itself – feeding into establishing a baseline of what these power relations looked like at the start of a programme. Unless impact evaluation addresses this it will 'continue asking the wrong questions and getting partial or wrong answers' (McGee and Gaventa 2011a: 31).

12 Following Rogers (2009), 'complex' in this context refers to 'appropriately dynamic and emergent aspects of interventions, which are adaptive and responsive to emerging needs and opportunities' (Rogers 2009: 218).

# 5 How effective are RCTs in measuring the impact of T&A programmes?

This section sets out an analytical framework for assessing how appropriate and effective RCTs are in evaluating the impact of T&A and T4T&A programmes in service delivery and social accountability. It outlines the search methods used to identify relevant studies to be included in this review, describes the identified studies and sets out the analysis guided by the framework set out in Section 5.1.

## 5.1 Analytical framework for assessing RCTs in IE of T&A initiatives

The analytical framework consists of four main groupings of questions under the headings of Design, Contribution, Explanation and Effects.[13] These questions were used to interrogate the studies.

Overarching question:

> **Does the evaluation plan use methods of analysis which are appropriate to the purpose of the impact assessment, taking into account its audience, the level of complexity involved, and positionality of those doing the study?**

### Design – general

- What evaluation questions are posed?
- What is the intended impact of the programme/intervention?
- Is the evaluation design 'randomisable'?
- Is there a clear counterfactual/mapping of causal influence?
- Does the counterfactual hold?
- Are ethical considerations clearly elucidated/taken into account?

### Contribution

- Does the evaluation design identify multiple causal factors (where relevant)?
- Does the design take into account whether the causal factors are independent or interdependent (where relevant)?
- Can the design analyse effects of contingent, adjacent and cross-cutting interventions (where relevant)?
- Are issues of 'necessity', 'sufficiency' and probability discussed?
- Is the choice of evaluation method informed by its success elsewhere?
- Does the evaluation design consider issues of timing, sequencing and durability?

### Explanation

- Is it clear how causal claims will be arrived at?
- Is the design able to answer 'how' and 'why' questions?

13 These questions draw on the probes devised by McGee and Gaventa (2011b: 48) for evaluating, designing or implementing T&A initiatives, combined with the set of generic questions set out in Stern *et al.* (2012) for judging the quality of IE design and methods (relating to validity and rigour), and some more general questions on evaluation design.

- Is a clear theory of change articulated? How is this derived?
- Is theory used to support explanation?
- Are alternative explanations considered and systematically eliminated?
- Does design take into account complex, contextual factors?

### Effects

- Does the evaluation design include methods for tracking change over time, including reference to a clear baseline?
- Are long-term effects identified?
- Are these effects related to intermediate effects and implementation trajectories?
- Is the question 'impact for whom' addressed in the design?
- Does the IE find evidence of impact?

These questions were used to guide the analysis of how effective RCTs are in measuring (which elements of) the impact of T&A programmes.

## 5.2 Search methods

A total of 15 evaluation studies of T&A initiatives in service delivery using an RCT impact evaluation design were found and each assessed against the framework set out above. This section describes the search strategy for finding the studies before analysing and discussing them in relation to the main questions or probes set out Section 5.1 under the four key domains of Design, Contribution, Explanation and Effects.

While not being a full systematic review (SR) the approach taken is to follow some of the tenets of an SR, by defining relevant search terms clearly, using a range of search strategies, described below, and reviewing systematically the most relevant literature.

The main search strategy combined database and web searches for evaluations of social accountability tools in service delivery, using key primary and secondary search terms and qualifiers, with hand-searches of bibliographies of key T&A studies and impact evaluation websites. Search terms are given in Table 1.

- Bibliographic databases via Metalib, including web of science/web of knowledge databases. Websites searched include 3ie, e-gap, J-PAL.
- Bibliographies consulted include: GSDRC Helpdesk Research Reports (GSDRC 2011, 2010); Joshi (2011), McGee and Gaventa (2011a, 2011b) and snowball searches of the bibliographies and reference lists of identified RCT evaluative studies for inclusion in the review.
- Subject matter experts[14] were also consulted.

14 Tiago Peixoto, Rosemary McGee, Chris Barnett, Maren Duvendack, Howard White, Clare Ferguson and Brendan Halloran.

Table 1 Literature search – primary and secondary search terms

| Primary search terms | Secondary search terms |
|---|---|
| Social accountability<br>Impact evaluation | Participatory budgeting<br>Public expenditure tracking<br>(Citizen) report cards<br>(Community) scorecards<br>Social audits<br>Citizen charters<br>Community monitoring<br>Information dissemination |
| Service delivery<br>Impact evaluation | Public services<br>Education<br>Water/sanitation/WATSAN<br>Health<br>Infrastructure |
| | AND |
| Randomised control trials<br>RCTs<br>Experiment<br>Impact evaluation | Transparency<br>Voice<br>Accountability |

## 5.3 The studies

The search uncovered 15 relevant RCTs evaluating T&A initiatives in the following areas of service delivery: education, health, infrastructure (health, education, road building), as well as in social accountability related to political representation and participation. The studies are summarised in Table 2. No RCT evaluations were found of complaints mechanisms, community scorecards or public hearings/social audits.

Three out of the 15 interventions involved technology: Duflo *et al.* (2008) – using cameras in schools to monitor teacher attendance; Grossman, Humphreys and Sacramone-Lutz (2013) – text messaging of political representatives by constituents; Callen and Long (2012) – 'Quick Count Photo Capture': vote counting using photographs.

