

BMZ Evaluation Division: Evaluation Working Papers

Micro-Methods in Evaluating Governance Interventions



Foreword by the Ministry for Economic Cooperation and Development

BMZ Evaluation Working Papers address a broad range of methodological and topical issues related to the evaluation of development cooperation. Some papers complement BMZ evaluation reports, others are free-standing. They are meant to stimulate discussion or serve as further reference to published reports.

The aim of this paper is to present a guide to impact evaluation methodologies currently used in the field of governance. It provides an overview of a range of evaluation techniques – focusing specifically on experimental and quasi-experimental designs. It also discusses some of the difficulties associated with the evaluation of governance programmes and makes suggestions with the aid of examples from other sectors. Although it is far from being a review of the literature on all governance interventions where rigorous impact evaluation has been applied, it nevertheless seeks to illustrate the potential for conducting such analyses.

This paper has been produced by Melody Garcia, economist at the German Development Institute (Deutsches Institut für Entwicklungspolitik, DIE). It is a part of a two-year research project on methodological issues related to evaluating budget support funded by the BMZ's evaluation division. The larger aim of the project is to contribute to the academic debate on methods of policy evaluation and to the development of a sound and theoretically grounded approach to evaluation. Further studies are envisaged.

The opinions presented in this study are those of the independent external expert and do not necessarily reflect the views of the BMZ.

This paper is available online at

http://www.bmz.de/en/what_we_do/approaches/evaluation/Evaluation/methods/index.html. It

should be cited as follows: Garcia, M. (2011): Micro-Methods in Evaluating Governance Interventions. *Evaluation Working Papers*. Bonn: Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung.

Michaela Zintl

Head of the division "Evaluation of Development Cooperation; Auditing"

Federal Ministry for Economic Cooperation and Development

Acknowledgements

This paper has been produced as a component of a more comprehensive research project at the German Development Institute on the “Evaluation of Budget Support as an Aid Instrument: Development and Applications of Evaluation Methods and Approaches,” financed by the evaluation and audit division of the German Federal Ministry for Economic Cooperation and Development (BMZ). This project forms part of a broader effort by Departments I (Bi- and Multilateral Development Cooperation) and III (Governance, Statehood and Security) of the German Development Institute to evaluate the effectiveness of new aid modalities.

I wish to thank primarily Jörg Faust, Stefan Leiderer and Michaela Zintl for the crucial advice and continued guidance they have given me. Valuable comments and suggestions on an earlier draft were provided by Christian von Haldenwang, Erik Lundsgaarde and Imme Scholz. I am also grateful for the comments received from the BMZ's division 203: Governance, Democracy and Rule of Law. I gratefully acknowledge the financial support of the BMZ. I take responsibility for any errors in this paper.

Melody Garcia

Contents

Abbreviations	vii
Executive summary	vii
1. Introduction	1
2. Governance as a core aspect of development cooperation	4
2.1. The concept of governance	4
2.2. Aid for governance	7
2.3. Design of governance programmes	9
3. Impact evaluation	10
3.1. Applying the general principles	10
3.2. The problem of attribution	13
4. Experimental or Randomised Control Trials (RCTs)	17
5. Quasi-experimental designs	17
5.1. Difference-in-differences (DID) approach	21
5.2. Instrumental variable approach	25
5.3. Propensity score matching (PSM)	29
5.4. Regression discontinuity design (RDD)	33
5.5. Combining techniques	35
5.6. Other evaluation approaches	35
5.7. Which technique to use?	36
6. Tackling the difficulties	40
6.1. Measuring outcome	40
6.2. Small sample size	42
6.3. No baseline data	43
6.4. Spillovers and contamination	43
6.5. Evaluating higher levels of government	43
6.6. Complex governance interventions	44
6.7. Intervention is full coverage	45
7. Practical implications	47
8. Summary and conclusions	49
9. References	51

Tables

Table 1 Country definitions of governance and good governance	5
Table 2 Governance dimensions, topics and activities.....	6
Table 3 Donor assistance in governance.....	8
Table 4 Difference-in-differences estimates of the effects on funding of having access to newspapers	25
Table 5 Summary toolbox	38
Table 6 Selected governance interventions, outcomes, measures and methods.....	41

Figures

Figure 1 Simplified design of governance programmes.....	9
Figure 2 Before-and-after scenario.....	15
Figure 3 Graphic illustration of difference-in-difference (DID).....	22
Figure 4 Illustration of biased DID	23
Figure 5 Stylised illustration by Reinikka and Svensson (2004).....	24
Figure 6 Region of common support.....	31
Figure 7 One-to-one matching, stylised example	32
Figure 8 Regression discontinuity design: an illustration.....	34

Boxes

Box 1 Principles of theory-based impact evaluation.....	12
Box 2 Impact evaluation techniques	17

Abbreviations and acronyms

ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
BMZ	Federal Ministry for Economic Cooperation and Development, Germany
CCT	Conditional Cash Transfer
CDD	Community Driven Development
CSO	Civil Society Organisation
DFID	Department for International Development, UK
DID	Difference-in-Differences
DIME	Development Impact Evaluation, World Bank
DRC	Democratic Republic of Congo
GBS	General Budget Support
GP	Gram Panchayat (local government in India at village/town level)
IE	Impact Evaluation
IEG	Independent Evaluation Group, World Bank
IMF	International Monetary Fund
IV	Instrumental Variable
JPAL	The Abdul Latif Jameel Poverty Action Lab
M&E	Monitoring and Evaluation
MCC	Millennium Challenge Corporation
OECD DAC	Organisation for Economic Co-operation, Development Assistance Committee
PROGRESA	Programa de Educación, Salud y Alimentación
PRS	Poverty Reduction Strategy
PSM	Propensity Score Matching
RCT	Randomised Control Trial
RIE	Rigorous Impact Evaluation
RDD	Regression Discontinuity Design
TEA	Traffic Enforcement Authority
UNDP	United Nations Development Programme
USAID	United States Agency for International Development
WB IEG	World Bank Independent Evaluation Group

Executive summary

Although billions of dollars have been spent on improving governance in the past decade, few of the programmes that have received funding have been subjected to strong and rigorous impact evaluation. Consequently, our understanding of the progress made by donors/governments in this field remains limited. This paper is designed as a working guide for practitioners who would like to conduct scientifically credible impact evaluations of donor-assisted and government-sponsored governance programmes. It advocates that policymakers pursue an “evidence-based” development policy in the decision-making on the important dimensions of governance.

Impact evaluation of governance programmes

The ultimate goal of impact evaluation is to understand the extent to which observed outcomes can be attributed to the programme, and to the programme alone. The term “rigorous” impact evaluation has grown in popularity over the past few years. It implies tackling the attribution problem through the use of experimental and quasi-experimental techniques in evaluation. Many economic and social sectors have been evaluated using these techniques: labour, health, nutrition, water, infrastructure, agriculture and education. Unlike these sectors, governance has been subjected to relatively little rigorous evaluation, some of the likely reasons being that:

- governance is characterised by complex interventions routinely combining various activities and producing outcomes that are difficult to measure;
- not many rigorous evaluators have focused on this field;
- insignificant and negative results are not published (publication bias);

- the topic is politically sensitive; and
- reliable baseline and longitudinal data are scarce.

There is considerable scope for evaluating certain types of governance interventions rigorously, perhaps far more than one would initially think. Admittedly, the applicability of quantitative techniques is subject to some limitations, but the interventions can be broken down into components suitable for this kind of analysis.

Which aspects of governance have been analysed using quantitative evaluation techniques?

Some aspects of governance programmes have been subjected to rigorous impact evaluation. Experimental design, and specifically randomised control trial have been frequently applied in evaluating programmes concerning corruption, elections and community development. Quasi-experimental designs – difference-in-differences, instrumental variables, propensity score matching and regression discontinuity designs – have also been used to assess issues related to government spending, local capture, voting behaviour, crime, citizen participation and transparency. A broader message conveyed by these examples is that experimental and quasi-experimental designs can be applied to evaluate difficult concepts, with the focus less on the success (or failure) of the intervention itself. While several donors have stepped up their efforts to evaluate governance, there is still significant room for rigorous impact evaluations, seeing that the number of evaluation studies that provide scientific evidence of impact is still small. Thus proof of the effectiveness of certain reforms and interventions remains weak, since they have not been replicated or evaluated under different conditions.

Why are governance programmes difficult to evaluate? What can be done about this?

Impact evaluation of governance programmes shares many of the challenges and data constraints facing other development programmes. In practice, the situation with which one is confronted is usually less ideal than that conveyed in theory. Some of the difficult points and issues encountered in the evaluation of governance programmes are outcome measurement, small sample size, presence of spillovers, lack of baseline data, evaluation at higher levels of government, complex interventions and full-coverage programmes.

Rigorous impact evaluations that have addressed these problems in the context of governance remain scarce. Alternative approaches to resolving these issues in which experience in other sectors is borrowed could therefore prove useful. For instance, outcomes which are difficult to measure, such as corruption, can be captured using perception surveys or some form of direct measurement. Single differences using propensity score matching or triple-difference methods, as well as baseline reconstruction, are options for compensating for the lack of baseline data. Spillovers can be potentially managed by changing the level of randomisation. Higher levels of government can be evaluated using cluster randomisation. Complex governance interventions can be investigated by unpacking the intervention into smaller dimensions and using randomised factorial design to analyse combinations of interventions. Programmes covering the entire population can be rolled out in stages to enable impacts to be measured. If that is not possible, treatment can be administered randomly with varying intensity. In the case of full programme roll-out and uniform treatment, time-series with repeated measures taken periodically can be used.

What practical lessons can we draw from previous experience?

The debate on the application of rigorous impact evaluation and the challenges it poses has implications for the design and implementation of governance interventions. The examples and challenges highlighted call for the discussion of several aspects:

Include the evaluator at the outset. The presence of an evaluator in the early phases of programme design can significantly increase the potential of rigorous impact evaluation (RIE). For practitioners, this means obtaining advice from technically skilled evaluators from the very start of the programme. However, including evaluators at the outset is easier said than done. As most evaluators are consultants hired externally by donor agencies, keeping them on for the entire duration of the programme is not economical. An alternative approach is to include the evaluator as a member of an advisory committee before project implementation, so that appropriate outcome measure and data collected have a better chance of remaining valid until the evaluation has been completed.

Randomise whenever possible. One of the most equitable ways of allocating interventions is through randomisation. Critics would say this is “unethical” since it results in benefits being deliberately withheld from those in greatest need. Nevertheless, if random allocation of interventions is done properly, it offers the simplest way of evaluating a programme.

Collect good-quality baseline data. There is a difference between collecting baseline data and *quality* baseline data. Quality baseline data come from careful planning of the programme. This is typically difficult to plan *ex ante* as some indirect outcomes are revealed only during or after implementation.

Collect data for the control groups. Funding the intervention does not neces-

sarily include the funding of the collection of control group data. There are cases where funding for control group surveys is not included in the programme. As the control group is a very important source of information in impact evaluation, it is important to provide a budget for this aspect.

Cooperate with the relevant government agencies. The involvement of government agencies in development programmes varies widely. Regardless of the depth of their involvement, the cooperation of the relevant agency must be ensured. Government agencies are often important sources of information on the nature of the target beneficiaries. Being the local experts, they can help to identify the likelihood of programme take-up.

Identify components of the programme that permit RIE. It is important to identify from the outset potential components of

the programme that permit quantitative analysis. Depending on the context, the inputs can be matched and combined in different ways to shed more light on their marginal contribution.

Next steps

As the next items on the research agenda, this paper suggests, firstly, a comprehensive scoping study of governance interventions that have been evaluated rigorously to identify evidence gaps which an evaluation might seek to close and, secondly, a systematic evaluation of a specific governance intervention to build on existing knowledge of ‘what works’, ‘what doesn’t’ and ‘why’. A few isolated evaluations are not enough to send a strong message on the usefulness of certain interventions. The effectiveness of interventions should therefore be tested under various conditions and in various settings.

1. Introduction

A nation's quest to alleviate poverty hinges crucially on its commitment to 'good' governance. Corruption, political instability, financial mismanagement, conflict, malfunctioning legal systems and human rights violations are extremely destructive and have the potential to cripple a country's socio-economic capacity and development. Consequently, 'good governance' has become a core theme of development policy over the years. The international donor community has responded to what it perceives as a lack of good governance by investing billions of dollars in development assistance targeted at helping partner countries to achieve a well-functioning government and to create a favourable institutional environment. In fact, over the past decade, the world has seen a remarkable growth of development aid under governance-related programmes. In 2008 alone, committed governance programmes in the form of grants, loans and technical assistance reached 18.62 billion US dollars, a three-fold increase compared to 1995.¹ The number of active donor agencies in this field has also grown.² The proliferation of donors engaged in the governance agenda, along with the need to develop strong institutions, has been partly driven by donor disbursement pressures, the controversies surrounding aid allocation and the active debate on new aid modalities (Booth 2008).

Governance is also used as a ground for implementing new aid modalities. Decades of research on the effectiveness of traditional of project-based aid suggest weak impact and lack of ownership from partner countries. However, the discussion on whether new aid modalities, such as general budget support (GBS), are more effective than project-based aid persists. This is because, as briefly described here, the practice of GBS is accompanied by opportunities and risks (for more details, see Leiderer 2010). On the one hand, by directly transferring funds from donors to partners, GBS gives recipient governments more freedom in managing their development programmes. They have the opportunity to support and implement their projects according to their national development goals and poverty reduction strategies. On the other hand, GBS is subject to such obvious fiduciary risks as fungibility, corruption and transparency issues.

Governance therefore plays two critical roles in budget support: first, donors use it as one of their allocation criteria, meaning that governance is one of their considerations when they are deciding whether or not to provide budget support. To qualify for GBS, most budget support donors implicitly or explicitly require recipient countries to demonstrate credible policy structures and adequate levels of good governance (BMZ 2008; European Commission 2010; World Bank 2006). Second, some donors intend governance to be one of their objectives (although the importance they attach to it varies). The aim of GBS is to strengthen state institutions and support activities that improve accountability, public financial management, public participation and political dialogues. Ultimately, the success

¹Source: Author's calculation using disaggregated project level data from AidData accessed on 23 June 2010. All amounts refer to commitments by all donors as reported in AidData, using constant 2000 prices. The amount committed in 1995 was 4.8 billion US dollars.

² They include the World Bank, the UK Department for International Development, the United States Agency for International Development, the Swedish International Development Co-operation Agency, the German Federal Ministry for Economic Cooperation and Development and, recently, the Millennium Challenge Corporation.

of GBS depends partly on the prudent exercise of power and responsibility by the government.³

The role of governance in GBS as a prerequisite and objective is an illustration of the major part that governance plays in modern development policy and cooperation. In general, there is a broad consensus among donors regarding the consideration to be given to governance in the overall allocation of aid. This notion stems from the World Bank's 1998 landmark publication *Assessing Aid: What works, what doesn't, and why*. It argues that aid can have a significant impact on development only if recipient countries rise above certain thresholds of "good" governance. In fact, allocating aid to countries with a favourable institutional environment has become an integral part of donors' strategic policies (Collier / Dollar 2004; IMF 1997; MCC 2010; USAID 2004; World Bank 1998).

In theory, governance programmes, as a target of development aid, offer significant opportunities to improve institutional quality and political stability. Examples of these programmes are interventions to combat corruption, promote democracy, improve public services, ensure a fair election process and increase civic participation. Yet, despite the efforts of donors in financing governance programmes, relatively little is yet understood about their impact. For practitioners and policymakers, evaluation could provide important lessons yet the increasing demand for it is still unmet. For one thing, it is essential to know "what works?", "what doesn't?", and "why" so that future interventions can be improved to ensure better results. Even though there is a growing consensus in the international community on the use of the qualitative and quantitative "mixed method" approach, the effectiveness of these programmes has relied mostly on descriptive case studies and anecdotal evidence. The desirability of adopting experimental or quasi-experimental approaches to evaluation is due to their ability to answer the question, "What would have happened without the intervention?"⁴

Rigorous impact evaluations are probably rare because, first, governance programmes are often characterised by complex interventions routinely combining several activities and producing outcomes that are difficult to measure. Second, the topic lies outside the realm of most rigorous evaluators (Blattman 2008). Third, there appears to be a suspicion that only interventions producing statistically significant results have been publicised, while negative and insignificant results have remained undocumented (Duflo et al. 2007, for instance, give a thorough explanation of publication bias in experimental and non-experimental evaluations). Fourth, such evaluation in governance is often associated with politically sensitive issues and is therefore difficult to undertake. And finally, there is a

³ To increase the effectiveness of budget support, donors also include non-financial inputs in the form of conditionality, technical assistance, policy dialogues and capacity-building. Achieving the goals of budget support set under the national development programmes or poverty reduction strategy (PRS) depends on factors on both sides of the aid-delivery chain. On the recipient's side, they include ownership, quality of strategies and commitment, while on the donor side, they consist of the quality of the non-financial contributions mentioned above. Direct financial assistance serves to "strengthen the effectiveness of non-financial inputs by acting as leverage for reforms, improvements of policy content, governance, and public financial management" (Leiderer 2010, 4).

⁴ The terms "quantitative" and "rigorous" should not be used as synonyms. The use of quantitative methods does not automatically introduce appropriate rigour into an analysis. The term "rigour" means that the analysis has credibly established a valid counterfactual, or "what would the outcome have been in the absence of intervention," through the use of experimental or quasi-experimental designs. To achieve this, the evaluation usually requires the use of certain quantitative evaluation methods.

shortage of reliable baseline and longitudinal data. As a result, evidence of effectiveness and of the lessons learnt is rather limited.

The paper revisits these issues by attempting to answer three key questions: first, what are the features of governance interventions that make rigorous impact evaluation difficult and challenging? Second, what aspects of governance have been evaluated by quantitative methods? And third, what evaluation lessons can we learn from previous experience or what practical implications does it have. While much of the review of the literature focuses either on evaluation methods and applications in general (Angrist / Pischke 2009; Imbens / Wooldridge 2009; Khandker et al. 2010; Ravallion 2008) or on randomised control trials in specific fields (examples being Glewwe / Kremer 2006 on education, Moehler 2010 on democracy promotion), few studies, as far as the author knows, have yet reviewed the quantitative methodologies that have been applied to governance evaluation (one example being Bollen et al. 2005).

This paper attempts to close this gap (i) by providing an overview of the experimental and quasi-experimental approaches that have been adopted to evaluate governance programmes and (ii) by considering some implementation issues associated with them. It is mainly a “methods” paper, elaborating on what has been done by Caspari and Barbu (2008) in a previous study and, of course, by others. It does not attempt to provide a complete review of all governance programmes, but rather to indicate when and how various rigorous impact evaluation techniques might be used. It will generally focus on the evaluation of specific interventions or components of a larger programme, whether government or donor interventions. The scope of the review will be limited to existing literature, typically academic or technical articles and/or manuscripts produced in the last fifteen years. The paper is primarily aimed at audiences working on governance topics who are interested in conducting impact evaluation in this field. Some parts of the paper are technical, but every effort is made to give examples to ensure that the reader grasps the main ideas. The whole agenda and the examples are largely research-driven, but important lessons can nevertheless be learnt from them to somehow bridge research and practice.

The paper suggests that there is considerable scope for evaluating certain types of governance interventions rigorously, perhaps far more than one would initially think. Admittedly, the applicability of quantitative techniques is subject to some limitations, but the interventions can be broken down into components suitable for this type of analysis.

Evidence suggests that some aspects of governance programmes have been subjected to rigorous impact evaluation. They concern accountability and corruption (Bertrand et al. 2007; Grimes / Wängnerud 2010; Olken 2007; Reinikka / Svensson 2005), elections (Collier / Vicente 2010; Gerber et al. 2008; Gerber / Green 2000; Imai 2005), participation (Capuno / Garcia 2010; Humphreys et al. 2006), empowerment (Chattopadhyay / Duflo 2004), service delivery (Duflo et al. 2005) and the rule of law (Ruprah 2008). However, this paper argues that the number of evaluation studies undertaking this type of analysis is still small. Consequently, there is very little evidence from which to infer with any confidence the effectiveness of certain reforms and interventions, given that they have not been replicated or evaluated under different conditions.

This study also suggests ways of overcoming recurrent obstacles encountered in the evaluation of governance programmes. They include a small sample size; outcome measurement; presence of spillovers; lack of baseline data; evaluation at higher levels of government; complex interventions; and full-coverage programmes. These problems are not new in the evaluation literature. In such heavily evaluated sectors as education, labour

and health these challenges have already been tackled. This paper thus borrows from the lessons and experience of these sectors.

The paper is structured as follows. Chapter 2 discusses governance as a core aspect of development cooperation. Chapter 3 explains the fundamental impact evaluation problem. Chapters 4 and 5 explain the standard experimental and quasi-experimental approaches and give an example in the governance context. Chapter 6 describes some of the main challenges encountered in the evaluation of governance programmes. The examples will be derived from heavily evaluated sectors (such as labour, education and health) that have experienced similar issues. Chapter 7 presents some practical implications for the design of governance programmes. Chapter 8 concludes the study.

2. Governance as a core aspect of development cooperation

2.1. The concept of governance

The concept of governance is complex, unbounded and intangible. Its operations vary widely – from national to subnational government, from institutions to corporations and individuals. Donors, practitioners and researchers have developed their own definitions, which ultimately reflect part of their own research or political agenda, as the examples below show. Table 1 also presents some definitions of (good) governance from donor countries.

- The BMZ (2009, 6) refers to governance as “the way decisions are taken and policies are framed and implemented in a state. It also includes political processes at supranational level and relevant regional organisations. The focus is on norms, institutions and procedures that regulate the actions of governmental, non-governmental and private-sector players. On the one hand, it is about the values that underlie governance and, on the other, about the institutional frameworks in which governance takes place. The normative and institutional dimensions of governance can only be understood in the light of the specific historical, cultural, social and economic context.”
- The World Bank (1992, 1) defines governance as “the manner in which public officials and institutions acquire and exercise the authority to shape public policy and provide goods and services.”
- The IMF (2007, 128) defines it as “the process by which decisions are made and implemented (or not implemented). Within government, governance is the process by which public institutions conduct public affairs and manage public resources.”
- A UNDP policy paper (1997) defines it as “the exercise of economic, political and administrative authority to manage a country’s affairs at all levels. It comprises the mechanisms, processes and institutions, through which citizens and groups articulate their interests, exercise their legal rights, meet their obligations and mediate their differences.”
- Kaufmann, Kraay and Zoido-Lobaton (1999, 1) define governance as the “traditions and institutions by which authority in a country is exercised. This includes the process by which those in authority are selected, monitored and replaced; the capacity of the government to effectively formulate and implement sound policies,

and the respect of citizens and the state for the institutions that govern economic and social interactions among them.”

- The Institute of Governance (2010) adopts a broader definition: “governance determines who has power, who makes decisions, how other players make their voice heard and how account is rendered.”

Table 1 Country definitions of governance and good governance

Country	Definition of (good) governance
AUSAID	‘Good governance’ means competent management of a country’s resources and affairs in a manner that is open, transparent, accountable, equitable and responsive to people’s needs.” (AusAid 2000, 3)
ADC	“In the context of a political and institutional environment that upholds human rights, democratic principles and the rule of law, good governance is the transparent and accountable management of human, natural, economic and financial resources for the purposes of equitable and sustainable development.” (ADC 2006, 5)
CIDA	"Good" governance is the exercise of power by various levels of government that is effective, honest, equitable, transparent and accountable.” (CIDA 1999, 21)
DFID	“Governance is about the use of power and authority and how a country manages its affairs. This can be interpreted at many different levels, from the state down to the local community or household.” (DFID 2007, 6)
SIDA	“Good governance and good public administration are important aspects of democracy. These concern the management and distribution of public resources, equality before the law and procedures to combat the abuse of power. ” (SIDA 2003, 24)
USAID	“Governance issues pertain to the ability of government to develop an efficient and effective public management process. Because citizens lose confidence in a government that is unable to deliver basic services, the degree to which a government is able to carry out its functions at any level is often a key determinant of a country’s ability to sustain democratic reform.” (USAID 1998, 19)

Source: Some parts extracted from UNDP (2007, p. 4)

The definition of the term ‘governance’ has evolved over time. A UNDP report (2007, 3) classifies the concepts in three broad categories:

“(1) governance, which is the most neutral and refers to sound public financial management; (2) good governance, which retains the economic and financial elements, but adds elements of accountability and transparency of decision-making, and the rule of law, especially the protection of property rights and respect for contracts; (3) democratic governance, which retains the elements of the previous definitions, but adds elements of democracy (especially horizontal and vertical accountability) and respect for human rights (civil, political, social and cultural).”

In this paper, the term governance will generally focus on what the government does, encompassing all three of the above categories. Although many distinctions are made between them, this paper will not go into the conceptual debate in any depth.

Table 2 Governance dimensions, topics and activities

Dimensions	Topics	Example of interventions
Political system	Democracy promotion (elections) Human rights Conflict Rule of law (judicial and legal development) Decentralisation	Construction of ministerial building; demobilise combatants and support their reintegration into civil life; freeing imprisoned child soldiers; training courses and workshops for officials of human rights institutions; training of court judges; encouraging youths to vote; supporting electoral procedures, such as computerised voting systems; supporting the enactment of decentralisation laws; strengthening competitive party systems in emerging democracies
Public administration	Corruption Public management Public financial management Public procurement Tax policy Transparency Fiscal control mechanisms	Annual audits of national entities; contributing to reallocation of public resources from military to socio-economic sectors; creating a regulatory framework to improve treasury management; capacity-building in the accounting units by training X accounting officers; improve financial transactions by computerising revenue collection and expenditure management; creating anti-corruption agencies.
Social governance	Efficient public service delivery Citizen empowerment (e.g. women, voice, participation) Community development	training and workshops for local CSOs*, women and elders; Community audits;
Market governance	Economic policy and planning Business environment	Training support, conferences and seminars

Source: Columns 1 and 2 are from OECD report (2008, p. 19), modified.

For the systematic organisation of the wide range of topics in the three categories, this paper adopts a modified version of the OECD's grouping (2008, 19) for governance assessment. Table 2 groups topics according to four broad dimensions, namely: political system, public administration, social governance and market governance. Within each dimension, 'political system' comprises democracy promotion, human rights, conflict, rule of law and decentralisation. 'Public administration' covers corruption, public management, public financial management, public procurement and tax policies. 'Social governance' consists of service delivery, citizen empowerment, community development and voice. And lastly, market governance involves the creation of a favourable business environment

and economic policy and planning. The third column contains examples of typical donor interventions within each governance dimension.

It is important to note that this categorisation is derived not from theory but rather from a survey of governance assessments made by donor agencies. Since the paper focuses on governance interventions, this classification serves the purpose of the paper. However, this is likely to change if some form of theoretical approach is adopted.

It should also be noted that each dimension may have overlapping topics. Projects on corruption, which is covered by 'public administration,' may have a rule-of-law component (under 'political system'). This list is not therefore intended to be definitive or exhaustive, but rather to act as a framework for the discussion, given the breadth and depth of the concept.

2.2. Aid for governance

Governance is at the heart of the aid effectiveness debate. It is both an objective of and condition for development assistance. In practice, donors have often used a combination of loans, grants and technical assistance to promote good governance. With the help of AidData,⁵ one of the most comprehensive aid databases, this section seeks to provide some insight into donor interest in the field of governance.

Table 3 shows the amount committed by donors to projects whose dominant sector is classified as governance. It reveals that donor interest moved from market governance in 1995 to political systems in 2008. It also demonstrates that 'political systems,' comprising such thematic issues as elections, human rights, conflict and legal and judicial development, has grown tremendously over the last decade, having accounted for 5.24 per cent of total governance assistance in 1995 and 38.63 per cent in 2008. The share of projects devoted to public administration peaked in 2005 and slowly declined thereafter (from 29.6 per cent in 2005 to 16.11 per cent in 2008). Nevertheless, it continued to have a modest share of development projects. Projects involving women, civil society and access to information have consistently captured about 10 to 14 per cent of governance aid since 1995.

Table 4 shows that development assistance was increasingly focused on more fragile, emerging conflict and post-conflict states, as evidenced by the larger share of projects classified under 'conflict' (rising from 1.51 to 16.31 per cent). The smallest share went to elections, public financial management and human rights.

As "governance" has become more important, name-changing or reclassification by donors may have taken place. For instance, what was classified as "education at subnational level" can be classified under 'decentralisation.' This highlights one of the problematic issues in aid data collection. Others are unclassified items, reporting errors and redundant accounting. Nevertheless, the initiative of donors and researchers in creating and improving a standardised project-level aid database to shed light on past and present trends in development assistance is already a huge leap forward. But more needs to be done.

⁵ Findley, M., Darren Hawkins, R. Hicks, D. Nielson, B. Parks, R. Powers, J. T. Roberts, M. Tierney & S. Wilson (2009) AidData: Tracking Development Finance, presented at the PLAID Data Vetting Workshop, Washington, DC, September

Table 3 Donor assistance in governance

Topic/Year	1995	2000	2005	2007	2008
<i>Political system</i>	5.2%	23.2%	25.5%	31.7%	38.6%
Conflict	1.5%	14.1%	11.2%	16.2%	16.3%
Elections	0.9%	1.1%	3.7%	2.5%	3.1%
Human rights	1.9%	4.1%	4.9%	4.2%	8.1%
Rule of law	1.0%	4.0%	5.7%	8.8%	11.2%
<i>Public administration</i>	19.4%	14.9%	29.6%	22.4%	16.1%
Public-sector financial management	1.0%	2.3%	2.2%	3.9%	5.0%
Government administration	18.3%	12.7%	27.4%	18.6%	11.1%
<i>Social development</i>	11.4%	9.8%	10.2%	11.5%	14.5%
Women	2.1%	1.1%	1.1%	1.9%	2.5%
Civil society	9.2%	8.2%	8.6%	8.2%	10.8%
Access to information	0.1%	0.5%	0.5%	1.5%	1.2%
<i>Market/economic governance</i>	33.9%	8.7%	12.6%	10.7%	9.8%
<i>Others (not classified)</i>	30.1%	43.3%	22.1%	23.6%	21.0%
TOTAL (in billion USD, constant 2000 prices)	4.86	7.32	12.93	15.30	18.62

Source: Author's calculation derived from AidData, accessed 23 June 2010. The figures reflect amounts committed by donors.

Notes:

1. As unclassified governance aid is quite substantial, commitment amounts in each dimension may have been over- or underestimated.
2. Owing to the huge amount of information that needs to be collected, some projects may have been unintentionally omitted.
3. A project may appear in the database more than once if multiple donors are involved, the donor commits new money or the donor splits the projects into several activities. Since there is no unique project identifier, this may cause overestimation of the number of development projects (Findley et al. 2009).
4. The amount reflected in these data does not include general budget support, much of which is devoted to improving such governance structures as public financial management.

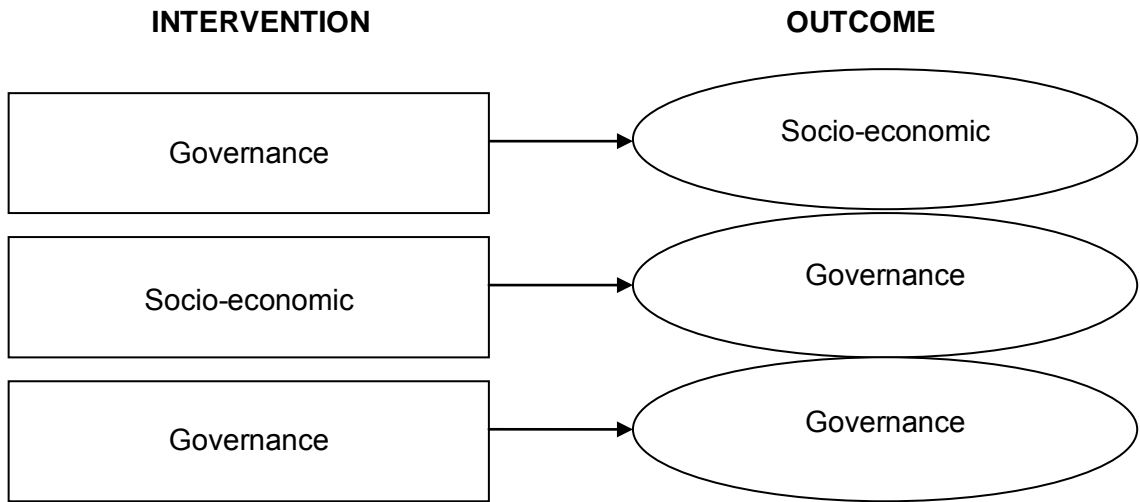
In sum, the current trend shows that investments in governance programmes have greatly increased over the last decade. Yet, despite the efforts of donors and the international community to finance governance programmes, relatively little is yet understood about their impact. The changes in the aggregate measurements of governance established by many index producers and the amount of money spent may be correlated and may substantiate donor efforts, but they are not sufficient proof of a tangible impact. The scale of investment presupposes careful evaluation of the impact of governance programmes, so that questions regarding its effectiveness and what could be done better in the future may be addressed.

2.3. Design of governance programmes

Donor assistance to the governance agenda typically combines policy advice, technical assistance, grants and loans. These interventions are designed to achieve certain outcomes, such as decreasing corrupt practices, increasing voter turnout and improving the delivery of public services. This agenda and those adopted for other sectors are not mutually exclusive. Some features of anti-poverty, education or infrastructure programmes embed governance in their design and execution. Governance interventions thus usually complement other sector projects or programmes.

The governance topics mentioned in the previous section come in three forms when evaluated: (i) as an intervention aimed at achieving a certain development outcome, (ii) as an outcome of an intervention or (iii) both. Thus the positioning of governance in the logical framework of inputs and outcomes is not fixed. Of course, it may also cover the whole range from inputs and processes to outputs, but for the purpose of this discussion a simplified version of input-outcome will be used. Governance evaluation may be examined using three separate linkages (as in Moehler 2010). The first linkage runs from assessing a socio-economic intervention with an expected governance outcome. The second linkage indicates the intrinsic value of governance to development. And the third identifies the link between governance intervention and specific governance goals. Figure 1 below illustrates this.

Figure 1 Simplified design of governance programmes



Linkage 1: Governance intervention with intended socio-economic (or environmental) outcome

In principle, governance reforms should lead to socio-economic development. Welfare should improve because governance interventions are meant to (i) increase the efficiency of public service delivery, (ii) decrease wastage of public funds (iii) empower citizens to make the government accountable, etc.

An example of a governance intervention with social outcomes is given in the study conducted by Reinikka and Svensson (2005). The intervention is what they termed a ‘govern-

ance innovation' in the form of a newspaper campaign serving as an anti-corruption instrument. The aims were to minimise the capture of public funds by district officials and to empower schools to monitor those funds and claim what they were entitled to. The ultimate objective of the campaign was to increase school enrolment and performance. In this sense, governance interventions are instruments for improving specific social and economic outcomes.

Linkage 2: Socio-economic (or environmental) intervention with intended governance outcome

Social programmes may also embed governance outcomes in their design. For example, Grimes and Wängnerud (2010) examined the effect of a conditional cash transfer (CCT) programme in Mexico on corruption. The main aim of CCT programmes is to alleviate poverty by increasing children's school attendance and health visits. However, combating administrative corruption, as in this case, was the secondary aim of the project. Thus socio-economic interventions do not necessarily reflect socio-economic outcomes: they may have direct or indirect governance consequences.

Linkage 3: Governance intervention with intended governance outcome

Governance could also be an outcome of a programme that seeks to address governance issues directly. Collier and Vicente (2010) investigated the impact of a grassroots anti-violence campaign on voters' behaviour. This intervention involved the following activities: town meetings, street theatres and the distribution of campaign materials. It is hoped that this will combat voter intimidation by decreasing the perceived threat to individual voters.

3. Impact evaluation

The ultimate goal of impact evaluation (IE) is to determine the extent to which observed outcomes can be attributed to the programme, and to the programme alone. According to the definition of the International Initiative for Impact Evaluation (3ie), IE measures "the net change in outcomes amongst a particular group, or groups of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question(s) being investigated and to the specific context" (3ie s.a., 1). Similarly, the World Bank's Independent Evaluation Group (IEG s.a., 1) defines IE as "the systematic identification of the effects, positive or negative, intended or not, on individual households, institutions, and the environment caused by a given development activity such as a program or project." IE provides information on whether the programme has had an impact and on the magnitude of that impact. Because of this, it is an important source of information for policymakers and development institutions seeking to justify the implementation and expansion of a programme.

The IEG lists four impact evaluation models: (i) rapid assessment or review, conducted ex-post, (ii) ex-post comparison of project beneficiaries with a control group using multivariate analyses, (iii) quasi-experimental design using matched control and treatment groups and (iv) randomised design. The IEG classifies the last two models as *rigorous impact evaluations*. It emphasises that "the strong advantage of these two methods is that they are the most reliable for establishing causality – the relationship between a specific intervention and actual impacts – and for estimating the magnitude of impact attributable to the intervention" (IEG s.a., 3).

A broader (though similar) description of rigorous impact evaluation is provided by 3ie: “rigorous impact evaluations are those which tackle the attribution problem. The main challenges to be addressed in attribution are: (1) allowing for confounding factors, (2) selection bias arising from the endogeneity of program placement, (3) spillover effects, (4) contamination of the control and (5) impact heterogeneity” (3ie s.a. a, 1).

Common to these descriptions is that RIE entails the inclusion of *quantitative* elements in the design of the evaluation, as well as qualitative analysis to tease out and validate the effects of the intervention. As will be argued in subsequent sections, addressing this attribution problem adequately usually entails the use of experimental or quasi-experimental approaches.⁶

The need to enhance development effectiveness has prompted a tremendous growth of interest in impact evaluation studies in recent years. At the same time, a better understanding of the technical difficulties of the appropriate attribution of outcomes has led to increasing calls for more ‘rigorous’ impact evaluations to address the problem of causality.

The push towards evidence-based development policy has consequently led to many rigorous impact studies being conducted in various economic and social sectors. Heavily evaluated programmes have been implemented in such subject areas as labour (Angrist / Krueger 1999), health and nutrition (Gaarder et al. 2010; Habicht et al. 2009), water and infrastructure (van de Walle 2009; Waddington et al. 2009), agriculture (Duflo et al. 2008) and education (Glewwe / Kremer 2006). By comparison, relatively little attention has been paid to the rigorous evaluation of governance. Despite the challenges posed by and limited experience of the evaluation of governance, rigorous evaluations of some promising aspects have already been undertaken. Moreover, the international community and donor agencies⁷ appear to have made significant efforts to bridge this gap in recent years.

3.1. Applying the general principles

Because of the growing popularity of quantitative approaches to solving the problem of attribution, donor agencies are under mounting pressure to comply. However, such approaches should not be overdone. Both evaluators and donors should establish whether or not a quantitative approach is feasible. As a rule, it is important first to identify the evaluation questions and to understand the design of the programme before the search for an appropriate method begins, and not vice versa (searching for questions that can be addressed by a method).