Table 2 Summary of included studies

| Author(s) | Description |
|---|---|
| Community monitoring | |
| Humphreys *et al.* 2012 | RCT evaluation of the first stage of a governance programme: community monitoring of local development projects in Eastern DRC. Participant communities randomly selected through public lotteries from a larger pool of potential participating communities. Coupled with an experimental element to measure behavioural change. |
| Barr *et al.* 2012 | Experiment to test the impact of two, related community-based monitoring interventions in accountability in service delivery in education, via School Management Committees (SMC) in Uganda, based on a school scorecard monitoring tool. RCT combined with lab experiments. |

cont./

Table 2 Summary of included studies (cont.)

| Author(s) | Description |
|---|---|
| Duflo *et al.* 2008* | RCT of community monitoring programme using ICT. Teacher attendance monitored by cameras; salaries paid made a function of attendance; randomised experiment and a structural model to test whether monitoring and financial incentives can reduce teacher absence and increase learning in rural India. |
| Information dissemination | |
| Banerjee *et al.* 2010 | Evaluation of three different interventions to encourage beneficiaries in resource allocation and monitoring and management of school performance via public school committees. Randomised experiment in the state of Uttar Pradesh (UP) in India. |
| Pandey *et al.* 2009 | Evaluation of a community-based RCT to determine the impact of an information campaign on learning and other school outcomes. RCT evaluation with FGDs. India. |
| Olken 2007 | Randomised field experiment on reducing corruption in over 600 Indonesian village road projects: (i) audit experiment and (ii) participation experiments. Two different experiments that sought to increase grass roots monitoring of the project via improved participation at village-level 'accountability meetings'. Indonesia. |
| Malesky, Schuler and Tran 2012 | A randomised experiment to test whether transparency initiatives (publishing transcripts and scorecards from legislative sessions on a newspaper website) can be exported to authoritarian regimes and provide the same beneficial effects observed in democratic contexts. Vietnam. |
| Lieberman, Posner and Tsai 2013 | Post-treatment, matched village evaluation design to measure the impact of an initiative providing information on citizen activism. Measures differences in citizen activism between households that received children's assessment scores and instruction materials on how to act to improve children's learning, and those that did not, as well as spillover effects. Kenya. |
| Chong *et al.* 2010 | RCT to examine the effects of an information campaign on electoral participation and incumbent parties' vote share in the 2009 municipal elections in Mexico. Randomly assigned electoral precincts in 12 municipalities in the states of Jalisco, Morelos and Tabasco to one of four groups receiving flyers containing different types of information. |
| Humphreys and Weinstein 2012 | RCT study of the impact of accountability mechanism (Uganda's Parliamentary Scorecard) on the behaviour of members of parliament (MPs), the attitudes of voters, and on electoral outcomes. |
| Report cards | |
| Lassibille *et al.* 2010 | Randomised impact evaluation of interventions to improve management of the teaching process in Madagascar. Testing impact of tools to streamline and tighten workflow processes of all actors along service delivery chain, including: workflow templates/tools, report cards, instruction guidebooks; facilitating meetings between school staff and community to develop/agree school improvement plan; training sessions. |

cont./

Table 2 Summary of included studies (cont.)

| Author(s) | Description |
|---|---|
| Björkman and Svensson 2009 | Randomised field experiment on increasing community-based monitoring in 50 'communities' from nine districts in Uganda, plus a quantitative survey. Piloting citizen report cards aimed at enhancing community involvement and monitoring in the delivery of primary health care. |
| **Public Expenditure Tracking Survey (PETS)** | |
| Reinikka and Svensson 2006 | Experimental evaluation of newspaper campaign to boost the ability of schools and parents to monitor local officials' handling of a large school grant programme (1990s); PETS. Natural experiment – difference-in-differences approach with instrumental variable analysis: distance to the nearest newspaper outlet used as an instrument of exposure to the campaign. |
| **Election monitoring** | |
| Callen and Long 2012* | Experimental evaluation of Quick Count Photo Capture – a monitoring technology designed to detect the illegal sale of votes by corrupt election officials to candidates. |
| **Other (voice)** | |
| Grossman *et al.* 2013* | RCT evaluation of text messaging of political representatives by constituents: sampled constituents in Uganda with an opportunity to send a text message to their representatives at one of three randomly assigned prices. To examine whether ICTs can flatten interest articulation and how access costs determine who communicates and what gets communicated to politicians. |

*T4T&A initiatives.

## 5.4 Analysis

The 15 relevant evaluations of T&A initiatives were analysed against the framework set out in Section 5.1. As the questions are clustered under four domains – Design, Contribution, Explanation and Effects – this section discusses each domain in turn.

### 5.4.1 Design

Questions related to the design of the evaluation combine basic descriptions of the intervention and evaluation design with appraisal of the degree to which the basic underlying assumptions of experimental approaches hold:

1 What evaluation questions are posed?
2 What is the intended impact of the programme/intervention?
3 Is the evaluation design 'randomisable'?
4 Is there a clear counterfactual/mapping of causal influence?
5 Does the counterfactual hold?
6 Are ethical considerations clearly elucidated/taken into account?

**Evaluation questions** are not posed explicitly in ten out of the 15 studies. This relates in part to the tendency for the evaluations and interventions to lack explicit theories of change (discussed earlier and returned to later), but also reflects the tendency in these studies towards hypothesis testing and a focus on measuring outcomes. In the latter respect there is

an implicit focus on whether or not the interventions have achieved whatever it was they set out to do.

Across all of the studies scant attention is paid to stating upfront the **intended impacts** of the interventions being evaluated, with the studies varying in the degree to which they couched impact in terms of intermediate outcomes or outputs (for example, Lassibille *et al.* 2010: 'strengthen administration of the education system'), or in terms of broader 'bigger picture' impact (for example, Reinikka and Svensson 2006: to reduce capture and corruption). A 'lack of clarity in what the intended impacts actually are' is also highlighted in the review by Joshi (2011). The stated impacts of the programmes are given below:

- 'bring[ing] about change in local accountability and social cohesion as well as improve the welfare of communities' (Humphreys *et al.* 2012);
- 'encourage collective action' (Barr *et al.* 2012);
- improved attendance of teachers, improved test scores of pupils (Duflo *et al.* 2008);
- 'make large group control work better by mobilizing public sentiments and providing better information. A second goal was to sensitize people about the importance of education for their children and the state of education in the village, with the expectation that this would encourage them to try to do something about it' (Banerjee *et al.* 2010);
- improve school outcomes by 'providing information through a structured campaign to communities about their oversight roles in schools' (Pandey *et al.* 2009);
- reduce corruption in road-building projects (Olken 2007);
- 'increased openness forces delegates to perform better in order to win over voters in an electoral democracy' and 'increased transparency enables voters to choose better candidates for office – tossing out the laggards and selecting delegates who are more likely to act with the constituency's interests in mind' but 'transparency in authoritarian assemblies may have two contradictory effects' (Malesky *et al.* 2012);
- 'citizens can play [a role] in improving public service provision through mechanisms other than voting – for example, by exerting pressure directly on (unelected) service providers or by taking individual or collective actions that substitute for inadequate actions taken by the state' (Lieberman *et al.* 2013);
- 'information can cause electoral retribution in at least two ways: To "throw the rascal out", voters can cast a ballot in favour of the opposition or voters can abstain from voting at all' (Chong *et al.* 2010);
- 'transparency initiatives [scorecards] plausibly strengthen the incentives for incumbent politicians to perform well' (Humphreys and Weinstein 2012);
- 'strengthen administration of the education system' (Lassibille *et al.* 2010;
- 'strengthen providers accountability to citizen-clients by enhancing communities ability to monitor providers' (Björkman and Svensson 2009);
- to reduce capture and corruption (Reinikka and Svensson 2006);
- 'Monitoring reduces the incidence of theft or damaging of election materials at polling centres' (Callen and Long 2012);
- 'Does the introduction of an ICT system result in representative information on constituency needs and preferences?' (Grossman *et al.* 2013).

With respect to RCT evaluation design – whether or not implementation is **randomisable**, whether there is a clear **counterfactual** and/or mapping of **causal influence** and whether these hold – all of the interventions evaluated in the studies could be considered 'randomisable'. It was possible to either set up an experiment, a quasi-experiment, or to take advantage of 'natural' experimental conditions due to underlying programme design. The

nature of the method implies that analysis is based on causal influence underpinned by the construction of a counterfactual or control. The mechanisms being evaluated all have 'excludability' – it is possible to apply the intervention to some and to withhold it from others, to construct a plausible counterfactual in the form of a control group, thus minimising selection bias and confounding factors. The exception here is arguably the intervention evaluated by Reinikka and Svensson (2006), a newspaper campaign to boost the ability of schools and parents to monitor local officials' handling of a large school grant programme. The plausibility of the RCT approach here depends on the credibility of the instrument (distance to nearest newspaper outlet) they use to control for selection bias and proxy for potential exposure to the campaign (contamination) of 'control' group participants.

Just one paper mentions **ethical considerations** explicitly. The Tuungane programme evaluated by Humphreys *et al.* (2012) used public lotteries for selection into treatment and control groups and this is described in the report as representing a degree of informed consent. The RCT of alternative pricing for text messages to MPs in Uganda mentions ethical concerns indirectly in relation to the low purchasing power of the cost of the text messages that one treatment group faced (Grossman *et al.* 2013). It is somewhat surprising that the ethics of using cameras as a monitoring tool in schools is not mentioned in Duflo *et al.* (2008).

### 5.4.2 Contribution

The group of questions under the banner 'contribution' focus on the way the evaluation design deals with and explains the contribution of the intervention to potential outcomes and impacts.

1 Does the evaluation design identify multiple causal factors (where relevant)?
2 Does the design take into account whether the causal factors are independent or interdependent (where relevant)?
3 Can the design analyse effects of contingent, adjacent and cross-cutting interventions (where relevant)?
4 Are issues of 'necessity', 'sufficiency' and probability discussed?
5 Is justification for the evaluation methods informed by success elsewhere?
6 Does the evaluation design consider issues of timing, sequencing and durability?

**Multiple causal factors**: The studies reviewed focus on testing hypothesised relationships but are very much restricted to what is in the model, decided *a priori* by the investigators. In line with the underlying causal model of an RCT, they explore single causal relationships. They are similarly restricted in their ability to uncover and include in the analyses 'unexpected' causal influences, as the studies have no remit to ask those kinds of questions. One exception is Pandey *et al.* 2009, which combines RCT with focus group discussions. The study by Humphreys *et al.* (2012) acknowledges potential confounding factors and multiple causal pathways, but still relies solely on an RCT as the evaluation method, and the sequencing of implementation and timing of the RCT as a basis for assuming that the potential bias from confounding factors is minimised. In one of the few papers on a T4T&A intervention, Grossman *et al.* test multiple hypotheses about the likely effects of ICT on political communication and discuss the likelihood of multiple interacting factors leading to different kinds of impacts. The discussion of the RCT of the tech-based intervention in election monitoring in Afghanistan in Callen and Long (2012: Quick Count Photo Capture designed to detect the illegal sale of votes by corrupt election officials to candidates) considers alternative interpretations of their results, but the authors did not have the data to be able to test alternative models and instead focused on the one model that offered a simplified framework for interpreting these results.

An RCT by design **assumes causal factors are independent** and in these cases the evaluations are not set up to analyse effects of contingent, adjacent and cross-cutting interventions. **Issues of 'necessity', 'sufficiency' and probability are not discussed**.

Whether the **justification for methods is informed by success elsewhere** relates to whether or not the initiative design understands enough the reasons for success of one set of tools or approaches in one context before adapting, replicating or scaling to other settings. All of the studies are grounded in a consideration of whether and how interventions had worked elsewhere, as well as considering the success of an RCT approach to evaluating similar interventions. Especially strong in this regard were Barr *et al.* (2012) and Duflo *et al.* (2008). However, it is notable that none of the evaluations consider the successful use of other methodological approaches such as case studies, which suggests that the choice of RCTs as the method may have followed a different selection logic than an impartial appraisal of a full range of possible methods in the light of the evaluation question(s).

In experimental design cause comes before effect. Therefore in this respect all of the study designs consider **issues of timing and sequencing** although issues of **durability** are unclear (in both intervention and evaluation design). In all cases a baseline was conducted, the programme implemented, and effect realised (or not) – although see later points on short-term focus of the impact evaluations: most are conducted within a very short time frame of the intervention being implemented.

### 5.4.3 Explanation

This set of questions is concerned with how well the evaluation is able to explain how the intervention has caused a particular effect – essentially, how well does it set out a clear theory of change that reflects the realities of T&A interventions. The main questions are:

1 Is it clear how causal claims will be arrived at?
2 Is the design able to answer 'how' and 'why' questions?
3 Is a clear theory of change articulated? How is this derived?
4 Is theory used to support explanation?
5 Are alternative explanations considered and systematically eliminated?
6 Does design take into account complex, contextual factors?