Although generalisation is often not possible, the advantage of using a qualitative approach is that it provides a good contextual basis, which the other approach frequently lacks. White (2006) argues that there should not be a trade off between quantitative and qualitative approaches. Qualitative data provide context and appropriate interpretation of

⁶ There are general debates regarding the roles of quantitative and qualitative impact evaluations as well as methodological disputes over the applicability/accuracy of quasi-experimental approaches over randomised designs. This study will not cover these debates: see White (2006) for a thorough discussion.

⁷ Some of the organisations that actively conduct rigorous governance evaluations are the Jameel Poverty Action Lab (JPAL), the UK Department for International Development (DFID), the United States Agency for International Development (USAID), the Swedish International Development Cooperation Agency and, recently, the Millennium Challenge Corporation. The International Initiative for Impact Evaluation has funded several evaluations in India, eastern Congo and Sierra Leone in 2010, <http://www.3ieimpact.org/openwindow/round2/>.

quantitative results. In fact, the combination of the two, known as the mixed methods approach, should produce “the strongest evaluative findings, combining well-contextualized studies with quantitative rigor” (White 2006, 2). Box 1 sets out important standards to be observed in impact evaluation.

Box 1 Principles of theory-based impact evaluation

The aim of theory-based impact evaluation is to determine why an intervention has had an impact, rather than knowing only that it has had one. The six steps in the successful adoption of this approach are as follows:

- (1) **Map out the causal chain (programme theory).** This step involves constructing a detailed flow chart of the causal chain from inputs to outcomes and impact. It seeks to test the underlying assumption embedded along the causal chain, also taking into account the changing dynamics of the intervention and of unintended impacts.
- (2) **Understand context.** Context is defined as the socio-economic and political setting in which an intervention takes place. Apart from revealing the factors that may explain why an intervention has had an impact, it plays an important role in indicating how similar interventions may have different impacts. This step requires the reading of project documents and of more general literature on anthropology or political economy.
- (3) **Anticipate heterogeneity.** This means that the evaluator must be aware of the possibility of an intervention having various impacts. The differences may be due to the social and political setting, the behaviour of the target groups, the existence of other interventions or the design of the intervention itself.
- (4) **Rigorous evaluation of impact using a credible counterfactual.** The construction of a credible counterfactual, or control group, involves the use of an experimental and quasi-experimental approach to tease out the effect of the treatment. This was briefly discussed earlier and is the core topic of the following chapters. It is a key aspect in theory-based impact evaluation.
- (5) **Rigorous factual analysis.** Apart from counterfactual analysis, factual analysis is needed to confirm whether the intervention has reached the targeted groups and whether it has actually changed their behaviour. This type of question reveals any potential breakdown in the causal chain that could lead to low impact.
- (6) **Use mixed methods.** This refers to the combination of quantitative and qualitative approaches in the same evaluation. The qualitative aspect involves a wide range of activities, including the reading of project documents, focus group discussions, literature review and field work.

Source: White (2009)

The application of experimental and quasi-experimental methods to solve the problem of attribution is subject to general limitations.

- First, they require good data, usually drawn from surveys.
- Second, the allocation of the treatment must be designed in certain ways, which may not be practicable.
- Third, a relatively large sample size is needed, which means that the unit of analysis must be large enough for statistical tests to be conducted.
- Fourth, they may not be applicable to certain interventions, this being especially true of reforms at national level.

In practice, quantitative analysis may have to be restricted to certain components of the programme, especially where interventions are highly complex. In such cases, unbundling various activities and measuring their impact will be one option.

This paper focuses on the fourth principle in Box 1, where impact is measured with the aid of a credible counterfactual. This does not imply that other principles must be neglected. In fact, the validity of findings may be greatly reduced if the focus is limited to this aspect only and context specificity is omitted. However, for the sake of brevity, the case studies presented in the following chapters will focus solely on impact measurement. For more details of the overall project aim, intervention logic and theory of change, the original papers should also be consulted.

3.2. The problem of attribution

Impact evaluation analyses the impact of an intervention on welfare outcomes. Intervention may take the form of policies, programmes or projects. In reality, changes in outcome may be only partly due to the intervention, and sometimes not at all. Thus, the fundamental problem with evaluation is how to establish attribution, i.e. to determine that the outcome is the result of the intervention and not of any other factors. It raises the issue of the counterfactual, “the comparison of what actually happened and what would have happened in the absence of intervention” (White 2006).

Formally, impact is the difference in outcome Y , with Y_1 denoting the outcome if a person is exposed to the intervention, and Y_0 is the outcome if he/she is not.

$$impact = Y_1 - Y_0 \quad (1)$$

At a given point in time, it is possible to observe only the outcome of the person being exposed to treatment, but not the outcome of his/her not being exposed. In other words, failure to observe both states at the same time poses a dilemma for equation 1. The main challenge to impact evaluation is therefore to find the *valid counterfactual*. In the search for a valid counterfactual, the two common comparison groups often considered, but insufficient if considered separately, are (i) data on the same individuals *before and after* the intervention and (ii) data on a group of individuals who participated in the programme and another group who did not or, in short, *with and without intervention*.

To illustrate the problem of impact evaluation and why these two comparison groups are not valid counterfactuals, the following is a simplified hypothetical scenario:

Imagine that developing country X suffers from widespread corruption and bribery in the enforcement of traffic regulations. Unlike most developed countries, any traffic violation in country X requires the motorist concerned to surrender his licence. Issuing tickets for traffic violations is often ineffective unless the fine is collected on the spot because of the poor information-tracking system that is a common feature of developing countries. Traffic enforcers (i.e. the police) accept and often solicit bribes to compensate for their low wages, while motorists offer bribes to avoid fines, the confiscation of their licences, long queues to recover their licences and, in the case of serious violations, mandatory seminars. Thus bribes generate extra income for the police and also reduce transaction costs for motorists. The Traffic Enforcement Authority (TEA) is a public agency responsible for enforcing traffic regulations. It has proposed to the president of country X that the problem of corruption and bribery among traffic enforcers should be solved. The inception report contains several hundred pages of proposed interventions. Of the many interventions proposed by the TEA, the following appeal to the president:

1. Set up a complaints system enabling motorists to identify erring policemen by text message.
2. Launch a quota system for traffic enforcers requiring them to catch and report 10 traffic violators per month. The number of violators caught will be linked directly to the enforcers' performance assessment.
3. Set up a rapid licence retrieval system that enables motorists to pay by credit card and the license to be sent to the address they specify.
4. Conduct corruption awareness programmes by training traffic enforcers.

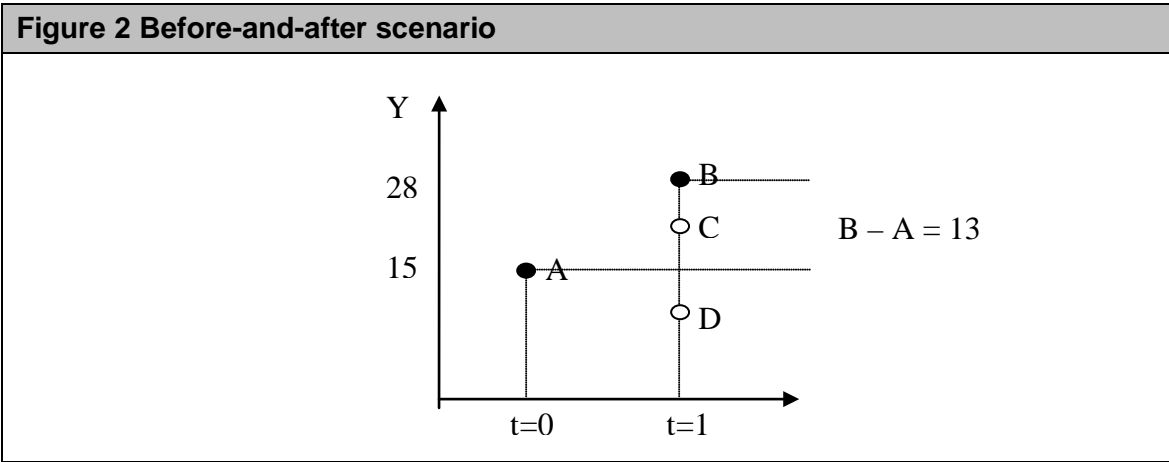
The president and the TEA appreciate the difficulty associated with quantifying governance outcomes and acknowledge that each intervention may require a different set of data – provided by the policeman and/or the motorist. The survey data on the police will provide information on their self-assessed performance, while data from motorists will reveal what the public think of police performance. These two sets of data can show whether and how the police perception and the public perception match. After careful consideration of the pros and cons of each intervention and of the limited budget, the group finally decides to launch a quota system for traffic enforcers. In this way, the police are obliged to increase their efforts and to report their results (in respect of the 10 violators at least), which may help to reduce the incentive to accept bribes. Although this may not rein in corruption altogether, the policymakers believe that the improvement of police performance will reduce inefficiencies in service delivery to a minimum.

Partly due to budget constraints the TEA decides to begin by gathering survey data on police performance (and to conduct the public survey later). To demonstrate the effectiveness of the intervention, the TEA selects 30 policemen whom they apply the quota system. They measure the success of the programme by comparing the number of reported violations the year before the quota system was introduced with the number of reported violations the following year. They find that the quota system increases the number of reported violations by an average of 87 per cent. From this result, the TEA concludes that the programme is effective.

The president wisely communicates her doubts about the calculation. Impacts of governance interventions are very likely to be confounded by unexpected political developments, such as the passing of a new law and elections and economic shocks such as food shortages, the discovery of oil or a drought. If only reported violations before and after are con-

sidered, the results will be biased owing to the failure to take account of time-varying factors that could affect the outcome over the period. This means that the factors responsible for the change cannot be attributed to the intervention alone, other plausible explanations possibly including the passage of time or the sudden occurrence of an event. The TEA asks, "Does this mean that the effect of the quota programme is rather small, since the policemen treated would have increased their catches anyway?" "That's possible," replies the president.

The president illustrates her explanation with a graph. The average number of violations reported by the individuals in the programme before the intervention is, for instance, 15 per month ($t=0$, point A). After a year, the average number of reported violations has risen to 28 per month ($t=1$, point B). The impact is equal to 13 calculated as B minus A. But is the increase in the number of reported violations due to the programme alone? That is not very likely. An increase in the number of reported violations following the dismissal of a corrupt high-level TEA official may generate confidence in the system and lead to a rise in the number of violations reported by the police ($t=1$, point C). The real impact is therefore the difference between points B and C, the impact having previously been overestimated. Likewise, promoting a highly corrupt TEA official reduces confidence in the political system and results in a lower number of reported violations ($t=1$, point D). The impact has therefore been initially underestimated, and the real impact is between points B and D. This before-and-after approach fails to take account of time-varying factors that may affect the impact measure.



Source: Author's representation.

To calculate whether the quota system has an impact or not, it must be known how many violations the policemen would have reported if they had not participated in the quota system. Unfortunately, the same policemen cannot be observed in both circumstances. The TEA therefore constructs a comparison group and finds that the participants in the quota system reported, on average, *more* violations than those who participated in the programme. Here again, the conclusion is that the intervention has had a positive impact.

But the president is still not convinced. Why not? She wants to know how the individuals in the comparison group are selected. She thinks it possible that the number of reported violations has increased because the quota group consists of motivated individuals and that the programme therefore works for them, but may not have the same effect on others. The president explains that, to solve this problem, it will be necessary to find a compari-

son group for whom, in the absence of intervention, the outcome would have been similar to the outcome for those who received treatment. Simply choosing a group of individuals who did not participate in the programme does not produce a *valid counterfactual*. In other words, evaluators cannot simply include policemen who do not take part in the programme because there may be underlying differences between the two groups that have not been taken into account. It also needs to be understood how certain individuals are assigned to the quota system. In a situation where policemen volunteer to join the programme, it is important to understand why *they* do so and others do not. Are the participants more motivated than the non-participants? The inherent difference, both observable and unobservable, between the two groups is usually known as *selection bias*. If this is not taken into consideration, the impact estimates could produce misleading results.

Solving the problem

Ideally, policemen should be randomly assigned to the quota system at the start of the programme. Eventually, two groups will be formed: those who belong to the quota group and those who do not. As the number of policemen in both groups increases, the difference in terms of extraneous factors should even out until the only remaining difference between the two is the intervention.

Since it is already too late for randomisation, *combining* the before-and-after approach and with-and-without comparison can itself yield credible estimates. To ensure that the pre-treatment differences in the control and treatment groups are taken into account, statistical techniques for correcting for selection bias, such as propensity score matching and instrumental variables, can be used to improve the impact estimate. (More on this in the next chapters).

The example highlights the importance of arriving at the right conclusion through an appropriate understanding of how impact is measured. From the perspective of both the government and aid agencies, the risk of discontinuing an anti-corruption project which is in fact effective or of institutionalising a policy which is in fact ineffective is too high. The resources and efforts expended on such projects can be justified only by means of impact evaluation.

In general, experimental and quasi-experimental designs seek to address the selection bias due to purposive programme placement or individual heterogeneity stemming from beneficiaries' self-selection to the programme (see Box 2 for a brief description of the quantitative techniques). It is important to note that each technique is accompanied by underlying assumptions of how the counterfactual problem is resolved. Further, the design of the programme usually dictates the most appropriate technique.

Box 2 Impact evaluation techniques

Impact is calculated as the difference in outcome with and without intervention. But since it is not possible to collect data from a single individual or subject in both circumstances – in the presence and absence of intervention – various IE techniques can be applied to circumvent the problem of missing data and to arrive at an unbiased estimate of the impact. The application of these techniques largely depends on how the intervention will be or has been implemented and on the design of the experiment in general.

Randomisation: This approach requires that participants be randomly assigned to treatment. The random assignment generates two groups of participants, the control and the treatment group. As the number of sample participants increases, the difference between the two groups in terms of extraneous factors should even out until the only remaining difference is the intervention.

Instrumental variable (IV): In simple terms, this approach involves identifying a special variable that affects outcome and intervention, but without the two having any causality. IV approach is needed if the causal relationship between intervention and outcome runs the opposite direction. Such a scenario occurs when (i) the intervention has been deliberately targeted or (ii) participants have joined the programme for specific reasons. This implies that potential unobserved characteristics or omitted variables have not been taken into account.

Difference-in-differences (DID): This double-difference approach calculates impact by utilising information before –and after the intervention and calculating the change in outcome over the two periods between the control and treatment groups.

Propensity score matching (PSM): To minimise confounding factors due to the non-random assignment of the intervention, matching requires finding a comparison group that matches the characteristics of the treatment group. The observable characteristics are used to generate a propensity score, which is the probability of participation. Each treated individual is matched to a non-participant on the basis of this propensity score.

Regression discontinuity design (RDD): RDD require a specific eligibility rule in the targeting of participants for the programme. The degree to which the intervention changes the outcome of the treatment group compared to non-participants near the eligibility cut-off is the impact.

4. Experimental or Randomised Control Trials (RCTs)

Randomisation removes the selection bias by randomly assigning individuals to treatment or comparison (control) groups. Since individuals are randomly assigned, the inherent differences between the two groups should, on average, be cancelled out as the sample size increases, and what remains is the effect of the treatment.

In theory, randomisation occurs in two stages. In the first stage a group of individuals is randomly selected from the entire population, and in the second stage these individuals

are randomly assigned to treatment and control groups. The first stage is necessary for external validity (applicability for scaling-up), while the second stage is needed to test for internal validity (Khandker et al. 2010).

Once the samples have been gathered, calculating the impact is straightforward. A simple test of means between the two samples can be made, although a simple regression framework is more appropriate, especially in cases where randomisation is conditional on some observed variables. To illustrate this, a simple regression framework in the case of the two-stage pure randomisation mentioned earlier can be formulated as (Duflo et al. 2007):

$$Y_i = \alpha + \beta T_i + \varepsilon_i \quad (2)$$

where Y is the individual outcome, T_i is a dummy equal to 1 if the individual belongs to the treatment group and 0 if he belongs to the control group. The treatment effect,⁸ which is the difference between the control and treatment group, is captured by $\hat{\beta}$. RCT is known as the “gold standard” in impact evaluation because of its high internal validity, and it is superior to quasi-experimental designs. By randomly assigning treatment, RCT is also considered to be one of the most equitable ways of allocating limited resources.

Corruption

The first example addresses the different monitoring schemes used to reduce corruption which have been promoted by many donor agencies. Some scholars argue that a top-down approach is effective in reducing corruption if individuals are provided with the right balance of monitoring and incentives (Becker / Stigler 1974). The problem with this method is that the very officials who are assigned to do the monitoring may also be corrupt. The alternative approach underscores the power of grassroots participation. Since citizens are the direct beneficiaries of public services, they have the incentive to monitor the performance of their public officials. But this, too, has a disadvantage. Reducing corruption in public affairs is a public good and subject to free-rider problems, which means that lazy citizens cannot be prevented from benefiting from the contribution made by active individuals who monitor corrupt officials. They may not want to share the cost of supervision and may leave the responsibility to others. In the end, nobody has any incentive to monitor. Whether or not top-down and bottom-up approaches reduce corruption therefore requires deeper investigation.

To answer this question, Olken (2007) conducted a field experiment in the context of road projects in Indonesia financed by the World Bank nationwide. Six hundred and eight villages were randomly selected and divided into four groups. Olken informed the selected villages in the first group that the road would be audited by central government. In the second group, he organised village-level accountability meetings. In the third group, he again organised accountability meetings and, in addition, anonymous survey forms were distributed to villagers. The fourth group served as the control group. Since it is difficult to capture corrupt practices, he constructed a simple measure of corruption. Instead of using

⁸ The treatment effect of a programme is usually represented by two aggregate measurements: (i) the average treatment effect on the treated (ATT), defined as the average gain of those who are treated conditional on their receiving treatment and (ii) the average treatment effect (ATE), which is the difference between treatment and control groups randomly drawn from the whole population. In this paper, the treatment effect referred to is the ATT, which will be the main concern in the discussion. See Khandker et al. (2010) and Ravallion (2008) for a further discussion and the derivation of the ATT and ATE.

perception-based corruption measure, he devised a direct measure of corruption as the difference between reported expenditure and the independent engineers' estimate. The independent estimate was derived from the samples taken by the engineers after project completion to measure the quantity of materials used from interviews with villagers to determine the wages paid and from a survey of suppliers carried out to estimate prices. He found that increasing government audits had a positive effect by reducing corruption. "Missing expenditures" fell by eight percentage points when the villages were audited by central government. He also found that grassroots participation in monitoring had no impact. Rapid assessments and M&E approaches to evaluation may lead to a different conclusion. Olken's result is convincing because, through randomisation, he ensured that the only difference between the treatment and control groups was the intervention.⁹

In another example the consequences of corruption are examined. Is corruption efficient? Is it harmful to society (Rose-Ackerman 1978), or does it merely reduce transaction costs by speeding up bureaucratic processes (Huntington 1968)? Understanding the implications of corruption can provide clues to the anti-corruption strategies that may improve the efficiency and fairness of governments.

Bertrand et al. (2007) set out to answer these questions by following a sample of candidates applying for driving licences in India. Eight hundred and twenty-two candidates were tracked, all of them eventually taking a surprise independent driving test. The authors wanted to know whether corruption can speed up the application process and whether it helps individuals unqualified to drive to obtain a licence. To this end, candidates were randomly assigned to one of the three groups. The "bonus group" received a financial bonus if they obtained their licences quickly (within 32 days). The "lesson group" was offered free driving lessons. And finally, the third was a control group that was simply monitored throughout the process. The results of the experiment showed that individuals in the bonus group were 24 percentage points more likely to obtain a licence than individuals in the control group. They were also likely to obtain a licence without taking the driving test and more likely to obtain a licence but fail the independent driving test, which shows that corruption may be harmful to society. The lesson group was only 12 percentage points more likely to obtain a license than the control group. This suggests that being a good driver is not necessarily a guarantee of obtaining a licence. In fact, individuals in the lesson group also made additional payments to "agents" despite being better drivers. Like Olken's approach, this experimental method allowed the authors to disentangle the impact of the intervention from other factors, whereas simply conducting a survey and asking individuals whether or not they had made extra payments to obtain a licence might have led to the conclusion that no corruption had taken place. Careful randomisation led the authors to the opposite conclusion.

Democracy promotion

The third example shows how RCTs can be used to evaluate interventions in democracy promotion. Democratic processes in developing countries are weak because some voters are less well informed or because of elite capture. Elections often feature ballot fraud, violence and vote-buying. To minimise this, donors have deployed international observers to new democracies to monitor the election process. While some argue that international

⁹ Note that this is a one-period intervention and results may change over time. For instance, community participation, when strongly cultivated over time, may *actually* have an impact in the long run.

observers help increase voters' confidence and discourage fraud, observers are not necessarily objective and may carry with them the political thrust of their own countries. Whether outside assistance helps or harms the election process is a very important empirical question.

To capture the impact of international election observers on election quality, Hyde (2010) randomly assigned international observers during the 2004 presidential election in Indonesia to villages identifiable on a local map, in each of which they visited one to four polling stations. The random assignment generated two groups of villages, those that were observed and those that were not. Her results show that, as international observers tend to increase the total number of votes cast for the incumbent, they may change voters' election-day behaviour disproportionately. She also found that election officials tend to take greater care to comply with election regulations if international observers are present.

Another experimental study of democracy promotion focuses on clientelism and vote-buying in developing countries. Clientelism consists in personal and material favours done by candidates for an individual voter or group of voters in exchange for political support. Do voters respond to the offer of favours more than public policy messages? Public policy messages are nationally oriented and concern such aspects as public health and child welfare. Are female voters more likely to respond to offers of favours? In most developing countries, candidates rely on personal favours to obtain political support. Whether this is an effective strategy or not has major implications for democracy promotion. Randomised designs have been increasingly used to enable some initial conclusions to be drawn.

Wantchekon (2003) investigated the voting behaviour of the citizens of Benin during the presidential elections of 2001. He randomly assigned villages to three groups, the first receiving a clientelistic message, the second "broad, nationally oriented" messages and the third both kinds of message. His experiment showed that clientelism increases the probability of votes going to regional and incumbent candidates, but is less popular among women, who are more attracted to the public-policy type of campaigning.

The examples described here show that RCTs provide new evidence that refutes what was initially hypothesised at the programming stage. In the case of corruption, the implicit assumption was that grassroots monitoring was central in eliminating local corruption. The evidence challenges this hypothesis and encourages further evaluation to see if the findings are similar in different settings.

A broader message conveyed by the examples concerns the feasibility of applying RCTs in evaluating such difficult concepts as corruption and democracy promotion, rather than the success of the intervention itself. These examples have proved that the method is not only theoretically appealing, but can also be managed in the field. As it is also very simple to understand, the RCT method is becoming increasingly popular in social policy. Moehler (2010) has identified 41 RCT studies on democracy and governance that used randomised field experiments. She found that substantial work has been done on elections, community-driven development and improved public service delivery.

However, she noted that that work largely focuses on interventions at village or community level, mainly because RCTs require large samples. This requirement highlights one of the drawbacks of conducting experimental (and, more so, quasi-experimental) techniques. Other important considerations include ethical issues arising from the withholding of treatment from the most needy, the time taken to prepare for evaluation before programme roll-out, the unwillingness of individuals to participate, the complexity of interven-

tion, institutional factors and costs (Burtless 1995; Heckman / Smith 1995). Needless to say, the rule of thumb is to randomise whenever possible, although a number of practical obstacles may prevent randomisation. The next chapter identifies alternative approaches to evaluating governance programmes.

5. Quasi-experimental designs

When random placement of the programme is not feasible, various quasi-experimental approaches can be adopted to construct the counterfactual or “what would have happened without the intervention.” Among the most popular of these approaches are difference-in-differences, instrumental variables, propensity score matching and regression discontinuity designs, which will be described in this chapter.

The suitability of each approach depends on the design of the programme and the nature of the dataset. The use of these methods requires assumptions regarding the characteristics of the data, which are not directly testable. Provided that this is borne in mind, a quasi-experimental approach is nonetheless a powerful tool if used in conjunction with a theory-based approach to contextualise the quantitative findings (White 2006). The next five sections describe these alternative approaches in detail and give examples relating to governance.

5.1. Difference-in-differences (DID) approach

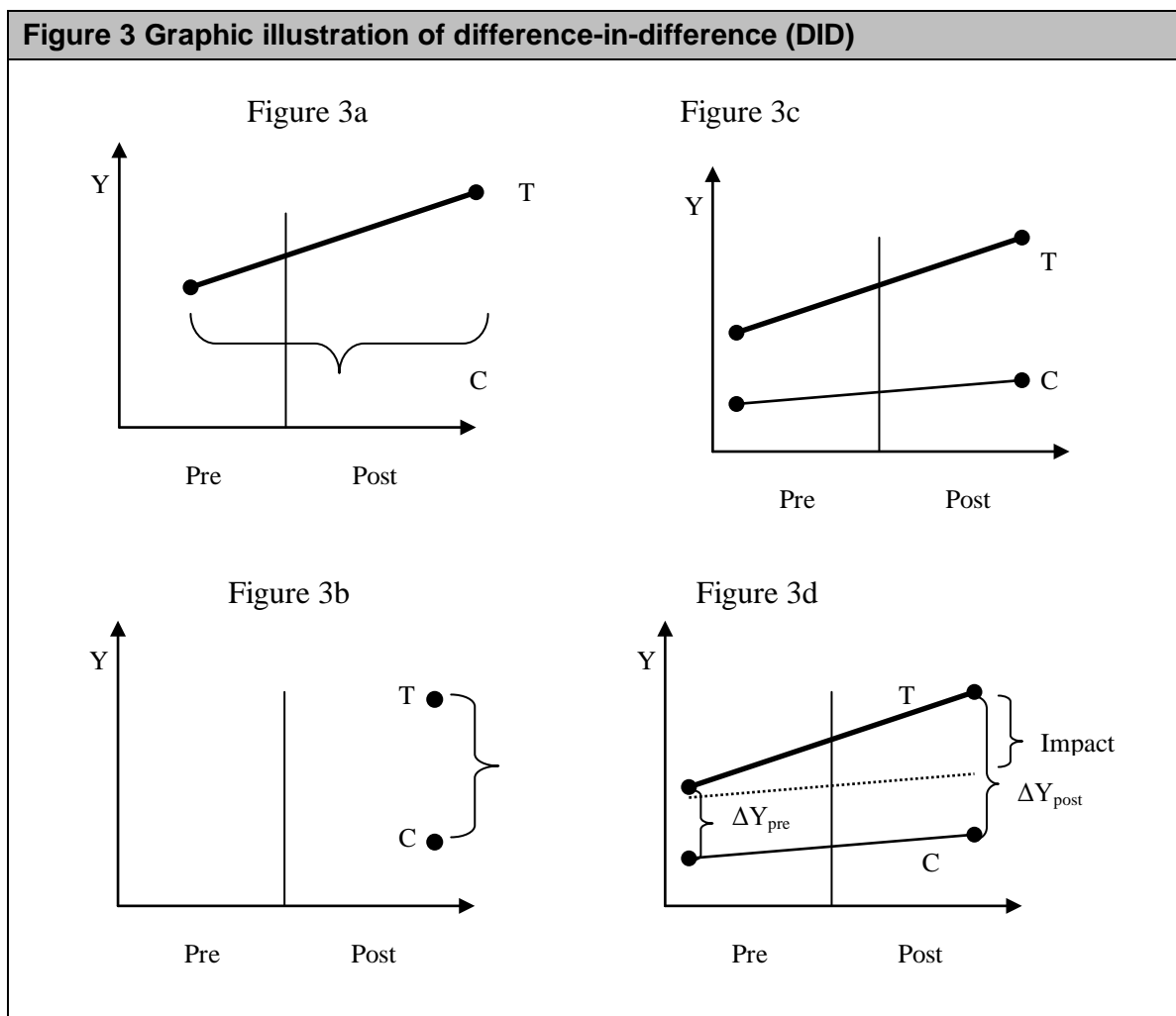
If randomisation is unlikely and the programme has already been rolled out, the difference-in-differences technique can be applied as long as baseline and post-intervention data are available for both the treatment and the control group. The idea behind the difference-in-differences approach (commonly known as DID) is to compare the outcome in the case of one group of individuals who received the programme with the outcome in the case of another group that did not and then to compare their before-and-after levels. Mathematically, the impact is calculated using the difference between pre- and post-intervention mean outcomes for the treatment and control groups and then subtracting the two differences.

$$impact = \delta = (Y_{post}^{treat} - Y_{pre}^{treat}) - (Y_{post}^{control} - Y_{pre}^{control}) \quad (6)$$

DID addresses the issue of non-random treatment assignment by controlling for inherent differences between the two groups. The first difference controls for time-invariant factors while the second difference controls for time-varying factors that are the same in both treatment and control groups. Thus, selection bias is eliminated due to differencing. In other words, this method controls for the unobserved differences between treatment and control group as long as their trends do not change over time. But why do the control and treatment groups require similar outcome trends? The graphs in Figures 3a to 3d illustrate DID and provide the answer.

The vertical axis in these graphs represents the outcome, and the horizontal axis represents time. Figure 3a shows data on individuals in the treatment (T) group before and after intervention. It reveals the ‘naïve’ impact of the intervention on the treatment group, calculated as a single difference before and after intervention ($Y_{post}^{treat} - Y_{pre}^{treat}$). It is a ‘naïve’ impact since considering only the data before and after the intervention means that time-varying factors have not been taken into account. Figure 3b presents the difference

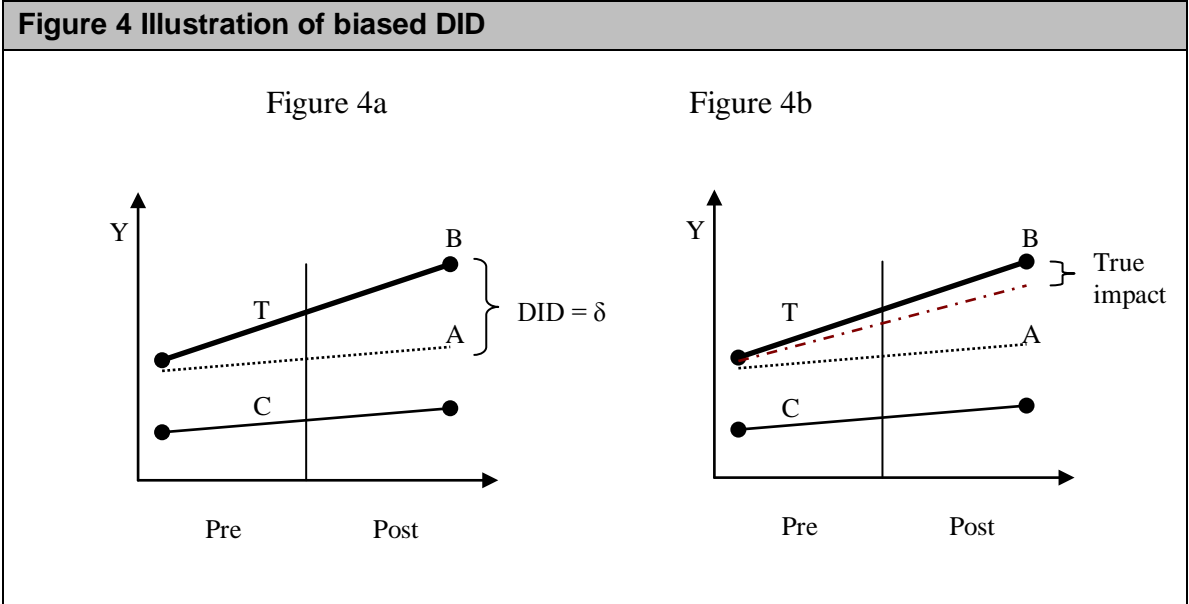
in post-intervention 'naïve' impact on the treatment and the control group ($Y_{\text{post}}^{\text{treat}} - Y_{\text{post}}^{\text{control}}$). A single difference *ex post* implies that pre-existing differences between the two groups have been ignored. Figure 3c shows the combination of the four crucial data points needed for the calculation of DID. The bold line represents the behaviour of the treatment group before and after intervention, while the normal line below depicts the trend followed by the comparison group, again in both the pre- and the post-intervention period. The impact is ΔY_{post} minus ΔY_{pre} . Figure 3d illustrates the manner in which the impact is generated from these points. It is important to note that DID assumes that, in the absence of intervention, the treatment group would follow the behaviour of the control group over time, as depicted by the dotted line. The dotted line represents the true counterfactual outcomes which are never observed. If outcome trends are systematically different, bias impact estimates will result.



Source: Author's representation.

To illustrate this, Figures 3d and 4a show the impact, δ , as the distance between points A and B. It shows the outcome of the control and treatment groups moving in the same direction. If the slopes between the two groups are not the same, then it will be difficult to capture the true average treatment effect. For instance, if the dotted line between A and B in Figure 4b is the true behaviour of the treatment group without intervention, then δ is overestimated.

Thus, the key assumption with DID is that the treatment and control groups should have similar growth rates in outcome. In reality, this assumption between the two groups cannot be directly verified. One way to increase the credibility of the results is to have two base-lines and to check whether trends in both groups prior to the intervention remain stable.



Source: Author's representation.

Difference-in-differences can also be applied using a simple regression framework. Following Wooldridge (2002), the regression equation takes the form:

$$Y = \alpha + \beta T_i + \gamma t_i + \delta(T_i * t_i) + \theta Z_i + \xi_i \tag{7}$$

where $T_i=1$ if the individual belongs to the treatment group, 0 if he belongs to the control group. $t_i = 1$ if the individual is observed in the second time period, 0 if he is observed in the first period. $T_i * t_i$ if the individual is observed in the treatment group and in the second time period. Z_i stands for observable characteristics of individual i . The parameter of interest is δ , the coefficient of the interaction terms T_i and t_i .

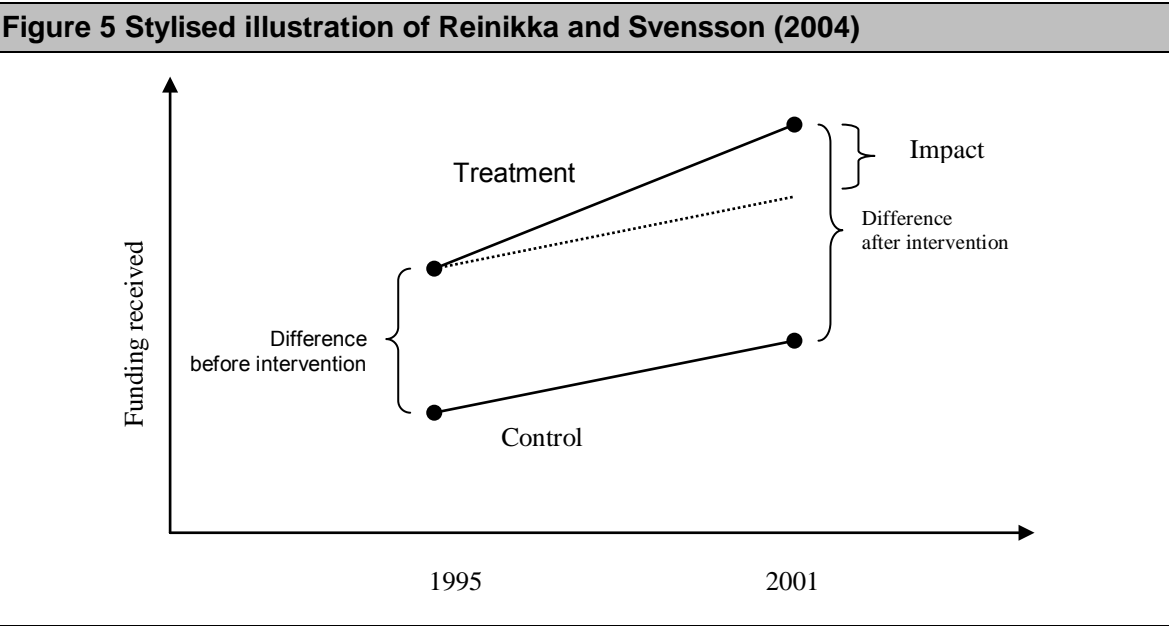
It is easy to assume that DID involves panel data, where the same individuals are tracked for both periods or even longer. According to Wooldridge (2002) and Ravallion (2008), the DID technique may work even if panel data are not available. It can also be applied with pooled cross-section data, as long as the individuals come from the same population and additional regressors are used to take individual-level characteristics into account.

Local capture

Local capture is considered to be one of the main shortcomings of decentralisation. It takes many forms; in some countries it is blatant, in others it is hardly visible and so difficult to pin down. What can be done to minimise it? In Uganda, there was a growing fear that school grants were being captured by local government officials and politicians (at district level) responsible for disbursing the funds to schools. In fact, in 1996 it was found that the some schools received only 20 per cent of the funds allocated to them, while others, tragically, received none.

The central government then launched a newspaper campaign in 1997 in which information was published on monthly transfers of capitation grants to districts and on how local officials had handled the grants programme. The newspaper was circulated to various districts, targeting both schools and parents. The aim was to empower the schools and the constituents to demand what they were entitled to.

To determine the impact, Reinikka and Svensson (2004) applied the DID approach to see whether access to public information would reduce the capture of school funds by district officials. Figure 5 below illustrates the design of their evaluation. The vertical axis represents the funding¹⁰ received by schools and the horizontal axis the time between the beginning and the end of the newspaper campaign. The treatment group consists of schools (represented by head teachers) reporting access to at least one newspaper in 2001, the control group of schools reporting no access to a newspaper in 2001. The four dots represent averages for the treatment group and control group in 1995 (229 schools) and 2001 (217 schools). It should be remembered that a single difference between the treatment and control groups in 2001 is not the impact, since that difference may be caused by selection bias – the inherent differences between the two groups. In addition, a single difference in the case of the treatment group between 1995 and 2001 is not the impact, either, since the change may have been caused by factors that unfold over time. Impact, as shown below, is the widening of the funding gap between the treatment and control groups.



Source: Author’s representation.

The difference-in-differences estimate shows that schools claimed a significantly larger proportion of their grant after the newspaper campaign started. The DID results obtained by Reinikka and Svensson (2004) are presented in Table 5. The “access to newspapers” row represents the treatment group, “no access to newspapers” the control group. The “access-no access difference” is the difference between the two groups. The figures are

¹⁰ Funding received by schools is calculated as the share of money received by the school from the district office and the grants disbursed by the central government for that school.

averages and the numbers in parentheses are standard errors. The last column and last row show the DID estimate. The table reveals that funding received by schools in both the treatment and the control group in 1995 was, on average, very meagre. This is a possible indication of local capture and corruption. When the newspaper campaign was introduced, funding increased for both groups, with the treatment group receiving higher funding. The DID estimate is 13.8 per cent and is significant at 5 per cent. Thus the share of the funding received by schools with access to newspapers rose by an average of 13.8 percentage points compared with those with no access to newspapers.