The studies are generally very clear in justifying and explaining **how causal claims will be arrived at** by virtue of the chosen methodology. Only one of the studies, Pandey *et al.* (2009), builds into the design methods for uncovering/deepening explanations for observed impacts – **'how' and 'why'** – by including focus group discussions. Barr *et al.* (2012) are also able to some extent to explain how the interventions worked due to the evaluation design. The study by Lieberman *et al.* (2013) of the treatment effects of the intervention providing educational outcome information as well as guidance on how to act to bring about further improvement, strengthens the findings of the quasi-experimental component of the evaluation by also encompassing in-depth interviews with village elders and teachers, and focus group discussions with village elites. These were originally designed in order to answer the 'how' and 'why' questions related to changing attitudes and behaviours. But because the study failed to find a treatment effect, these served to confirm that the programme in fact had no impact on citizen activism in the timescale considered.

In terms of **theories of change** the studies vary in the extent to which they articulate clearly a theory of change and disentangle common assumptions about the links between

transparency, accountability and participation. The studies that are more successful in this regard are: Humphreys *et al.* (2012), Barr *et al.* (2012), Grossman *et al.* (2013), Malesky *et al.* (2012) and Pandey *et al.* (2009). All of these studies are hypothesis based and the theories of change cascade from here. The theory of change underlying the school monitoring initiative in Duflo *et al.* 2008 is noted to be 'ambiguous' by the authors, drawing on empirics in the absence of a strong underlying theory to draw on. The theory of change underlying Björkman and Svensson (2009) is also reasonably clear: improving citizen voice and accountability via dissemination of information on social services at the local level will improve service delivery. This is also one of only two studies (the other is Pandey *et al.* 2009) that **considers and eliminates alternative explanations** – although neither are systematic in this regard.

Few of the evaluations **take into account complex, contextual factors**, including the capacities and incentives on both the citizen and state side of the equation, and the linking mechanisms across the two. By using two different measures of fraud, Callen and Long take account of complex environments with 'highly adaptive political agents' (2012: 3). In the experiment to test outcomes using different types of scorecards (Barr *et al.* 2012), the participatory nature of developing scorecards captures contextual factors at the school level for that particular treatment arm, even though overall contextual factors are only differentiated at district level. In including political economy and 'constraints and opportunities resulting from the existing policies of the state government', Banerjee *et al.* (2010) also go some way towards taking into account context in explaining the impact of the intervention. However, this was notably missing from the other studies (or is somewhat superficial in the case of Pandey *et al.* 2009, in the purposive sampling of regions to get a balance of contexts). This suggests the external validity (generalisability) of these studies is somewhat limited.

### 5.4.4 Effects

This part of the framework refers to the degree to which the evaluation explains how the intervention worked.

1 Does the evaluation design include methods for tracking change over time, including reference to a clear baseline?
2 Are long-term effects identified?
3 Are these effects related to intermediate effects and implementation trajectories?
4 Is the question 'impact for whom' addressed in the design?
5 Does the IE find evidence of impact?

All of the studies had a **baseline** prior to implementation of the intervention(s). Most of the evaluations took place within too short a time frame from baseline to implementation to endline to be able to identify **long-term effects**. The longest time frame between intervention and valuation was the newspaper campaign evaluation conducted by Reinikka and Svensson (2006) with the PETS carried out six years following the launch of the campaign. In relation to the impact of information provision on learning and other school items, Pandey *et al.* (2009) suggest long-term effects are important but are not proven. Just Duflo *et al.* (2008) reported on impact beyond the life of the evaluation, stating that for the school monitoring initiative using a combination of cameras and salary deductions to monitor and sanction teacher (non-)attendance, the treatment effect remained strong even after the post-test, which marked the end of the formal evaluation. After four years, teacher attendance was still significantly higher in the schools with cameras. Similarly, only Duflo *et al.* consider **intermediate effects and implementation trajectories**: the evaluation looks at impact after one year and again after two to two-and-a-half years.

The evaluators themselves highlight the problem of conducting the evaluation too soon after the start of implementation on the T&A intervention and not allowing sufficient time to unfold for impacts to be felt:

> However, the greatest limitation of the findings reported here is that outcomes are measured soon after the intervention. Changing behaviour to change school outcomes requires time. Barriers to collective action are apparent from focus group discussions and may take time to be overcome, especially in light of the recent studies that document institutional inertia (Sokoloff and Engerman 2000; Banerjee and Iyer 2005). Future research is needed to examine whether behavioural changes translate into learning and whether a campaign sustained over longer time generates greater impact and whether there are sustained differences in impacts across states (Pandey *et al.* 2009: 374).

> Even for this one-time intervention, certain results will become apparent only with time. For example, after several years, it will be apparent whether the increased scrutiny imposed by the audits affects who chooses to become involved in project management, and whether negative audit findings affect the re-election probabilities of village officials. Reducing corruption may also reduce campaign expenditures for village offices, since the rents from obtaining these positions will have declined. Whether the reduced campaign expenditures take the form of fewer cash hand-outs to villagers, or fewer banners advertising the candidates' names, will determine the ultimate general equilibrium social welfare implications of the reduction in corruption. The efficiency impact of the reduction in corruption will also become clearer with time since we can observe changes in how long the road lasts. Understanding the long-run implications of anticorruption policies remains an important issue for future research (Olken 2007: 244).

Too short time horizons are also suggested by the authors to be a possible factor in explaining their failure to find an impact on citizen activism of the Uwezo education intervention initiative, in Kenya (Lieberman *et al.* 2013):

> A third possibility is that inadequate time had elapsed between the assessment and our measurement of its impact. Real behavioral change may require reflection, discussions with other community members, and a rearrangement of commitments and prior obligations to make room for new activities and behaviors. Three months may simply have been too short an interval for these processes to work themselves through (Lieberman *et al.* 2013: 23).

They also float the possibility that the impact could have been so short-lived that three months' interval could equally have been too long. This is something that the qualitative work ought to have been able to shed light on, by asking people directly about their actions. In general, though, from the range of studies of T&A consulted for this review, it is felt that behavioural impacts of T&A initiatives are most likely to manifest themselves over the longer term. The implications for evaluation design, whether based on an RCT or other methods, is that in order to take account of this 'unknown' evaluations need to be both forward and backwards looking – either via a staged/sequenced design or by building in components with a well-designed recall element (in cases where impact may have been shortlived).