Table 4 Difference-in-differences estimates of the effects on funding of having access to newspapers

Group	Year		
	1995	2001	2001-1995 difference
<i>Campaign experiment</i> (number of observations: 444)			
Access to newspapers	24.5*** (2.87)	83.7*** (1.94)	59.2*** (3.46)
No access to newspapers	29.6*** (5.40)	75.0*** (3.11)	45.4*** (6.22)
Access-no access difference	-5.12 (6.10)	8.68** (3.66)	13.8** (7.13)

Source: Reinikka and Svensson (2004), Table 4.

The important assumption made in the DID model is that, in the absence of intervention, the behaviour of the treatment group would be the same as that of the control group over time. In their paper, Reinikka and Svensson validated this assumption by conducting robustness tests. Firstly, they used earlier baseline data to determine whether outcome trends change systematically across groups. Secondly, they checked to see if funding shares changed when schools differed in other ways. In other words, they were interested in establishing if the treatment schools had other specific characteristics.

Although Reinikka and Svensson (2004) tested the plausibility of assuming that the outcome trends are similar between control and treatment groups in the DID approach, estimating the impact still causes concern. This stems from the endogeneity problems possibly associated with access to newspapers. In other words, (i) “there may be some unobserved school characteristics correlated with both newspaper access and the efficiency in which the school can articulate its case to the district officials,” (ii) “schools (head teachers) may be informed about the grant (program) even if it does not have a newspaper if parents in the community where the school is located have access to one,” and (iii) “newspaper readership (frequency, time spent, etc.) may vary greatly across schools reporting access to at least one newspaper” (Reinikka / Svensson 2004, 13-14). This issue of endogeneity and how to solve it is explained in the next section.

Oil discovery and corruption

How does an oil discovery affect governance mechanisms in fragile countries? One theory is that countries tend to grow especially slowly if they have a weak institutional framework (Mehlum et al. 2006). But why is this the case? Vicente (2010) explores the effect of an oil discovery announced in Sao Tome and Principe, which led to heightened corruption among public servants. He used the DID approach to identify the difference between Sao Tome and Principe to Cape Verde in corruption perceptions in public services before and after the oil discovery. He carried out retrospective household surveys in both island countries and also asked the respondents questions about their personal histories. His findings showed strong evidence of perceived corruption in education and customs, as well as an increase in vote-buying.

As evidenced from these examples, DID is a highly intuitive and flexible method to use. It accounts for the unobservable characteristics, due to nonrandom assignment of intervention, as long as they are time-invariant. However, DID may not be plausible in practice if baseline data have not been obtained for both treatment and control groups.

5.2. Instrumental variable approach

“Doctors are observed to be frequently in the presence of people with fevers, but doctors do not cause the fevers; it is the third variable (the illness) that causes the two other variables to correlate (people with fevers and the presence of doctors).”

Leeuw and Vaessen (2009)

The instrumental variable approach (or IV) is a popular technique for addressing (i) selection bias – when participation in a programme may have been deliberately targeted at certain groups of individuals – and/or (ii) endogeneity – when the dependent variable affects the independent variables owing to unobserved individual differences between the treatment and control groups. The latter occurs, for example, when one group (treated or control) is more “motivated” than the other. Unlike DID, IV can check for factors that vary over time. Solving these problems requires the identification of an instrument (or a variable) that is correlated with the intervention, but not with the outcome. The use of the instrument eliminates the endogenous assignment of the treatment or participation. The IV approach entails the use of two-stage least squares (2SLS), the first stage involves the regression of endogenous treatment variable T on instrument Z and other exogenous regressors X_i . The second stage involves the regression of outcome Y on predicted treatment \hat{T} and X_i .

$$Y_i = \alpha + \beta T + \theta_i X_i + \varepsilon \quad (3)$$

$$\text{Stage 1: } T = \lambda + \delta Z + \pi_i X_i + \xi \quad (4)$$

$$\text{Stage 2: } Y_i = \alpha + \beta(\hat{T}) + \theta_i X_i + \varepsilon \quad (5)$$

The IV estimator (or parameter β) is a consistent estimate of the average treatment effect on the treated and is expressed as $\beta = Cov(Y, Z) / Cov(T, Z)$. The main problem with the IV approach is the difficulty of identifying the instrument. Its validity hinges on the use of good instruments, instruments that are closely correlated with the treatment and do not

affect the outcome. Weak instruments may exacerbate the bias and lead to incorrect inferences (Glazerman et al. 2003; Ravallion 2008).¹¹

Local capture, continued

Returning to the problems identified by Reinikka and Svensson (2004) and discussed in the previous section, the authors adopted an IV approach to determine how the impact would evolve if access to newspapers was considered endogenous. In doing so, they needed to identify an instrument that affected exposure to the newspaper campaign, but should not be directly correlated with the 'ability to claim funds from the district.' They identified the third variable (Z) as the distance between the nearest newspaper outlet and the schools. This variable was chosen because distance captured the 'cost and ease of accessing a newspaper.'

In the first stage, they regressed the head teachers' knowledge of the distance from the nearest newspaper outlet (the instrument). The teachers' knowledge was proxied by the aggregate score achieved by the teachers in questions put to them. The test consisted of questions on the formula used to arrive at the capitation grant and the timing of disbursements. They found that teachers in schools closer to newspaper outlets knew more about the grant programme and the timing of disbursements. In the second stage, they regressed the grants received by the schools on the predicted teachers' knowledge, the post-intervention year, the interaction of these two variables and income.¹² The result shows that schools exposed to newspapers, or 'more informed schools,' increased their funding by 44.2 percentage points between 1995 and 2001. The result lends credibility to the previous finding generated by DID.

But how do we know if the distance from the newspaper outlet is a legitimate instrument? It should be noted that this instrument must affect the head teacher's knowledge of the grant as a result of increased access to newspapers, but not the funds he claims. The authors conducted several tests:

- To check that distance correlated with access to newspapers, they undertook a regression to show that the relationship between distance and access to newspapers was negative and significant, meaning that the shorter the distance from the outlet, the greater the access to newspapers. This suggests that proximity to newspapers increased head teachers' knowledge of the grant programme.
- To verify that the head teachers in the treatment group were not more knowledgeable than those in the control group (and so better able to claim funds), they were subjected to a written test to compare their general knowledge of politics and public affairs. The results show that distance from the nearest newspaper does not correlate with the head teachers' test scores in terms of general knowledge.

Overall, these tests show that the instrument correlates closely with the endogenous variable (teacher's knowledge of grants), but not with the outcome (funding).

¹¹ Weak instruments can be verified using the test of overidentifying restrictions. See Wooldridge (2009) for detailed explanation.

¹² The structural regression is: $schoolshare = f(\hat{teacher}, 2001, \hat{teacher} * 2001, income)$

Voting behaviour

Why do citizens need to vote? The simple answer is that government decisions have a direct impact on each individual and the community as a whole. Voting gives individuals the power to choose and to express their opinions. It is an essential process in the promotion of democracy. For a variety of reasons, voter turnout and political participation have continued to decline in established democracies in recent decades (Niemi and Weisberg 2001). What could be done to rekindle the enthusiasm of voters? Can voter mobilisation campaigns do the job?

Arceneaux et al. (2006) studied the impact of a “large-scale voter mobilisation experiment” in Iowa and Michigan before the 2002 mid-term elections. The treatment consisted of “get-out-the-vote” phone calls encouraging citizens to vote and reminding them of the election date. Voters were randomly assigned to treatment and control groups.¹³ Those in the control group did not receive any calls. Since some individuals in the treatment group refused to listen to the message or did not answer the calls, only 41.8 per cent received treatment. The failure to administer the intervention to a fraction of individuals assigned to the treatment group created a selection problem.

As not everyone assigned to the treatment group received the message, an instrumental variable regression (2SLS) was employed. The authors exploited the characteristics of their data, distinguishing cases where individuals were actually treated from those assigned to the treatment group. Thus the endogenous treatment variable, T , is a dummy variable equal to 1 if the individual received the treatment, while the instrument is a dummy variable, Z , equal to 1 if the individual is assigned to the treatment group.

After the election, the voting data on each individual were collected to see whether the treatment had been effective. The first stage of regression involves regressing treatment assignment T with instrument Z . The second stage involves regressing the outcome variable (dummy equals one if the individual voted in 2002) with the predicted T and covariates. Interestingly, the authors found that get-out-the-vote phone calls did not increase voter turnout.

Crime

In the third example the question is how best to reduce crime. Donors often make substantial investments in police training in developing countries. But whether the police can effectively reduce crime is still an open question, since studies on its effect show conflicting conclusions.

Studying the effect of police on crime is plagued with endogeneity problem, since increases in crime are likely to induce the government to recruit more police officers. Hence it is not clear which one affects the other. Most cross-sectional studies found that the police had no impact on crime (Cameron 1988), while studies addressing the specification issue more carefully (that there is reverse causality between policemen and crime) identified a substantial, negative impact (Corman / Mocan 2000; Levitt 1997; Marvell / Moody 1996). In his contribution to this debate, Levitt (2002) identifies the impact of the police on crime using annual, city-level panels covering the period 1975-1995 in 59 large US cities. He proposes the number of firefighters per capita as the instrument, since firefighters and

¹³ In detail, the two states are composed of districts which were divided into two groups: competitive and uncompetitive. Within each group, households with two or more registered voters were randomly assigned to treatment and control groups. Only one person per household was selected.

police officers are highly correlated over time and there is little reason to believe that fire-fighters have a direct impact on crime. He acknowledges, however, that the latter assumption may be questioned. He found that the police reduce violent and property crimes.

Overall, these examples illustrate that the IV approach is useful if programme placement is non-random and the data exhibit endogeneity. IV can check for both observable and unobservable characteristics as long as the available instrument is found. And unlike DID, IV can control for time-varying selection bias.

5.3. Propensity score matching (PSM)

If a given governance intervention starts without the treatment being allocated randomly, individuals who receive treatment can be matched with one or more individuals who have not received treatment on the basis of some observable characteristics. This is the essence of PSM. Reverting to the police and corruption story in the previous section as an example, the TEA has selected several policemen to join the quota programme (the treatment group). It should be noted that, in this case, the TEA has conducted the intervention without randomly assigning the treatment. It is possible that the TEA chooses only those policemen who volunteer or who happen to be deliberately assigned. And as the president suggested, it has also included policemen who did not participate in the programme as a control group. The problem here is that the TEA cannot just select *any* policemen who did not participate in the programme, since those who participated may be inherently different from those who did not. Thus, to ensure that the two groups do not differ systematically across various observable characteristics (gender, income, family size, etc.), matching is required. Matching one characteristic, such as gender, of all policemen in the treatment group with that of one or more policemen in the control group is relatively easy, but matching ten characteristics of one policeman in the treatment group with those of one or a hundred policemen in the control group is a time-consuming exercise. Furthermore, an exact match of a huge number of characteristics of individuals is virtually impossible. Rosenbaum and Rubin (1983) therefore recommended matching based on a propensity score. This means that matching is based on the probability that the individual will participate in the programme.

$$\Pr(\text{propensity}(X_i)) = \Pr(\text{Treatment} | X)$$

Propensity score matching eliminates selection bias by pooling individuals from the control group who have similar characteristics to those of the treatment group. The objective is to increase the similarity of participants and non-participants in the programme. The counterfactuals are the non-participants with characteristics similar to those of the participants.

Some caveats are necessary: first, implementing PSM requires a rich set of control variables and comparable surveys of treated and control groups. Second, matching is based on observable characteristics, not on unobservables – behaviour that cannot be observed and/or measured, such as motivation and enthusiasm. The key assumption in PSM is therefore that there is no selection bias due to unobservable characteristics or that participation is independent of outcomes.¹⁴

¹⁴ The Rosenbaum and MH bounds are tests that can provide some indication of the validity of this assumption, although it cannot be directly tested. See Becker and Caliendo (2007).

As the propensity score is a continuous variable, the probability of obtaining two similar scores from individuals in the treatment and control groups is infinitely small. Consequently, various algorithms have been invented in which propensity scores of the treatment and control observations are selected and matched on the basis of some tolerance level, weights, strata or neighbourhood. There is no superior algorithm. The selection of the algorithm is a trade off between bias and efficiency. Considering more than one algorithm together can provide a robustness check on the estimates. PSM proceeds in the following stages:¹⁵

1. Collect comparable surveys of treated and control groups.
2. Pool the samples and estimate the probit of programme participation.
3. Restrict samples to ensure area of common support. Common support ensures that the characteristics in the treatment groups are similarly observed as in the control group (see Figure 6)
4. Choose matching algorithms.
5. For each treated individual find controls with similar propensity scores.
6. Calculate the increase/decrease in outcome for that observation.
7. Calculate the average of the individual results to obtain impact.
8. Carry out a sensitivity analysis using Rosenbaum or MH bounds.

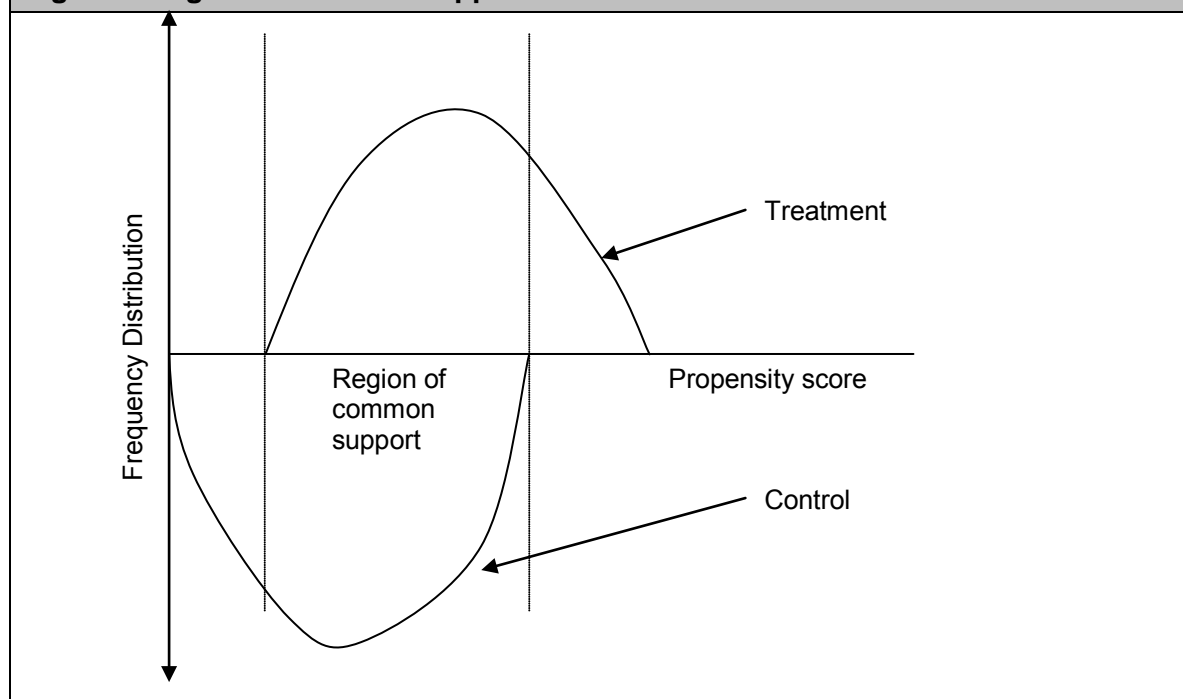
Transparency

An example of its application has been done to determine whether transparency of government performance motivates people to participate or engage in civic activities. In many countries, donors actively participated in developing performance benchmarks for local government at the height of decentralisation. One of the most popular is the Citizens Report Card in India, Bangladesh and Mexico. In the Philippines, the Report Card Survey is just one of 30 different performance assessment systems funded by donor agencies. Most of these systems are aimed at improving public service delivery and welfare. However, public service delivery can be improved only if people become more active members of the community. Anecdotal evidence has provided some clues to the effectiveness of these measures. But a causal link between information on local government performance and civic participation is yet to be established.

Capuno and Garcia (2010) investigate the effect of information concerning the performance of local government officials on people's participation during decentralisation in the Philippines. A local governance performance index was introduced in 12 cities and municipalities. At the eight treatment sites, the index scores were announced by means of public presentations, posters, magazines and stickers, while the index scores were not announced at the four control sites. Since not all individuals in the treatment municipalities receive treatment, there is reason to suspect selection bias. Propensity score matching is therefore used to match individuals who actually received treatment (in the treatment municipalities) with those in the control municipalities.

¹⁵ See Caliendo and Kopeinig (2008) and Becker and Ichino (2002) for more discussion on matching algorithms and practical guidance.

Figure 6 Region of common support



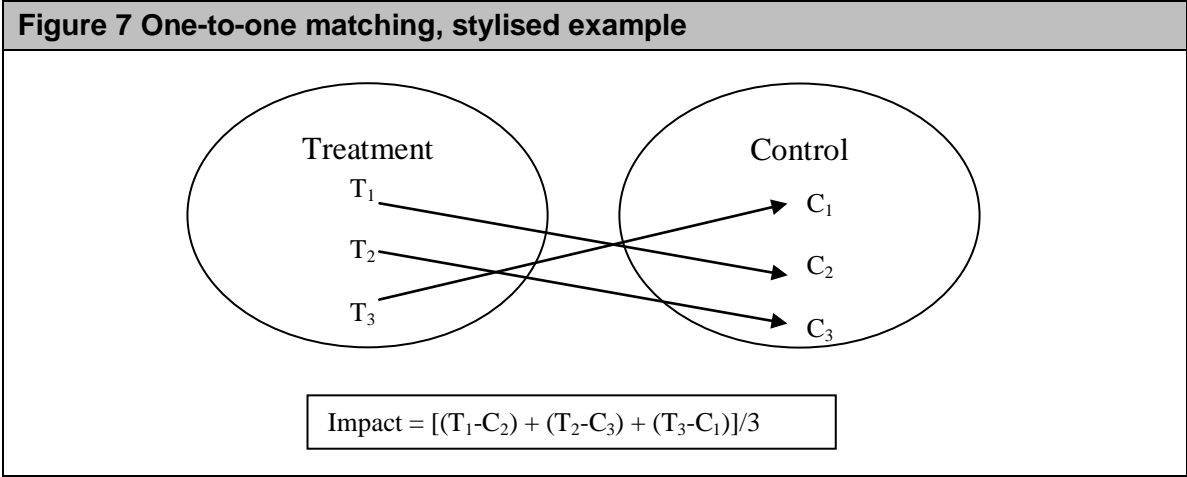
Source: Author's representation.

The data were derived from three rounds of random household surveys conducted at all the pilot sites. At the eight treatment sites, 178 individuals reported exposure to the treatment – 95 in the 2002 mid-pilot survey and 83 in the 2003 post-pilot survey. The treated individuals were those who had read the magazines, seen the posters or attended the public presentation. There were three comparison groups: (i) 400 individuals from the four control municipalities in 2002, (ii) 400 individuals from the four control municipalities in 2003 and (iii) 800 individuals in the two years combined. A number of individual and village-level characteristics, including gender, age, schooling, income, family size, employment and civil status, were used to determine participation. For each comparison group, the result shows that information on local government performance leads to increased civic participation. The calculation is as follows:

- The 95 individuals in the treatment group were matched with the 400 individuals in the control group in 2002 using various matching algorithms.¹⁶ Similarly, the 83 treated individuals in 2003 were compared with the 400 individuals in the control group for the same year. On the assumption that there were no time-varying factors that might affect the outcome, the combined number of treated individuals in 2002 and 2003 (178) were matched with 800 individuals in the control group in the combined years. Some individuals who had very few characteristics in common with the others were dropped, thus restricting the sample to individuals in both the treatment and the control group who shared an overlap. This is also known as the region of common support. It ensures that the characteristics observed in the treatment groups are the same as those in the control group (see Figure 6).

¹⁶ Matching algorithms used are nearest 1-to-1, nearest neighbour, kernel, radius and stratification. See Becker and Ichino (2002) for a description.

After the sample had been restricted, one or more control individuals were matched with each treated individual. Assuming one-to-one matching as illustrated in Figure 7, this implies that one individual in the treatment group will be matched with one in the control group who has the closest propensity score. The difference between the two individuals' outcomes was calculated. After this procedure had been repeated for all individuals in the restricted sample until all treated individuals were matched, the impact was calculated from the average difference in the outcomes of the treated and control units.



Source: Author's representation.

Deliberation

Deliberation is a venue for individuals to express a wide range of views and is considered a crucial component of the democratic process. It builds knowledge based on the arguments presented by participants so that a sound judgement or decision may be made. But whether deliberation can in fact increase learning is an empirical question.

Barabas (2004) employed propensity score matching to assess the effect of people's deliberations on social security reform in the United States. A large 'deliberative forum' on possible social security reforms was carried out to see whether participants' learning would increase and whether their opinions would change. Participants were provided with information materials before the forum. During the forum, the organisers reiterated the importance of having an open mind and of abandoning strongly held views. Four hundred and eight citizens from a random sample of registered voters in the county of Maricopa participated. Invitations to participate were also publicised locally in the media and on the internet. Participants in the treatment group were those who attended the forum (or forum group), and the control groups were (i) those who were invited but did not attend and (ii) a random sample of people in the county. Because assignment to treatment and control groups was non-random, it is difficult to tell why some citizens participated and others did not. To overcome this problem, the author employed PSM to match the forum and control groups. To determine whether deliberation had changed the knowledge and opinions of the participants, Barabas calculated the net effect as the 'difference between the pre- and post-forum survey measures for the group that attended the deliberative forum subtracted from the difference in the pre- and post-forum scores of the matched comparison group members.' He found that deliberation increases knowledge and can change opinions on condition that the information provided is of good quality and participants have diverse views and an open mind.

Unlike DID, PSM does not require baseline data. The impact is calculated based on a single difference of outcome after the intervention. The main drawback of this approach is the difficulty in justifying that selection bias on unobservables is small enough to have any effect on the impact estimate. On a more practical note, PSM can also be computationally challenging.

5.4. Regression discontinuity design (RDD)

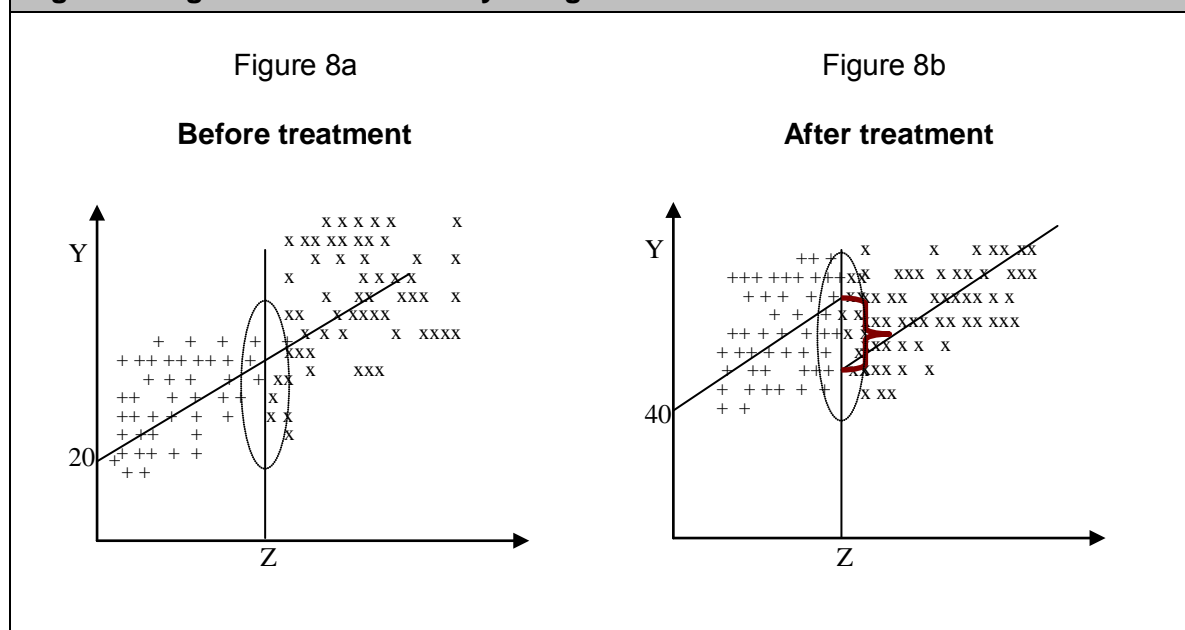
Some donor programmes target groups of individuals on the basis of certain eligibility criteria (such as test scores, age or poverty index). Individuals below a certain cut-off receive the treatment and those above it do not. This division automatically generates the treatment and control groups. The characteristics of individuals in the neighbourhood of the eligibility cut-off should be more or less the same except for, say, income (in the case of the poverty index). Since individuals are considered homogeneous in this subset, RDD checks for both observed and unobserved individual characteristics. Thus, to some extent, the closest approximation to randomisation in producing unbiased estimates is achieved with RDD (Rossi and Freeman 1993). Finally, the impact is calculated through a comparison of the shift between pre-treatment and post-treatment outcomes at the cut-off.¹⁷

If we consider an election education campaign that seeks to increase the probability of registration and assume that this intervention targets participants on the basis of a poverty indicator, those below the cut-off level Z will receive the treatment and those above that level will not (see Figure 8a). Along the eligibility cut-off, there is no reason to believe that individuals are different except for their incomes. Figure 8a illustrates the outcome (Y) for individuals below and above the poverty cut-off (Z). Those who are below the cut-off are participants in the programme and those above it are not (the control group). Figure 8b shows the discontinuity in the regression line of the targeted group after the information campaign.

Along the boundary Z , impact, or the local average treatment effect is derived. This is calculated by taking the outcome of individuals who are at the boundaries – i.e., who are marginally below and above the cut-off point. The estimate is assumed to be free of selection bias, both observable and unobservable, since individuals would tend to exhibit similar characteristics along the boundaries of this cut-off.

¹⁷ The RDD approach requires that the eligibility criterion is continuous and adhered to consistently, i.e. participants should not be able to manipulate their assignment to either treatment or control group.

Figure 8 Regression discontinuity design: an illustration



Source: Author's representation.

Local spending

Do local governments spend more under representative democracy than direct democracy? Are there policy consequences from opting one than the other? The debate on the virtues of direct democracy as providing citizens with great amount of participation in the legislative processes to that of representative democracy, where rights are transferred to group of people, persists as strong as ever.

To contribute to this debate, Pettersson-Lidbom and Tyrefors (2007) analysed the effects of direct and representative democracy on per capita spending of local government in Sweden using pooled cross-section data from 1930 to 1950. Whether localities adopt direct or representative democracy is highly dependent on their population size. The cut-off population size is mandated by law. From 1919 to 1938, municipalities with a population of 1,501 or more were required to have representative democracy. In 1938, the population cut-off was reduced to 701. Municipalities falling below the cut-off could choose either direct or representative democracy. Because the nature of the data allows those who are below the cut-off to choose, eligibility rules are not strictly followed. To circumvent this problem, a variation of RDD design called “fuzzy discontinuity” has been applied. With fuzzy discontinuities, the average local treatment effect is the ratio of the difference in mean outcomes above and below the cut-off to the difference between the mean treatment status of eligible individuals and ineligible individuals. In this example, the authors used the ratio of the difference in spending to the difference in the probability of treatment. According to their preliminary RDD results, local government spends less under direct democracy than representative democracy.

Social assistance

RDD has been widely applied in labour market studies, but not many studies can be found in the governance realm. To illustrate RDD further, one of the many labour market studies

is taken as an example. The study links the working behaviour of citizens to the social assistance provided by government.

Governments are always keen to strike a balance between social benefits and disincentives to work. This is evidenced by a series of labour market reforms in the USA, Canada and Europe. Analysing the impact of reforms is challenging owing to macroeconomic changes that alter the demand for labour, as well as isolating the effects of other policies (Blank 2002). Lemieux and Milligan (2008) analysed the impact of a policy in the province of Quebec, which provides less social assistance for childless individuals under age 30 than recipients over 30. The policy break at age 30 provided a natural opportunity to analyse the impact of welfare payments using RDD. By using RDD, the authors were able to circumvent the problems mentioned above. First, by exploiting the discontinuity in a static policy, they were able to ensure that no other reforms contaminated the evaluation of the low-benefit policy. Second, no assumptions were needed regarding the comparability of the treatment and control groups, even when the economic environment changed, since, the two groups should be more or less identical, except for their age, where the discontinuity occurs. The average local treatment effect is simply the difference between outcomes below and above the cut-off within a certain range. Using census data, the authors found strong evidence of generous social benefits having a negative impact on employment.

These examples show that, as average treatment effects are calculated at the boundary, some concerns need to be addressed when RDD is applied:

- First, the result may not always be true for the entire population.
- Second, it is questionable whether the observations at the cut-off would be of interest to policymakers.
- Third, the units of observation at the boundary may be too few in number, which could make it difficult to arrive at a robust inference.
- Lastly, but more generally, strict adherence to the cut-off should be ensured, meaning that individuals should not be able to influence their assignment above or below the cut-off whenever they choose.

5.5. Combining techniques

Depending on the nature of the data, two impact evaluation techniques are usually combined to increase the accuracy of the impact estimate. This may minimise an identified cause of bias that one approach cannot altogether eliminate. In the previous example of local capture (see sections 5.1 and 5.2), Reinikka and Svensson (2005) combined the DID and IV approaches in studying the effect of a newspaper campaign on corruption in Uganda. As previously shown, the authors used DID in calculating the impact of the programme. But because they suspected endogeneity in programme participation, IV was used to solve the problem and confirmed the positive results.

Apart from using RDD in analysing the effects of social insurance on employment, Lemieux and Milligan (2008) also employed DID to check the results obtained with RDD. They have utilized data after the cancellation of the low benefit policy. The cancellation has produced a pre and post comparison for both treatment and control groups. They found that DID does not perform well if the control groups used do not come from the

same area. This implies that, in this context, the DID results could not take broader economic changes into account.

More examples of evaluation not related to governance, but routinely carried out in the various sectors, include the combination of PSM/DID and of DID/RDD.¹⁸ Specifically, Ravallion and Chen (2005) assessed the impact of the Southwest China Poverty Reduction Project on household saving and consumption behaviour in rural villages. They used PSM to match villages, so that initial heterogeneity in the treatment and control villages could be removed. They then applied DID to the matched villages. This technique is used to increase the precision of impact estimates, especially in cases where changes over time are dependent on initial conditions. Further programme placement is also based on these baseline characteristics and is typically not randomly assigned (Jalan / Ravallion 1998; Ravallion 2008).

Another technique is the combination of DID and RDD undertaken by Jacob and Lefgren (2004) in education. They evaluated whether the school remedial measures, such as summer schools have an impact on a student's future achievement. They used the test scores as eligibility criteria and a DID estimator to determine the gain above and below the cut-off score as well as the gain before and after.

5.6. Other evaluation approaches

Other methods can also be applied to evaluate programme impact. Some of these techniques include quantile regressions and general equilibrium analysis. Quantile regression involves analyzing the effect of the program across distribution of outcomes. Even if the average impact of the programme is zero, policymakers are often interested in how the outcomes vary for specific groups of targeted individuals or households. General equilibrium analysis is useful when it comes to assessing the effect of a certain policy change. It models the effects of macroeconomic policies and even economic shocks on households or firms. Such modelling is often very complex and requires many identifying assumptions. In both techniques, how the programme was implemented remains an important factor in the analysis since selection bias must always be taken into account. This is beyond the scope of this paper, but reference can be made to Khandker et al. (2010) for a detailed explanation.

5.7. Which technique to use?

Now that the various evaluation designs have been presented, the question of when one technique is more appropriate than another remains. There is no quick and easy answer to this question. And typically, there is no definitive approach to choosing the best evaluation technique. Each method has its own particular requirements, advantages and disadvantages.

To a certain extent evaluators must rely on their own judgment to strike the right balance. However, going back to programme theory helps a great deal to identify potential methods, since a policy intervention is almost always case-specific and, in some respect, unique. Its design varies with the beneficiaries of the programme and the socio-economic

¹⁸ Due to the limited impact evaluation studies in the governance area, examples from other sectors are also presented to illustrate how the combination of techniques can be performed.

characteristics of the population. Such peculiarities must therefore be addressed in the research design.

Nonetheless, a rough classification of the applicability of the methods discussed in this paper is possible: the toolbox below (see Table 6) is a simple guide to determine when each quantitative method can be used, where it has been applied and what are the key requirements, advantages and disadvantages.

Table 5 Summary toolbox

Methods	Usage	Key requirements/assumptions	Advantages	Disadvantages
RCT	<p>When randomisation of treatment assignment is possible</p> <p>Ex: Olken (2007) on corruption, Wantchekon (2003) on voting behaviour</p>	<ul style="list-style-type: none"> • Full compliance of individuals who are offered treatment 	<ul style="list-style-type: none"> • Simple and powerful • Gold standard or most preferred choice • Baseline survey typically not needed • Can easily trace the type of individuals in the group where the programme is most effective 	<ul style="list-style-type: none"> • Ethical issues • Not always feasible • Individuals may not always be willing to participate • Potential contamination issues • External validity
DID	<p>Pre- and post-intervention data are available for both treatment and control group</p> <p>Ex: Reinikka and Svensson (2004) on local capture</p>	<ul style="list-style-type: none"> • The growth trends in outcome should be the same for treatment and control group <p>or</p> <ul style="list-style-type: none"> • Unobserved characteristics are time-invariant between the two groups 	<ul style="list-style-type: none"> • Addresses non-random assignment of intervention • Simple and intuitive • Takes unobservable characteristics into account as long as they are time-invariant 	<ul style="list-style-type: none"> • Biased if trends change between the two groups • Assumption that selection bias is time-invariant may be implausible in practice • Participants must already be identified at baseline
PSM	<p>Good-quality data for matching No baseline data collected</p> <p>Ex: Capuno and Garcia (2010) on transparency</p>	<ul style="list-style-type: none"> • No unobservable characteristics in either treatment or control group 	<ul style="list-style-type: none"> • Does not require randomisation or baseline data 	<ul style="list-style-type: none"> • Biased if unobservable characteristics exist • Must have enough cases in the common support region
IV	<p>When data exhibit endogeneity If available instrument is found</p> <p>Ex: Reinikka and Svensson</p>	<ul style="list-style-type: none"> • The instrument is a variable correlated with participation but not with outcome 	<ul style="list-style-type: none"> • Allows for time-varying selection bias • Controls for both observable and unobservable characteristics 	<ul style="list-style-type: none"> • Difficult or sometimes impossible to find good instruments

Methods	Usage	Key requirements/assumptions	Advantages	Disadvantages
	(2004) on local capture, Levitt (2002) on crime		tics	
RDD	<p>If programme placement is based on a certain cut-off. Cut-off must be continuous</p> <p>Example: Lemieux and Milligan (2008) on social assistance Pettersson-Lidbom and Tyrefors (2007) on local government spending</p>	<ul style="list-style-type: none"> Eligibility criteria must be defined. Individuals should not be able to influence assignment above or below the cut-off 	<ul style="list-style-type: none"> Allows for observed and unobserved characteristics Mimics randomisation to a certain extent 	<ul style="list-style-type: none"> Provides treatment effects on a highly selected subsample Sufficiently large sample around the threshold

6. Tackling the difficulties

Impact evaluation in governance shares many of the challenges and data constraints of other development programmes. The previous chapter considered the ideal situation where sample sizes are large, no spillovers occur, the unit of analysis in most cases is the individual, the interventions are clear-cut and can be manipulated, outcome measure is straightforward, and there is room to collect control groups. What happens if an intervention does not correspond to this ideal situation? This problem is not new in the evaluation literature. Considerable experience has been accumulated in such heavily evaluated sectors as labour, education and health, showing how to tackle some of the challenges that arise when not all requirements are met or not all assumptions prove correct.

The aim of this chapter is to compile a list of some of the difficult points and issues arising in the evaluation of governance programmes. In most cases, the solutions are variations on the standard techniques discussed earlier. *The possible solutions are discussed in detail and draw on cases in sectors other than governance*, since examples of impact evaluation of governance are relatively few in number.

6.1. Measuring outcome

An important obstacle encountered in the evaluation of governance interventions is quantifying outcome that is largely characterised by human behaviour and difficult to pin down. Unlike some development programmes, there is no easy or straightforward measure of governance outcomes. For impact studies in the traditional sectors, the measure of outcomes can be defined and is not as controversial. For instance, the effects of education on labour are captured by wages and employment; nutritional improvements are measured by such anthropometric indicators as weight, height and body mass index; and the effectiveness of fertilisers is measured by farm yield. But what is the best way to capture a decline in corruption? Or an increase in civic participation? Or an improvement in law enforcement?

It is important to acknowledge that governance outcomes cannot be measured with a single, all-encompassing indicator. The practical approach would be to unpack the intervention and assess which of its dimensions can be rigorously analysed. Although measuring outcome is a challenging task, some authors have been very creative in solving this problem. A very good example is the randomised experiment conducted by Olken (2007), which was discussed in an earlier chapter of this paper. Instead of using a perception-based corruption measure as the outcome variable, he devised a direct measure of corruption: the difference between reported expenditure and an independent engineer's estimate. Other studies also use creative outcome measures. Table 4 presents the outcome measures used in the selected governance interventions discussed in this paper.

Table 6 Selected governance interventions, outcomes, measures and methods

Intervention	Outcome	Outcome measure	Authors
(i) Increasing government audits and (ii) grassroots participation	Decrease in corruption in the context of a road project	Difference between reported expenditure and the independent engineers' estimate	Olken (2007)
Conditional cash transfer	Corruption	Survey question whether respondents have been asked for bribes in the public services they have utilised	Grimes and Wängnerud (2010)
Message on clientelism	Voting preference between clientelism and public policy	Dummy variable from post-election survey depicting the voters' choice	Wantchekon (2003)
Direct versus representative democracy	Spending behaviour of local governments	Expenditure per capita	Pettersson-Lidbom and Tyrefors (2007)
Get-out-the-vote phone calls	Voting behaviour	Voting attendance	Arceneaux et al. (2006)
Newspaper campaign	Reduction of local corruption in the context of school grant programme	Ratio of grants received by school to funds disbursed to local government	Reinkka and Svensson (2004)
Deliberative forum	Change in learning and opinion	Scores from various knowledge questions	Barabas (2004)
Performance assessment system	Increase in citizen participation and membership in local groups	Survey questions on (i) individual's participation in planning, implementation, monitoring and evaluation of local government activities and (ii) membership in local groups	Capuno and Garcia (2010)
Women's leadership (mandated representation of women)	Direction of policy decision	Investments in public goods	Chattopadhyay and Duflo (2004)
Cash bonus and driving lessons	Corruption	Driving test (pass/fail)	Bertrand et al. (2007)
Conditional cash transfer	Empowerment of women and spending behaviour	Food, clothing and schooling expenditures	Handa et al. (2009)

6.2. Small sample size

A common problem encountered in the evaluation of governance interventions (especially at higher levels of government) is that the units of observation are often too few for the assumptions associated with parametric statistics to hold. In other words, the accuracy of the inferences based on the methods described hinges on statistical “laws” that require a large sample size (central limit theorem). The methods presented in the last chapter typically have large sample sizes. That is, sufficient cases or observations are collected from the treatment and control groups. Deviating from the large sample size requirement is often inevitable in some governance programmes. There are some alternative approaches to dealing with sample sizes.