**Impact for whom** is only narrowly addressed in some of the evaluation designs. For example, Duflo *et al.* (2008) have as key outcomes test scores for children, and attendance of teachers,

but there is no scope to explore other impacts, such as community-wide effects. The interventions in Lassibille *et al.* (2010) were packaged and targeted to two main sets of actors: midlevel bureaucrats (subdistrict and district administrators) and frontline service providers at the school level (teachers and school directors), representing two to three categories of potential beneficiary. The Tuungane evaluation (Humphreys *et al.* 2012) makes a distinction between individual, household and village-level impacts but does not consider impacts at the institutional level or at the level of other key actors/stakeholders in the programme.

All of the evaluations but Lieberman *et al.* (2013) and Humphreys *et al.* (2012) found **evidence of impact**. The RCT evaluation of the Tuungane community monitoring programme in DRC found no evidence of impact. Reasons suggested by the authors to explain lack of impact relate to both the evaluation design and the design of the intervention itself and manifest in three domains: level of analysis, measurement of outcomes, and scale (see Table 3).

Table 3 'Accounting for null effects'

| | Research | Intervention |
|---|---|---|
| Level | Research focused on general populations but perhaps the treatment had strongest impacts on community leaders only. Perhaps treatment had impacts at VDC or CDC level but not village level. | Treatment should target governance at levels higher than local communities. |
| Outcomes | Research measured governance in unstructured environments, but perhaps treatment affected governance in more structured environments only. Perhaps measurement of outcomes is flawed. | The treatment did not effectively address fundamentals – such as the material distribution of power. |
| Scale | Measures were taken too soon after completion of VDC projects. Perhaps block randomisation strategy led to risks of spillovers. | The treatment was too small and too short. |

*Source* Reproduced from Humphreys *et al.* (2012: 76, Item 18).

Explanations for the lack of treatment effect found by Lieberman *et al.* (2013), also discussed above, focused on the evaluation design: (i) impact was so small that the sample size was correspondingly not large enough for the impact to be seen; (ii) not enough households in a village received the intervention ('treated') which means there was no critical mass that would enable action to happen; (iii) as discussed above, the impact evaluation was conducted just three months following the intervention which did not leave enough time for people to act and for impact to be felt. However, the reason for lack of impact considered to be most important by the authors was an apparent 'absence in our study setting of a set of key conditions that must be present for an informational intervention to plausibly generate citizen activism'. Using a framework to think through the answers to the question: *What must be true for us to reasonably expect the provision of information to an individual to cause him/her to change his/her behaviour?*, the authors suggest a number of conditions, the absence of any of which can lead to a break in the causal chain. Thus, for information to generate citizen action:

- it must be understood;
- it must be new;
- it must cause people to update their prior beliefs and assumptions;
- it must relate to an issue that people care about and feel it is their responsibility to address;
- the people at whom the information is directed must possess the skills and knowledge to know what to do in light of the information;
- they must have the efficacy to believe that their actions will generate results; and,
- to the extent that the outcome in question requires collective action, they must believe that others in the community will act as well (Lieberman *et al.* 2013: 3).

This elucidation of a theory of change was *post hoc* in that the evaluators engaged in the exercise in order to explain the null effects found using quasi-experimental/quantitative approaches. Other programmes and contexts can learn from this – not only in terms of analysis of the links between voice and accountability initiatives and impacts, but also in evaluation design, whereby a theory of change developed earlier in the process can help to uncover important factors such as expected timescales and ensure the evaluation design is appropriate and effective. However, in the cases where impact was found the studies do not tell us why they worked – just that they did – limiting their relevance to other contexts.

## 5.5 Summary

Turning to the overarching question: *Does the evaluation use methods of analysis which are appropriate to the purpose of the impact assessment, taking into account its audience, the level of complexity involved, and positionality of those doing the study?* (McGee and Gaventa 2011b: 48), the included studies are mixed in this regard. They are, on the whole, well-designed and well-executed RCTs that address the (narrow) evaluation questions or hypotheses posed. However, they have key limitations. The nature of the RCT method means that the evaluations focus strongly on 'who?' and 'what?' (and 'how much') questions and little on 'why' and 'how'. Given the complexity of T&A interventions, and their implicit reliance on behavioural change, it is questionable that the right outcomes are being measured. With the exception of the Lieberman *et al.* study (2013), where findings contradict prior beliefs or hypotheses there is generally no unpacking of why this might be so, thus limiting learning from the evaluation, even though further investigation using complementary qualitative methods would strengthen the findings of the IE. Where impact is found the studies on the whole neglect contextual contributions to impact, they fail to analyse or take into account power and political economy, or how and why measures have worked (and how and why others have not). Finally, the timescales within which many of these evaluations have taken place are relatively short, calling into question the likelihood that an impact has begun to manifest itself.

# 6 Conclusion

This review has explored the overarching question: *How and in what ways might RCTs and other experimental methods be appropriate in the context of evaluating T4T&A programmes?* The literature search identified 15 evaluations of T&A initiatives in service delivery and in social accountability in the context of political participation and representation, based on an experimental/RCT design. Three of these were for technology-based (T4T&A) interventions. The 15 evaluations were analysed against criteria encompassing evaluation design, the way the evaluation deals with contribution of the intervention to impact, how well the evaluation is able to explain how the intervention caused an effect, and the extent to which the evaluation explains how the intervention worked. The RCTs considered here tend to test alternative mechanisms for T&A, in the context of pilot programmes, to evaluate which of these mechanisms are likely to achieve the desired impact. In some cases there is more than one 'treatment' arm where modifications of the same basic design are evaluated against a control group. Otherwise the studies are basic 'with–without' evaluations with one treatment arm and one control.

T&A work via multi-stranded, embedded programmes, with multiple mechanisms, and it is these things that (a) make it hard to evaluate and (b) do not lend themselves well to RCT methodology. The intended impacts of the T&A initiatives in these studies tend towards the 'bigger picture', whereas RCTs focus on clearly defined, narrow outcomes. It is possible to identify a host of additional evaluation questions to those posed by the experiments that one would need to be able to answer in order to get a full picture of the effectiveness (impact) of the intervention(s) – too many questions for an RCT to address. We therefore need to bear in mind this is just one tool with a very specific focus, that may well not tell you enough of what you need to know about the full impact of (T4)T&A initiatives, particularly how and why changes happen.

In this respect, the findings of this review chime with broader reviews of IE in this field. Here we try to draw out lessons which will be of use to the Making All Voices Count programme, its major stakeholder groups and other practitioners in the T&A field, and the broader community of those involved in designing, implementing, funding and evaluating T4T&A initiatives.