First, in cases where the *sample has moderate size* (say 20 observations), the importance of obtaining baseline data must be emphasized in this context. Remember that when randomizing, as sample size increases, the different characteristics between the treatment and control groups would cancel out and the remaining difference would only be the intervention. If sample size is of moderate size, this means that the comparison group may not be very similar to the treatment. Thus, the more pre- and post treatment data available, the more credible the impact estimate could be. Fewer observations also imply that collecting more than just one pre-treatment set of data is logistically easier than where samples are larger.

Second, *if randomisation is possible*, a technique known as ‘repeated measures’ can sometimes be used. ‘Repeated measures’ means that baseline data and post-treatment outcomes will be collected repeatedly for each unit over a period of time. Imagine a country with ten provinces in which a programme is to be implemented. After the baseline data have been obtained, the programme can be gradually “phased in” in at least one of the provinces for a certain period (say six months), with the others used as controls (NRC 2008). Repeating this process until only one control observation is left will produce a total of 10 pre- and at least 10 post-treatment controls. Considering the uniformity of the treatment effects across provinces already enables the reliability of the impact estimate to be gauged. Although the estimated effect may still be biased because some confounding factor has not been taken into account, it may nevertheless allow some inference as to the effectiveness of the programme.

Third, in the case of *very small sample size*, the evaluation needs to conform to a more qualitative approach. In case of non-random assignment of treatment, instead of using propensity score in matching of the control or treatment groups, one can match qualitatively as accurately as possible the key factors that are important in affecting outcome. Again, collection of baseline data is highly advisable.

In the case of $N=1$, i.e. when there is only one treated unit and no controls, pre- and post-treatment information must again be underscored. The frequency with which the data are collected before and after the intervention is crucial for the calculation of the impact estimate. If the intervention can be implemented at the time when external factors could be expected to be least important, then confounding factors may be minimised. However, it is difficult to predict an event or a sudden policy change that may affect the outcome.

Greater efforts should therefore be made to collect qualitative data and conduct interviews/stakeholder analyses.

6.3. No baseline data

Collecting baseline data or pre-intervention information is a challenging task for development practitioners, since (i) interventions may tend to change over time; (ii) all potential outcomes – expected and unexpected – must be taken into account at the outset; (iii) collecting baseline data requires additional financial resources; and (iv) the implementation of the project could be delayed. Those who are to collect baseline data must have an overall idea about how the project will be evaluated later on. Otherwise, the baseline data collected may be insufficient or no longer relevant at the time of evaluation.

If baseline data are not available, using a triple-difference (DDD) method is an option under certain conditions. Ravallion et al. (2005) adopted this approach for the Trabajar programme in Argentina, which provides work for the unemployed poor for six months. They examined the difference in the incomes of participants who had left and participants who had remained in the programme. A simple DD between “stayers” and “leavers” would lead to biased results, since work opportunities are not the same for both groups. Ravallion et al. therefore formed an entirely separate control group who had never participated in the programme. DDD is the DD of stayers with matched non-participants minus the DD of leavers with matched non-participants. The result is the income gained by stayers for participating in the program. DDD technique has allowed changes in the economy or labour market to be controlled for.

Another possible approach if no baseline data are available is to reconstruct a baseline in retrospect, a practice common to many clinical studies. However, the accuracy of recall, i.e. collecting data from individuals' recollections of what happened a year or more earlier, is a serious problem in this type of study.

As discussed above, some quasi-experimental techniques, namely single difference (control versus treatment group) approaches such as PSM or instrumental variable techniques, can be applied even in the absence of baseline data.

6.4. Spillovers and contamination

Spillovers and contamination occur when treating an individual or group of individuals affects the behaviour of individuals who are not in the programme (that serve as controls).

This is one of the most difficult issues in governance evaluation. For instance, the central government has randomly conducted a performance assessment to municipalities. In preparation for the next wave of assessment, municipalities in the control groups started improving their public service delivery than they would normally do. It is then possible that an intervention implemented in one municipality may have unintended impacts on others. If that is the case, the true impact may not be captured as control groups are contaminated and pure with-and-without comparisons cannot be performed.

Miguel and Kremer (2004) present an example of how contamination has been controlled for in the design of the evaluation in the context of child schooling. Earlier medical find-

ings show that deworming drugs have an insignificant effect on the health and school attendance of young children (Dickson et al. 2000). However, Miguel and Kremer believe that this result is underestimated. They argue that the effect of the deworming drugs is greater than what the earlier findings claim, since young children belonging to the control group and attend the same school as the children in the treatment group may also benefit from the positive effects of the treatment (low possibility of disease transmission or fewer infections). The contamination happens because the treatment is randomised at the individual level and positive spillovers can easily occur between the treatment and the control group. Thus, in order to circumvent contamination from their experiment, they randomised at school level. Randomising at school level means that all children in the selected schools receive deworming drugs, while children in the control schools do not. This is in contrast to individual-level randomisation, where both treated and untreated children attend the same school. This approach should help to minimise the spillover effects of the deworming drugs. Miguel and Kremer (2004) also identified cross-externalities in neighbouring schools, where control schools benefited from treatment schools. They found that deworming reduced absenteeism in treatment schools by 25 per cent. But it also improved health and school attendance among untreated children in the control and neighbouring schools. It is easy to imagine how spillovers can occur in governance, and this example describes an excellent approach to tackling this problem.

6.5. Evaluating higher levels of government

In clinical trials and experiments with new drugs, randomisation is often conducted at individual level. However, where the governance agenda is concerned, interest focuses on information obtained not only on the individual citizen but also on government units. Randomisation at city or provincial rather than individual level requires clustered randomisation. The clusters might consist of such geographical entities as cities, communities and provinces or states. This kind of randomisation is sometimes easier and may preclude contamination. Clustered randomisation ideally requires more clusters (and sometimes more individuals within each cluster) to take inter-cluster variations into account, so that a given power may be achieved (Duflo et al. 2007). The power of a test is the probability of a programme having a significant effect, provided that there is *truly* an effect. Increasing the sample size increases this power. Thus, in clustered randomisation, increasing the number of clusters increases the validity of the experiment.

If the number of clusters is small, arranging them in matched pairs is a common strategy. This means that clusters from treatment and control groups are matched one-to-one. The matching is based on factors believed to be correlated with outcome. This allows better comparability between treatment and control groups and reduces variation across clusters. Hayes and Bennett (1999) illustrate a simple way of calculating the appropriate number of clusters for cluster randomisation. They present a method of choosing the sample size for both unmatched and matched pair trials.

Chattopadhyay and Duflo (2004) conducted randomisation at *Gram Panchayat* (GP) level in their study on the impact of women's leadership on policy decisions in India. A GP is composed of 5 to 15 villages of about 10,000 inhabitants. Treatment was assigned at GP level, since it was not possible at village level. As the unit of analysis is the village, standard errors in the authors' regression analyses were corrected for clustering at GP level.

They found that mandated representation does have an effect on policymaking and that women leaders (occupying the special seats reserved for them) invest in public goods about which they are concerned.

Glewwe et al. (2009) adopted a randomised evaluation approach in rural Kenyan primary schools. They studied the effect of textbook provision on student test scores and found that textbooks increased the test scores of only the best students, having no effect on average students. The level of randomisation was the school, but the data used in analysing the effectiveness of textbook were collected at the level of the individual student.

6.6. Complex governance interventions

Sometimes policymakers are interested in not one but a range of policy interventions. In conducting RCTs, one or a combination of treatments can be accommodated with factorial design. Factorial design is simply an experimental design that allows each treatment, say A, B and C, to be combined to produce A, B, C, A+B, A+C, B+C, C+A and A+B+C. This means that various treatments can be tested in a single experiment, thus reducing evaluation costs. Factorial design can also investigate the interaction effects of treatment. However, sample size must be adjusted significantly, depending on the number of interventions to be examined. An example of this factorial design is taken from the study of corruption in the issue of driving licences discussed earlier, which was conducted by Bertrand et al. (2007). They recruited and randomly assigned applicants to three groups: (i) a comparison group (ii) applicants offered cash bonuses if they could obtain their licences within 32 days and (iii) a group offered free driving lessons. This allowed the authors to make comparative assessments of the various types of intervention.

Complex interventions also mean that, as a whole, the programme combines several inputs and conditions. It is sometimes difficult to isolate the effect of one input from the effect of another. One example of a complex programme that combines cash transfers, conditionality and incentives (Duflo et al. 2007) is the Mexican educational programme known as PROGRESA (subsequently renamed Oportunidades). This programme involves the offer of cash transfers to mothers on condition that their children attend school and visit health centres. The primary aim of PROGRESA is to prevent the intergenerational transmission of poverty through investment in human capital. How, then, can cash transfers and conditionality be evaluated?

Using data on PROGRESA, Handa et al. (2009) evaluated the impact of conditionality and of cash transfers directly to women on their spending behaviour. In contrast to previous evaluations that attested to the significant and positive impacts of PROGRESA on various educational and health outcomes (Dubois et al. 2002; Schultz 2004), Handa et al. focused their study on the effects of conditionality and gender targeting to determine whether their inclusion in cash transfer programmes was logical.

As the RCT method has been applied to PROGRESA, evaluating conditionality and transfer income is straightforward. No computationally challenging task needs to be performed. To test the effectiveness of conditionality, Handa et al. simply looked into the spending behaviour of households. They expected the transfer income to be spent primarily on schooling or children's clothing, since that was directly associated with the conditionality. They found, however, that the transfer income was used like earned income, indicating that it was not spent on human capital investment any more than earned in-

come. Conditionality was therefore ineffective. As regards the targeting of women for the cash transfer, they found that it did not significantly contribute to overall decision-making by women. These examples show that even complex interventions can be rigorously evaluated as long as key evaluation requirements have been embedded in the design of the programme.

6.7. Intervention is full coverage

A difficult scenario that often arises in the evaluation of governance interventions is one in which policies and programmes apply to all constituencies – examples being new reforms, the passage of new legislation or new implementation rules for audits or criminal codes. Evaluating a programme that covers the entire population of the country is quite difficult. In general, both experiments and quasi-experiments cannot be used since there is no scope for a control group. Two potential approaches could be considered to address this problem.

- If the programme is to be rolled out in stages, the RCT method can be applied. Some groups can be assigned to receive treatment now, while the others receive it a later date. By rolling out the program in stages, ethical considerations of deliberately excluding some participants as controls are eliminated. Refer to Hoddinott and Skoufias (2004) and Schultz (2004) analyzing the impact of PROGRESA by capitalizing on the phased-in design of the program.
- If the programme is of the full-coverage type, but treatment can be administered randomly in varying intensity, its impact can be measured at the different levels of treatment.
- If the programme is to be rolled out all at once and the treatment is to be administered uniformly, a time-series analysis consisting of repeated measures taken periodically before and after the intervention could be used (Rossi and Freeman 1993). For example, a new reform agenda has been implemented nationwide. Time-series analysis entails predicting the post-intervention outcome using the pre-intervention trend. The trend of the predicted outcome becomes the counterfactual, i.e. what would have happened without the reform. The impact of the reform is the comparison between the predicted outcome and the actual post-intervention outcome. The drawback of this approach is that pre-intervention observations must be large enough to produce a robust prediction. Secondly, any treatment effects may not be fully credible owing to the bias resulting from various factors not taken into account. Apart from external factors, another problem is the presence of implementation lags, “announcement effects” and uncertainty as to when the programme actually took effect. This makes it difficult to pin down the exact timing of the programme. Fortunately, such structural breaks in the outcome can be formally tested (see Piehl et al. 1999).

7. Practical implications

The discussion of the application of rigorous impact evaluation and of the challenges encountered in its conduct has implications for the design and implementation of governance interventions. Results of impact evaluations offer reliable knowledge and are important for evidence-based policymaking. The critical question is, then, “How can governance programmes be designed that rigorous impact evaluation becomes possible?” The studies and examples that have been highlighted in this paper call for the discussion of various aspects:

Include evaluator at the outset. The presence of an evaluator in the early phases of programme design can significantly increase the potential of conducting RIE. Evaluations of new programmes call for the evaluator to be involved in all aspects of the programme – including observations on changes in interventions and potential outcomes. This makes it easier to identify components of the programme that can be evaluated quantitatively. Programmes are more difficult to evaluate when they are being run or have been completed, since the evaluator’s assessment is limited to project reports and interviews with donor staff. In addition, the initial indicators and data collected may no longer be appropriate at the time of evaluation. More importantly, the evaluator’s knowledge and hands-on expertise regarding the dynamics of the intervention and the characteristics of the target population may be limited if the evaluation is conducted under severe time constraints. For practitioners, this means obtaining advice from technically skilled evaluators right from the start of the programme.

However, including evaluators at the outset is easier said than done. In most cases, evaluators are consultants hired externally by donor agencies, and keeping them on for the entire duration of the programme is not economical. An alternative approach is to include the evaluator as a member of an advisory committee before project implementation so that appropriate outcome measurements and data collected have a better chance of remaining valid for the evaluation design. For instance, the World Bank’s Development Impact Evaluation (DIME) Initiative provides support teams of experts and research groups to help government counterparts to design and carry out impact evaluations.

Randomise whenever possible. One of the most equitable ways of allocating the intervention is through randomisation. Some programmes have held lotteries to select programme beneficiaries. In some cases, this may not be suitable, especially if policymakers are interested only in the outcome for certain groups. Critics even say that it is “unethical,” since benefits are deliberately withheld from those in greatest need. Nevertheless, if interventions are assigned randomly and if done properly, it is the simplest way of evaluating a programme. This line of thinking also has its drawbacks, as Moehler (2010) notes. The popularity of RCTs should not prevent other quasi-experimental designs from being adopted. Impact evaluation means that the evaluator uses the best possible method (not only RCT), given all the information available to him and all the constraints he faces.

Collect good-quality baseline data. There is a difference between collecting baseline data and *quality* baseline data. Quality baseline data come from careful planning of the programme. The data are collected from the relevant beneficiaries of specific interventions as well as from stakeholders. The data should include both initial characteristics of

the target groups and expected outcomes. This is typically difficult to plan *ex ante* as some indirect outcomes are revealed only during or after implementation.

Collect data on the control groups. Funding the intervention does not necessarily include funding for the collection of control group data. There are cases where funding for control group surveys is not included in the programme. As discussed in the previous chapters, since the control group is a very important source of information in impact evaluation, it is important to provide a budget for its conduct

Operationally, forming control groups is always tricky and difficult in practice. The first difficulty lies in identifying the appropriate control group. Second, once identified, there may be instances in which households or individuals have to be offered incentives to participate in the survey. For instance, to determine the employment probabilities of graduates in schools with and without intervention, a graduate tracer study must be conducted on both. A graduate tracer study requires graduates to be identified, contacted and interviewed so that relevant information can be collected. Schools that are not beneficiaries of the programme (but knew about it) may be reluctant to participate in the survey, especially if they are required to make an additional effort – in terms of time and personnel – to contact their graduates and to convince them to participate. In this case, a letter from a government department (Ministry of Education) endorsing the survey should result in the control schools becoming more cooperative. There are many ways to circumvent this problem, mostly relying on the evaluator's creativity.

Cooperate with the relevant government agencies. The involvement of government agencies in development programmes varies widely. Regardless of the depth of its involvement, the cooperation of the relevant agency must be obtained. The relevant government agencies are often important sources of information on the nature of the target beneficiaries. Being the local experts, they can help identify the likelihood of programme take-up.

Government agencies also have institutional information on such matters as the presence of other donors in the area and their respective interventions. Such information could be useful in preventing bias and contamination issues. In the schooling example above, the cooperation of a government body could help support the evaluation to ensure its successful completion. Although it may not be necessary, the agency's confirmation also helps to lend credibility to the overall result of the evaluation.

Identify components of the programme that permit RIE. At the outset, it is important to identify potential components of the programme to which quantitative analysis can be applied. Early identification means that the best possible method can be used for the evaluation and that the required data can already be gathered.

For some donors and governments, the question whether or not to conduct rigorous impact evaluation is still open. In analyzing the impact of a high profile experimental intervention in African countries called Millennium Villages, Clemens and Demombynes (2010) enumerated the reasons when to consider impact evaluation a "necessity" than a "luxury": first, if the cost of conducting a rigorous evaluation is relatively low; second, if evaluation results can be produced before a policy decision needs to be made; third, if implementing the wrong policy can be disastrous; fourth, when it is impossible for all intended recipients to receive treatment due to constraints (meaning that there would be room for control group); fifth, there are strong interests from policymakers; and finally,

there is reason to believe that the pilot test area is as similar as the scaled up intervention (external validity possible).

There are, of course, interventions where RIE and the methods described earlier may be difficult (even impossible) to use. Consider development interventions, like for instance general budget support, where treatment is typically a mixture of financial and non-financial inputs provided at various levels of government. How can the changes in welfare be attributed to budget support? It may be possible to design certain aspects of budget support *ex ante* to permit appropriate attribution. Whether this is possible or not, further research is clearly needed to answer these questions.

8. Summary and conclusions

Governance is at the heart of the aid effectiveness debate. This debate stems from various empirical studies claiming that aid works only if recipient countries have the appropriate governance structures (Burnside / Dollar 2000; Collier / Dollar 2002; McGillivray et al. 2006; World Bank 1998). Apart from being a prerequisite, governance is also regarded as an objective for aid. The popularity of such new aid modalities as general budget support highlights these two governance functions. While it seeks to strengthen state institutions and public financial management, it also ensures that a minimum level of governance structure has already been committed to minimise fiduciary risk and corruption.

Although billions of dollars have been invested in governance programmes, sound evidence of their impact is still very scarce. This is not to say that good outcomes have not been achieved, but rather that verifying those outcomes is difficult. This paper has therefore discussed the various quantitative methods of governance evaluation and given genuine examples from the field. The scope of the paper covers experimental and quasi-experimental designs in which governance acts: (i) as an intervention aimed at achieving certain development outcomes, (ii) as an outcome of an intervention or (iii) as an intervention and outcome. The paper has presented evidence that certain aspects of governance programmes are suitable for rigorous impact evaluation. Of the four broad dimensions of governance, randomised control trials have proved popular in evaluating the programmes related to political systems and social governance, with specific reference to elections and community development (Moehler 2010). As this paper shows, difference-in-differences and instrumental variables have been applied to public administration and political systems, particularly in the control of corruption (Reinikka / Svensson 2004), voter mobilisation (Arceneaux et al. 2006) and measures to combat crime (Levitt 1997; 2002). Propensity score matching has also been useful in assessing social governance topics, specifically citizen participation (Barabas 2004; Capuno / Garcia 2010). Lastly, the application regression discontinuity design has been rather limited in this field but this paper presented examples on employment and government spending behaviour (Lemieux / Milligan 2008; Pettersson-Lidbom / Tyrefors 2007). The paper has explained each method with real-life examples and has found that rigorous impact evaluation of various governance dimensions is feasible.

As a rule, numerous technical problems emerge when governance programmes are evaluated. They include such evaluation issues as outcome measurement; small sample size; spillovers; a lack of baseline data; evaluation at higher levels of government; com-

plex interventions; and full-coverage programmes. As impact evaluations addressing these problems in the governance context remain scarce, alternative approaches to solving these problems by borrowing from the experience of other sectors have been presented. As previously argued, these problems were not new in the evaluation literature. Such heavily evaluated sectors as education, labour and health have already tackled these challenges.

There are ways of designing governance programmes to enable rigorous impact analysis to be undertaken. As most examples show, it is important to include an evaluator at project inception, to randomise whenever possible, to collect baseline and control group data, to cooperate with relevant government agencies and to identify programme components at the outset.

As regards further research, this paper suggests, firstly, a comprehensive scoping study of governance interventions that have been evaluated rigorously to identify evidence gaps which an evaluation might seek to close and, secondly, a systematic evaluation of a specific governance intervention to build on existing knowledge of 'what works', 'what doesn't' and 'why'. A few isolated evaluations are not sufficient to deliver a strong message on the usefulness of certain interventions. The effectiveness of interventions should therefore be tested under various conditions and in various settings.

In sum, more needs to be done to broaden the application of rigorous impact evaluation of governance in the field. Many questions still remain, but addressing this gap is one step forward in convincing donors to continue supporting governance initiatives and demonstrating to practitioners and policymakers that governance programmes can indeed be rigorously evaluated.

9. References

3ie (International Initiative for Impact Evaluation) (s.a.): Impact evaluation practice: a guide for grantees, New Delhi, London, Washington; online:

<http://www.3ieimpact.org/strategy/pdfs/3ie%20impact%20evaluation%20practice.pdf>
(accessed Dec. 2010)

– (s.a.): Principles for impact evaluation, New Delhi, London, Washington; online:

<http://www.3ieimpact.org/strategy/pdfs/principles%20for%20impact%20evaluation.pdf>
(accessed Dec. 2010)

ADC (Austrian Development Cooperation) (2006): Good governance: policy document, Vienna: ADC; online:

http://www.entwicklung.at/uploads/media/PD_Good_governance_01.pdf
(accessed Jan. 2011)

Angrist, J. / A. Krueger (1999): Empirical strategies in labor economics, in: *Handbook of Labor Economics* 3 (1), 1277–1366

Angrist, J. / J.-S. Pischke (2009): Mostly harmless econometrics: an empiricist's companion, Princeton, NJ: Princeton University Press

Arceneaux, K. / A. Gerber / D. Green (2006): Comparing experimental and matching methods using a large-scale voter mobilization experiment, in: *Political Analysis* 14 (winter), 37–62

Barabas, J. (2004): How deliberation affects policy opinions, in: *American Political Science Review* 98 (4), 687–702

Becker, G. S. / G. J. Stigler (1974): Law enforcement, malfeasance, and the compensation of enforcers, in: *Journal of Legal Studies* 3, 1–19

Becker, S. / M. Caliendo (2007): Sensitivity analysis for average treatment effects, in: *The Stata Journal* 7 (1), 71–83

Becker, S. / A. Ichino (2002): Estimation of average treatment effects based on propensity scores, in: *The Stata Journal* 2 (44), 358–377

Bertrand, M. / S. Djankov / R. Hanna / S. Mullainathan (2007): Obtaining a driver's license in India: an experimental approach to studying corruption, in: *The Quarterly Journal of Economics* 122 (4), 1639–1676

Blank, R. M. (2002): Evaluating welfare reform in the United States, in: *Journal of Economic Literature* 40, 1105–1166

Blattman, C. (2008): Impact evaluation 2.0. Presentation to the Department for International Development (DFID), London

BMZ (Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung) (2008): Budget Support in the Framework of Programme-oriented Joint Financing (PJF), Bonn: Federal Ministry for Economic Cooperation and Development (BMZ); online:

http://www.bmz.de/en/publications/type_of_publication/strategies/konzept181.pdf

- (2009): Promotion of good governance in German development policy, Bonn: Federal Ministry for Economic Cooperation and Development (BMZ); online: http://www.bmz.de/en/publications/type_of_publication/strategies/konzept178.pdf
- Bollen, K. / P. Paxton / R. Morishima* (2005): Assessing international evaluations: an example from USAID's democracy and governance program, in: *American Journal of Evaluation* 26 (2), 189–203
- Booth, D.* (2008): Good governance, aid modalities and poverty reduction: linkages to the Millennium Development Goals and implications for Irish Aid, The Advisory Board for Irish Aid, online: <http://www.odi.org.uk/resources/download/1524.pdf> (accessed Dec.2010)
- Burnside, C. / D. Dollar* (2000): Aid, policies, and growth, in: *American Economic Review* 90 (4), 847–868
- Burtless, G.* (1995): The case for randomized field trials in economic and policy research, in: *The Journal of Economic Perspectives* 9 (2), 63–84
- Caliendo, M. / S. Kopeinig* (2008): Some practical guidance for the implementation of propensity score matching, in: *Journal of Economic Surveys* 22 (1), 31–72
- Cameron, S.* (1988): The economics of crime deterrence: a survey of theory and evidence, in: *Kyklos* 41 (2), 301–323
- Capuno, J. / M. Garcia* (2010): Can local government performance induce civic participation? Evidence from the Philippines, in: *Journal of Development Studies* 46 (4), 624–643
- Caspari, A. / R. Barbu* (2008): Wirkungsevaluierungen: Zum Stand der internationalen Diskussion und dessen Relevanz für Evaluierungen der deutschen Entwicklungszusammenarbeit, *Evaluation Working Papers*, Bonn: Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
- Chattopadhyay, R. / E. Duflo* (2004): Women as policy makers: evidence from a randomized policy experiment in India, in: *Econometrica* 72 (5), 1409–1443
- CIDA (Canadian International Development Agency) (1999): Government of Canada Policy for CIDA on human rights, democratization and good governance, Quebec: CIDA; online: [http://www.acdi-cida.gc.ca/INET/IMAGES.NSF/vLUIImages/HRDG2/\\$file/HRDG-Policy-nophoto-e.pdf](http://www.acdi-cida.gc.ca/INET/IMAGES.NSF/vLUIImages/HRDG2/$file/HRDG-Policy-nophoto-e.pdf) (accessed Jan. 2011)
- Clemens, M. / G. Demombynes* (2010): When does rigorous impact evaluation make a difference? The case of the Millennium Villages, Working Paper 225, Washington, DC: Center for Global Development
- Collier, P. / D. Dollar* (2002): Aid allocation and poverty reduction, in: *European Economic Review* 46 (8), 1475–1500
- Collier, P. / D. Dollar* (2004): Development effectiveness: what have we learnt?, in: *Economic Journal* 114 (496), 244–271
- Collier, P. / Vicente, P.* (2010): Votes and violence: evidence from a field experiment in Nigeria; online: <http://users.ox.ac.uk/~econpco/research/pdfs/VotesandViolence-2010-10.pdf> (accessed Dec. 2010)

- Corman, H. / N. Mocan* (2000): A time series analysis of crime, deterrence, and drug abuse in New York City, in: *American Economic Review* 90 (3), 584–604
- DfID (Department for International Development) (2007): Governance, development and democratic politics: DfID's work in building more effective states, London: DfID; online: <http://webarchive.nationalarchives.gov.uk/+http://www.dfid.gov.uk/pubs/files/governance.pdf> (accessed Jan. 2011)
- Dickson, R. / S. Awasthi / P. Williamson / C. Demellweek / P. Garner* (2000): Effect of treatment for intestinal helminth infection on growth and cognitive performance in children: systematic review of randomized trials, in: *British Medical Journal* 320 (June 24), 1697–1701
- Dubois, P. / A. de Janvry / E. Sadoulet* (2002): Effects of school enrollment and performance of a conditional transfer program in Mexico, Working Paper University of Toulouse
- Duflo, E. / G. Fischer / R. Chattopadhyay* (2005): Efficiency and rent seeking in local government: evidence from randomized policy experiments in India, Cambridge, MA: The Abdul Latif Jameel Poverty Action Lab, Massachusetts Institute of Technology
- Duflo, E. / R. Glennerster / M. Kremer* (2007): Using randomization in development economics research: a toolkit, in: *Handbook of Development Economics* 4, 3895–3962
- Duflo, E. / M. Kremer / J. Robinson* (2008): How high are rates of return to fertilizer? Evidence from field experiments in Kenya, in: *American Economic Review* 98 (2), 482–488
- European Commission* (2010): Second revision of the Cotonou agreement – Agreed Consolidated Text – 11 March 2010, European Commission, Brussels
- Findley, M. / Darren Hawkins / R. Hicks / D. Nielson / B. Parks / R. Powers / J. T. Roberts / M. Tierney / S. Wilson* (2009): AidData: tracking development finance, presented at the PLAID Data Vetting Workshop, Washington, DC, September
- Gaarder, M. / A. Glassman / J. Todd* (2010): Conditional cash transfers and health: unpacking the causal chain, in: *Journal of Development Effectiveness* 2 (1), 6–50
- Gerber, A. / D. Karlan / D. Bergan.* (2008): Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions, New Haven, CT: Yale University
- Gerber, A. S. / D. P. Green* (2000): The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment, in: *The American Political Science Review* 94 (3), 653–663
- Glazerman, S. / D. Levy / D. Meyers* (2003): NX versus experimental estimates of evaluating eaming impacts, in: *Annals of the American Academy of Political and Social Science* 589, 63–93
- Glewwe, P. / M. Kremer* (2006): Schools, teachers, and education outcomes in developing countries, in: *Handbook of the Economics of Education* 2 (Chapter 16), 945–1017

- Glewwe, P. / M. Kremer/ S. Moulin (2009): Many children left behind? Textbooks and test scores in Kenya, in: *American Economic Journal: Applied Economics* 1 (1), 112–135
- Grimes, M. / L. Wängnerud (2010): Curbing corruption through social welfare reform? The effects of Mexico's conditional cash transfer program on good government, in: *American Review of Public Administration* 40 (6), 671–690
- Habicht, J.-P. / G. Pelto / J. Lapp (2009): Methodologies to evaluate the impact of large scale nutrition programs, Doing Impact Evaluation Series, Poverty Reduction and Economic Management Unit, Washington, DC: World Bank
- Handa, S. / A. Peterman / B. Davis/ M. Stampini (2009): Opening up Pandora's box: the effect of gender targeting and conditionality on household spending behavior in Mexico's Progresa program, in: *World Development* 37 (6), 1129–1142
- Hayes, R. / S. Bennett (1999): Simple sample size calculation for cluster-randomized trials, in: *International Journal of Epidemiology* 28, 319–326
- Heckman, J. J. / J. A. Smith (1995): Assessing the case for social experiments, in: *Journal of Economic Perspectives* 9 (2), 85–110
- Hoddinott, J. / E. Skoufias (2004): The Impact of PROGRESA on Food Consumption, in: *Economic Development and Cultural Change* 53 (1), 37–61
- Humphreys, M. / M. William / M. E. Sandbu (2006): The role of leaders in democratic deliberations: results from a field experiment in São Tomé and Príncipe, in: *World Politics* 19 (1), 105–133
- Huntington, S. (1968): Political order in changing societies, New Haven, CT: Yale University Press
- Hyde, S. (2010): Experimenting in democracy promotion: international observers and the 2004 presidential elections in Indonesia, in: *Perspectives on Politics* 8, 511–527
- IEG (Independent Evaluation Group) (s.a.): OED and impact evaluation: a discussion note, World Bank, Washington, DC; online: http://www.worldbank.org/oed/docs/world_bank_oed_impact_evaluations.pdf (accessed Dec. 2010)
- Imai, K. (2005): Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments, in: *American Political Science Review* 99 (2), 283–300
- Imbens, G. / J. Wooldridge (2009): Recent developments in the econometrics of program evaluation, in: *Journal of Economic Literature* 47 (1), 5–86
- IMF (International Monetary Fund) (2007): Manual on fiscal transparency, Washington, DC: IMF
- (1997): Good governance: The IMF's role, Washington, DC: IMF
- Institute of Governance (2010): Governance definition, Ottawa: Institute on Governance. 26 April 2010; online: <http://iog.ca/en/about-us/governance/governance-definition> (accessed Dec. 2010)

- Jacob, B. / L. Lefgren (2004): Remedial education and student achievement: a regression-discontinuity analysis, in: *The Review of Economics and Statistics* 86 (1), 226–244
- Jalan, J. / M. Ravallion (1998): Are there dynamic gains from a poor-area development program?, in: *Journal of Public Economics* 67 (1), 65–85
- Kaufmann, D. / A. Kraay / P. Zoido-Lobaton (1999): Aggregating governance indicators, Policy Research Working Paper 2195, Washington, DC: World Bank
- Khandker, S. / G. Koolwal / H. Samad (2010): Handbook on impact evaluation, Washington, DC: World Bank,
- Leeuw, F. / J. Vaessen (2009): Impact evaluations and development: NONIE Guidance on Impact Evaluation, Network of Networks of Impact Evaluation, online: <http://www.worldbank.org/ieg/nonie/guidance.html> (accessed Dec. 2010)
- Leiderer, S. (2010): Budget support as an aid instrument – neither pandemonium nor panacea, Bonn: Deutsches Institut für Entwicklungspolitik (Briefing Paper 9/2010)
- Lemieux, T. / K. Milligan K. (2008): Incentive effects of social assistance: a regression discontinuity approach, in: *Journal of Econometrics* 142, 807–828
- Levitt, S. (1997): Using electoral cycles in police hiring to estimate the effect of police on crime, in: *American Economic Review* 87 (3), 270–290
- Levitt, S. (2002): Using electoral cycles in police hiring to estimate the effects of police on crime: reply, in: *American Economic Review* 92 (4), 1244–1250
- Marvell, T. / C. Moody (1996): Police levels, crime rates, and specification problems, in: *Criminology* 34, 609–646
- MCC (Millennium Challenge Corporation) (2010): Guide to the MMC indicators and the selection process. Fiscal year 2011; online: http://www.mcc.gov/documents/reports/reference-2010001040503-_fy11guidetotheindicators.pdf (accessed Jan. 2011)
- McGillivray, M. / F. Simon / N. Hermes / R. Lensink (2006): Controversies over the impact of development aid: it works; it doesn't; it can, but that depends, in: *Journal of International Development* 18, 1031–1050
- Mehlum, H. / K. Moene / R. Torvik (2006): Institutions and the resource curse, in: *Economic Journal* 116, 1–20
- Miguel, E. / M. Kremer (2004): Worms: identifying impacts on education and health in the presence of treatment externalities, in: *Econometrica* 72 (1), 159–217
- Moehler, D. (2010): Democracy, governance, and randomized development assistance, in: *Annals of the American Academy of Political and Social Science* 628 (30), 30–46
- Niemi, R. G. / H. Weisberg (2001): Controversies in voting behavior, Washington, DC: CQ Press
- NRC (National Research Council of the National Academies) (2008): Improving democracy assistance: building knowledge through evaluations and research, Washington, DC: National Academies Press

- OECD (*Organisation for Economic Co-operation, Development*) (2008): Survey of donor approaches to governance assessment, Paris: OECD
- Olken, B. (2007): Monitoring corruption: evidence from a field experiment in Indonesia, in: *Journal of Political Economy* 115 (2), 200–249
- Pettersson-Lidbom, P. / B. Tyrefors (2007): The policy consequences of direct versus representative democracy: a regression-discontinuity approach, Stockholm University
- Piehl, A. M. / S. Cooper / A. Braga / D. Kennedy (1999): Testing for structural breaks in the evaluation of programs, NBER Working Paper No. 7226, Cambridge: NBER
- Ravallion, M. (2008): Evaluating anti-poverty programs, Amsterdam, North-Holland: Elsevier
- Ravallion, M. / S. Chen (2005): Hidden impact? Household saving in response to a poor-area development project, in: *Journal of Public Economics* 89, 11/12 (2005), 2183–2204
- Ravallion, M. / E. Galasso / T. Lazo / E. Philipp (2005): What can ex-participants reveal about a program's impact?, in: *The Journal of Human Resources* 40 (1), 208–230
- Reinikka, R. / J. Svensson (2004): The power of information: evidence from a newspaper campaign to reduce capture, World Bank Policy Research Working Paper No. 3239, Washington, DC: World Bank
- Reinikka, R. / J. Svensson (2005): Fighting corruption to improve schooling: evidence from a newspaper campaign in Uganda, in: *Journal of the European Economic Association* 3 (2-3), 259-267
- Rose-Ackerman, S. (1978): Corruption: a study in political economy, New York: Academic Press
- Rosenbaum, P. R. / D. B. Rubin (1983): The central role of the propensity score in observational studies for causal effects, in: *Biometrika* 70, 41–55
- Rossi, P. / H. Freeman (1993): Evaluation: a systematic approach, 5th edition, Los Angeles, CA: Sage Publications
- Ruprah, I. (2008): An impact evaluation of a neighbourhood crime prevention program: does safer commune make chileans safer?, Office of Evaluation and Oversight Working Paper, Inter-American Development Bank, Washington, DC
- Schultz, P. (2004): School subsidies for the poor: evaluating the Mexican Progresa poverty program, in: *Journal of Development Economics* 74 (1), 199–250
- SIDA (Swedish International Development Cooperation Agency) (2003): Shared responsibility: Sweden's policy for global development, Stockholm: SIDA (Government Bill); online: <http://www.sweden.gov.se/content/1/c6/02/45/20/c4527821.pdf> (accessed Jan. 2011)
- UNDP (*United Nations Development Programme*) (2007): Preliminary survey on donor use of governance assessments, Colchester: University of Essex
- (1997): Governance for sustainable human development, New York
- USAID (US Agency for International Development) (1998): Democracy and governance: a conceptual framework, Washington, DC: Center for Democracy and Governance

USAID (Technical Publication Series); online:
http://www.usaid.gov/our_work/democracy_and_governance/publications/pdfs/pnacd395.pdf (accessed Jan. 2011)

– (2004): U.S. foreign aid: meeting the challenges of the twenty-first century, White Paper, Washington, DC: U.S. Agency for International Development

van de Walle, D. (2009): Impact evaluation of rural road projects, in: *Journal of Development Effectiveness* 1 (1), 15–36

Vicente, P. (2010): Does oil corrupt? Evidence from a natural experiment in West Africa, in: *Journal of Development Economics*, 92, 28-38

Waddington, H. / B. Snilstveit / H. White / L. Fewtrell (2009): Water, sanitation and hygiene interventions to combat childhood diarrhoea in developing countries, synthetic review 1, International Initiative for Impact Evaluation (3ie)

Wantchekon, L. (2003): Clientelism and voting behavior: evidence from a field experiment in Benin, in: *World Politics* 55 (3), 399–422

White, H. (2006): Impact evaluation: the experience of the Independent Evaluation Group of the World Bank, Washington, DC: World Bank

White, H. (2009): Theory-based impact evaluation: principles and practice, in: *Journal of Development Effectiveness* 1 (3), 271–284

Wooldridge, J. (2002): *Econometric analysis of cross section and panel data*, Cambridge, MA: MIT Press

Wooldridge, J. M. (2009): *Introductory econometrics: a modern approach*, 4th edition, Mason, OH: South-Western Cengage Learning

World Bank (1992): *Governance and development*, Washington, DC: World Bank

World Bank (1998): *Assessing aid. What works, what doesn't, and why?* New York: Oxford University Press

World Bank (2006): *Budget support as more effective aid? Recent experiences and emerging lessons*, Washington, DC: World Bank

Federal Ministry for Economic Cooperation and Development
Division "Evaluation of Development Cooperation; Auditing"

Bonn office

Federal Ministry for Economic Cooperation and Development
Post-office box 12 03 22
53045 Bonn
Dahlmannstraße 4
53113 Bonn (Germany)

Email: eval@bmz.bund.de

Tel.: + 49 (0) 18 88 5 35 0

Fax: + 49 (0) 18 88 10 5 35 35 00

Berlin office

Federal Ministry for Economic Cooperation and Development
Stresemannstr. 94
10963 Berlin (Germany)
Tel.: + 49 (0) 18 88 25 03 0
Fax: + 49 (0) 18 88 10 25 03 25 95

www.bmz.de

www.bmz.de/en/evaluation/index.html

Editor in Chief and Official Responsible

Michaela Zintl

Asat

February 2011