The evaluations reviewed here are on the whole well-designed and well-executed RCT studies. In analysing these against the framework set out in Section 5.2, based on a set of guiding questions relating both to quality of design and methods and to the specific characteristics of T&A programmes, it is possible to make some general observations about the applicability and usefulness of RCTs in evaluation in this context.

Overall, when a programme strategy is to be politically transformative, RCTs do not work in evaluating impact in terms of transformation. Neither do they work as a sole method of evaluation where a programme is taking an adaptive and iterative approach. However, as an evaluative tool, RCTs can be an effective and useful means of deciding between different variations in an intervention design in the context of piloting a programme, especially in evaluating a subset of an intervention: for example, where a pilot within an adaptive programme is tested on a small scale before rolling out. In programme design, implementation (treatment) has to be randomisable and there need to be exclusionary factors. This can be difficult in T&A especially if policies apply nationally, for example, or regionally, or at other administrative levels for implementation.

A recurring theme in this review, drawing on the literature as well as the analysis of the included evaluations, is the importance of a clearly defined and well-articulated theory of change – both for effective programme implementation and evaluation of impact. While all the reviewed papers had implicit theories of change underlying the T&A initiatives, in many cases articulated via a series of hypotheses that the evaluation aimed to test, the assumptions that lead from cause (the intervention) to effect (the intended impact) are not made explicit, with no alternative, plausible, causal links identified. This makes it difficult to assess how and where initiatives work in the intermediate stages.

Consideration of context is also limited. For example, the evaluations here do not take account of important factors such as social effects and social networks on people's behaviour and are therefore limited in drawing learning more broadly as well as missing critical components and potentially key contributory factors in the impact pathway. If context is not taken adequately into account then evaluations are of limited external validity as we cannot be sure if they would work if scaled up.

There is in particular a strong case here for not only paying more attention to the theory of change, encompassing impact pathways, context and underlying assumptions, but also re-examining and evolving these approaches, both in the context of T&A/T4T&A and more broadly. Rather than seeing the development of a theory of change as being something that is 'done for the donors', instead it could be instituted as a crucial and evolving tool in design and planning, as a framework for learning and for assessing and managing risk – considering not only impact pathways to positive change but also theories of 'negative' change.

Overall, an RCT approach does lend itself well to technology-based initiatives, even given the challenges in the T&A context described in detail in this and other reviews. It is an ideal method for testing alternative mechanisms and/or technologies – although the three T4T&A RCTs analysed here do not do this (Duflo *et al.* 2008; Grossman *et al.* 2013; Callen and Long 2012). However, the value of this exercise needs to be balanced against cost. For small, one-off, low-cost interventions the generally high cost of doing a high-quality RCT could well preclude its use in evaluation. But if the programme is intended to be scaled up – say to national or even regional level – it could well be worth the investment.

This review also suggests that an RCT design for T&A evaluation needs to be improved by:

- Longer time frames between implementation and endline evaluation to allow sufficient time for impacts to be manifest, especially where technology is relatively new to users with little known about their propensity to take it up;
- Evaluations at intermediate stages of the implementation process – midline – in order to gauge intermediate impacts;
- Overall design based on a range of methods – qualitative and quantitative, experimental and non-experimental – to complement the RCT component;
- Clearly articulated theories of change, in order to ensure underlying models are correctly specified and to help identify the most appropriate 'package' of methods.

# References

Alsop, R. and Heinsohn, N. (2005) *Measuring Empowerment in Practice: Structuring Analysis and Framing Indicators*, World Bank Policy Research Working Paper 3510, February, http://siteresources.worldbank.org/INTEMPOWERMENT/Resources/41307_wps3510.pdf

Amin, Samia; Das, Jishnu and Goldstein, Markus (eds) (2008) *Are You Being Served? New Tools for Measuring Service Delivery*, Washington DC: World Bank

Avila, Renata; Feigenblatt, Hazel; Heacock, Rebekah and Heller, Nathaniel (2010) *Global Mapping of Technology for Transparency and Accountability: New Technologies*, Transparency and Accountability Initiative Report, November, http://ict4peace.org/wp-content/uploads/2011/05/global_mapping_of_technology_final.pdf (accessed 27 August 2014)

Banerjee, Abhijit and Iyer, Lakshmi (2005) 'History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India', *American Economic Review* 95.4: 1190–1213

Banerjee, A.; Duflo, E.; Glennerster, R.; Banerji, R. and Khemani, S. (2010) 'Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation of Education in India', *American Economic Journal: Economic Policy* 2.1: 1–30

Barahona, C. (2010) *Randomised Control Trials for the Impact Evaluation of Development Initiatives: A Statistician's Point of View*, ILAC Working Paper 13, Rome: Institutional Learning and Change Initiative, www.cgiar-ilac.org/content/working-papers (accessed 3 June 2014)

Barr, A.; Mugisha, F.; Serneels, P. and Zeitlin, A. (2012) 'Information and Collective Action in the Community Monitoring of Schools: Field and Lab Experimental Evidence from Uganda', unpublished manuscript, www.tilburguniversity.edu/upload/397359d2-09a0-40d5-9502-8c3592aaae40_zeitlin.pdf (accessed 27 August 2014)

Barrett, Christopher B. and Carter, Michael R. (2010) 'The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections', *Applied Economic Perspectives and Policy*, Winter, 32.4: 515–48

Bellver, A. and Kaufmann, D. (2005) 'Transparenting Transparency: Initial Empirics and Policy Applications', paper presented at IMF Conference on Transparency and Integrity, World Bank, Washington DC, 6–7 July

Björkman, M. and Svensson, J. (2009) 'Power to the People: Evidence from a Randomized Field Experiment on Community-based Monitoring in Uganda', *Quarterly Journal of Economics* 124.2: 735–69

Calland, R. (2011) 'Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Annex 2 Freedom of Information', prepared for the Transparency and Accountability Initiative Workshop, IDS, Brighton, October 2010, www.transparency-initiative.org/ workstream/impact-learning (accessed 9 November 2011)

Callen, M. and Long, J.D. (2012) 'Institutional Corruption and Election Fraud: Evidence from a Field Experiment in Afghanistan', unpublished working paper, www-igcc.ucsd.edu/assets/001/502329.pdf

Camfield, L. and Duvendack, M. (2014) 'Are We Off the Gold Standard?', *European Journal of Development Research* 26.1: 1–11

Carlitz, R. (2011) 'Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Annex 2 Budget Processes', prepared for the Transparency and Accountability Initiative Workshop, IDS, Brighton, October 2010, www.transparency-initiative.org/workstream/impact-learning (accessed 9 November 2011)

Cartwright, N. and Munro, E. (2010) 'The Limitations of Randomized Controlled Trials in Predicting Effectiveness', *Journal of Evaluation in Clinical Practice* 16: 260–6

CGD (2006) *When Will We Ever Learn? Improving Lives through Impact Evaluation*, Report of the Evaluation Gap Working Group, Washington DC: Center for Global Development

Chong, A.; De La O, A.; Karlan, D. and Wantchekron, L. (2010) 'Information Dissemination and Local Governments' Electoral Returns: Evidence from a Field Experiment in Mexico', unpublished paper, www.povertyactionlab.org/publication/information-dissemination-and-local-governments%E2%80%99-electoral-returns-evidence-field-experi (accessed 1 September 2014)

Claasen, M. and Alpín-Lardiés, C. (eds) (2010) *Social Accountability in Africa: Practitioners' Experience and Lessons*, Cape Town: IDASA/ANSA Africa

Coryn, C.; Noakes, L.; Westine, C. and Schroter, D. (2011) 'A Systematic Review of Theory-driven Evaluation Practice from 1990 to 2009', *American Journal of Evaluation* 32: 199–226

Deaton, Angus (2010) 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature* 48.2: 424–55

Duflo, E.; Hanna, R. and Ryan, S.P. (2008) *Incentives Work: Getting Teachers to Come to School*, www.povertyactionlab.org/sites/default/files/publications/Incentives%20Work,%20Getting%20teachers%20to%20come%20to%20school.pdf (accessed 3 June 2014)

Fox, J. (2014) 'Social Accountability: What does the Evidence Really Say? – May 14 2014 – GPSA Global Forum', powerpoint presentation, Global Partnership for Social Accountability, http://gpsaknowledge.org/knowledge-repository/social-accountability-what-does-the-evidence-really-say/ (accessed 27 August 2014)

Fox, J. (2007) 'The Uncertain Relationship Between Transparency and Accountability', *Development in Practice* 17.4: 663–71

Grossman, G.; Humphreys, M. and Sacramone-Lutz, G. (2013) 'Does Information Technology Flatten Interest Articulation? Evidence from Uganda', http://e-gap.org/wp/wp-content/uploads/EGAP9_Grossman_etal.pdf (accessed 1 September 2014)

GSDRC (2010) *Helpdesk Research Report: Evaluations of Voice and Accountability Instruments*, Governance and Social Development Resource Centre, April, www.gsdrc.org/docs/open/HD675.pdf (accessed 1 September 2014)

GSDRC (2011) *Helpdesk Research Report: RCTs for Empowerment and Accountability Programmes*, Governance and Social Development Resource Centre, April, www.gsdrc.org/docs/open/HD756.pdf (accessed 1 September 2014)

Guijt, I. (2007) *Assessing and Learning for Social Change: A Discussion Paper*, Learning by Design and IDS, www.ids.ac.uk/files/dmfile/ASClowresfinalversion.pdf (accessed 3 June 2014)

Harrison, Glenn W. (2013) 'Field Experiments and Methodological Intolerance', *Journal of Economic Methodology* 20.2: 103–17

Hawken, Angela and Munck, Gerardo L. (2009) 'Measuring Corruption: A Critical Assessment and a Proposal', in Rajivan, Anuradha K. and Gampat, Ramesh (eds), *Perspectives on Corruption and Human Development*, Vol. 1, New Delhi: Macmillan India for UNDP: 71–106

Holland, J. and Thirkell, A., with Trepanier, E. and Earle, L. (2009) *Measuring Change and Results in Voice and Accountability Work*, Working Paper 34, London: Department for International Development (DFID)

Houtzager, P.; Joshi, A. and Gurza Lavalle, A. (eds) (2008) 'State Reform and Social Accountability', *IDS Bulletin* 38.6, Brighton: IDS

Humphreys, M. and Weinstein, J. (2012) *Policing Politicians: Citizen Empowerment and Political Accountability in Uganda*, Working Paper, March, London: International Growth Centre

Humphreys, M.; Sanchez de la Sierra, R. and van der Windt, P. (2012) *Social and Economic Impacts of Tuungane: Final Report on the Effects of a Community Driven Reconstruction Program in Eastern Democratic Republic of Congo*, New York: Columbia University

Ibrahim, S. and Alkire, S. (2007) *Agency and Empowerment: A Proposal for Internationally Comparable Indicators*, OPHI Working Paper 4, Oxford: University of Oxford

Jonas, N.; Jonas, H.; Steer, L. and Datta, A. (2009) *Improving Impact Evaluation Production and Use*, Occasional Paper 300, London: Overseas Development Institute (ODI), www.odi.org.uk/resources/docs/4158.pdf

Jones, H. (2009) 'The "Gold Standard" is not a Silver Bullet for Evaluation', *Opinion*, Overseas Development Institute (ODI)

Joshi, A. (2011) 'Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Annex 1 Service Delivery', prepared for the Transparency and Accountability Initiative Workshop, IDS, Brighton, October 2010, www.transparency-initiative.org/workstream/impactlearning (accessed 10 November 2011)

Joshi, A. (2008) 'Producing Social Accountability? The Impact of Service Delivery Reforms', *IDS Bulletin* 38.6: 10–17

Jupp, D. and Ibn Ali, S., with Barahona, C. (2010) *Measuring Empowerment? Ask Them – Quantifying Qualitative Outcomes from People's Own Analysis*, Sida Evaluation Series 1, Stockholm: Sida

Lassibille, Gérard; Tan, Jee-Peng; Jesse, Cornelia and Van Nguyen, Trang (2010) 'Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions', *The World Bank Economic Review* 24.2: 303–29

Lieberman, Evan; Posner, Daniel and Tsai, Lily (2013) *Does Information Lead to More Active Citizenship? Evidence from an Education Intervention in Rural Kenya*, Working Paper 2013–2, Massachusetts MA: Massachusetts Institute of Technology Political Science Department

Malesky, E.; Schuler, P. and Tran, A. (2012) 'The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly', *American Political Science Review* 106.04: 762–86

Mayne, J. (2012) 'Making Causal Claims', *ILAC Brief* 26, October, www.cgiar-ilac.org/files/publications/mayne_making_causal_claims_ilac_brief_26.pdf (accessed 27 August 2014)

McGee, R. and Carlitz, R. (2013) 'Learning Study on "The Users" in Technology-for-Transparency-and-Accountability Initiatives: Assumptions and Realities', Hivos Knowledge Programme, https://hivos.org/sites/default/files/ids-userlearningstudyont4tais_0.pdf (accessed 27 August 2014)

McGee, R. and Gaventa, J. (2011a) *Shifting Power? Assessing the Impact of Voice and Transparency Programmes*, IDS Working Paper 383, November, Brighton: IDS

McGee, R. and Gaventa, J. (2011b) 'Review of Impact and Effectiveness of Transparency and Accountability Initiatives', prepared for the Transparency and Accountability Initiative Workshop, IDS, Brighton, October 2010

McNeil, M. and Malena, C. (eds) (2010) *Demanding Good Governance: Lessons from Social Accountability Initiatives in Africa*, Washington DC: World Bank

Mejía Acosta, A. (2011) 'Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Annex 4 Natural Resource Governance', prepared for the Transparency and Accountability Initiative Workshop, IDS, Brighton, October 2010, www.transparency-initiative.org/workstream/impact-learning (accessed 10 November 2011)

Moehler, D. (2010), 'Democracy, Governance and Randomised Development Assistance', *Annals of the American Academy of Political and Social Science* 628: 30

Olken, B.A. (2007) 'Monitoring Corruption: Evidence from a Field Experiment in Indonesia', *Journal of Political Economy* 115.2: 200–49

O'Neil, T.; Foresti, M. and Hudson, A. (2007) *Evaluation of Citizens' Voice and Accountability: Review of the Literature and Donor Approaches Report*, London: Overseas Development Institute (ODI)

Pandey, P.; Goyal, S. and Sundararaman, V. (2009) 'Community Participation in Public Schools: Impact of Information Campaigns in Three Indian States', *Education Economics* 17.3: 355–75

Pawson, R. (2007) *Causality for Beginners. NCRM Research Methods Festival 2008*, Leeds University, http://eprints.ncrm.ac.uk/245/ (accessed 3 June 2014)

Pawson, R. (2006) *Evidence-Based Policy: A Realist Perspective*, London: Sage Publications

Peruzzotti, E. and Smulovitz, C. (2006) 'Social Accountability: An Introduction', in E. Peruzzotti and C. Smulovitz (eds), *Enforcing the Rule of Law: Social Accountability in the New Latin American Democracies*, Pittsburgh PA: University of Pittsburgh

Ravallion, M. (2009) 'Should the Randomistas Rule?', *The Economists' Voice* 6.2: 1–5

Reinikka, R. and Svensson, J. (2006) 'The Power of Information: Evidence from a Newspaper Campaign to Reduce Capture of Public Funds', http://people.su.se/~jsven/information2006a.pdf (accessed 1 September 2014)

Rogers, P. (2009) 'Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation', *Journal of Development Effectiveness* 1.3: 217–26

Sokoloff, Kenneth L. and Engerman, Stanley L. (2000) 'Institutions, Factor Endowments, and Paths of Development in the New World', *Journal of Economic Perspectives* 14.3: 217–32

Stern, E.; Mayne, J.; Befani, B.; Stame, N.; Forss, K. and Davies, R. (2012) *Developing a Broader Range of Rigorous Designs and Methods for Impact Evaluations*, Final report, London: Department for International Development (DFID)

Taylor, P.; Deak, A.; Pettit, J. and Vogel, I. (2006) *Learning for Social Change: Exploring Concepts, Methods and Practice*, Workshop Report, Brighton: IDS

Tisné, M. (2010) 'Transparency, Participation and Accountability: Definitions', unpublished background note for Transparency and Accountability Initiative

Transparency International (2009) *The Anti-Corruption Plain Language Guide*, Berlin: Transparency International

Vogel, I. (2012) *Review of the Use of 'Theory of Change' in International Development*, Review Report for UK Department for International Development, April, http://r4d.dfid.gov.uk/pdf/outputs/mis_spc/DFID_ToC_Review_VogelV7.pdf (accessed 2 June 2014)

White, H. (2009) *Theory-Based Impact Evaluation: Principles and Practice*, Working Paper 3, Delhi: International Initiative for Impact Evaluation

White, H. and Phillips, D. (2012) *Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework*, 3ie Working Paper 15, June, New Delhi: International Initiative for Impact Evaluation

# Annex 1 Making All Voices Count grants

| Scaling-up grants ID | Amount | Applicant | Title |
|---|---|---|---|
| 1300 | £93,083 | Indonesia Corruption Watch (ICW) | e-Procurement Potential Fraud Analysis – Network Expansion 1 yr |
| 1009 | £99,987 | Cooperative Housing Foundation doing business as Global Communities | Our City: Our Say – Increasing Women's Voice in Governance |
| – | £100,000 | Ghana Integrity Initiative | Your Voice Matters – Report Corruption! |
| 1159 | £99,765 | UK Citizens online Democracy (operating as the mySociety Project) | Accelerating and improving FOI implementation and monitoring in Liberia |
| 939 | £100,000 | International Rescue Committee (IRC) Inc. | Every Voice ('*Etoil Daang*' in Turkana language) |
| 1369 | £99,843 | Transparency International – Kenya | Scaling Up the Mobile Drug Tracking System |
| 1416 | £100,000 | Catholic Agency for Overseas Development (CAFOD) | Justice and Peace actors as catalysts of change |
| 1320 | £100,000 | The Black Sash Trust | Scaling up presidency-aligned citizen-based performance monitoring |
| 1080 | £ 90,000 | Health-e News Service | Our Health Citizen Journalism Programme |
| 750 | £98,080 | Centre for Municipal Research and Advice (CMRA) | Municipal Barometer Project |
| – | £39,891 | The Carter Center | Access to Information Legislation Implementation Assessment Tool |
| | £1020,649 | | |

MAKING ALL VOICES COUNT

Web www.makingallvoicescount.org
Email info@makingallvoicescount.org
Twitter @AllVoicesCount

RESEARCH, EVIDENCE AND LEARNING
WORKING PAPER